

# **ECONOMETRICS**

**BRUCE E. HANSEN**  
©2000, 2019<sup>1</sup>

**University of Wisconsin**  
**Department of Economics**

This Revision: August, 2019  
Comments Welcome

<sup>1</sup>This manuscript may be printed and reproduced for individual or instructional use, but may not be printed for commercial purposes.

# Contents

<b>Preface</b>	<b>xv</b>
<b>About the Author</b>	<b>xvi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 What is Econometrics? . . . . .	1
1.2 The Probability Approach to Econometrics . . . . .	1
1.3 Econometric Terms and Notation . . . . .	2
1.4 Observational Data . . . . .	3
1.5 Standard Data Structures . . . . .	4
1.6 Econometric Software . . . . .	6
1.7 Replication . . . . .	6
1.8 Data Files for Textbook . . . . .	7
1.9 Reading the Manuscript . . . . .	9
1.10 Common Symbols . . . . .	10
<b>I Regression</b>	<b>11</b>
<b>2 Conditional Expectation and Projection</b>	<b>12</b>
2.1 Introduction . . . . .	12
2.2 The Distribution of Wages . . . . .	12
2.3 Conditional Expectation . . . . .	14
2.4 Log Differences* . . . . .	16
2.5 Conditional Expectation Function . . . . .	17
2.6 Continuous Variables . . . . .	18
2.7 Law of Iterated Expectations . . . . .	20
2.8 CEF Error . . . . .	21
2.9 Intercept-Only Model . . . . .	23
2.10 Regression Variance . . . . .	23
2.11 Best Predictor . . . . .	24
2.12 Conditional Variance . . . . .	24
2.13 Homoskedasticity and Heteroskedasticity . . . . .	26
2.14 Regression Derivative . . . . .	27
2.15 Linear CEF . . . . .	28
2.16 Linear CEF with Nonlinear Effects . . . . .	29
2.17 Linear CEF with Dummy Variables . . . . .	29
2.18 Best Linear Predictor . . . . .	32
2.19 Illustrations of Best Linear Predictor . . . . .	36
2.20 Linear Predictor Error Variance . . . . .	38
2.21 Regression Coefficients . . . . .	39
2.22 Regression Sub-Vectors . . . . .	40

2.23	Coefficient Decomposition . . . . .	40
2.24	Omitted Variable Bias . . . . .	41
2.25	Best Linear Approximation . . . . .	42
2.26	Regression to the Mean . . . . .	43
2.27	Reverse Regression . . . . .	44
2.28	Limitations of the Best Linear Projection . . . . .	45
2.29	Random Coefficient Model . . . . .	46
2.30	Causal Effects . . . . .	47
2.31	Expectation: Mathematical Details*	51
2.32	Moment Generating and Characteristic Functions*	53
2.33	Moments and Cumulants*	54
2.34	Existence and Uniqueness of the Conditional Expectation*	55
2.35	Identification*	55
2.36	Technical Proofs*	57
	Exercises . . . . .	60
<b>3</b>	<b>The Algebra of Least Squares</b>	<b>63</b>
3.1	Introduction . . . . .	63
3.2	Samples . . . . .	63
3.3	Moment Estimators . . . . .	64
3.4	Least Squares Estimator . . . . .	65
3.5	Solving for Least Squares with One Regressor . . . . .	66
3.6	Solving for Least Squares with Multiple Regressors . . . . .	67
3.7	Illustration . . . . .	72
3.8	Least Squares Residuals . . . . .	73
3.9	Demeaned Regressors . . . . .	74
3.10	Model in Matrix Notation . . . . .	75
3.11	Projection Matrix . . . . .	76
3.12	Orthogonal Projection . . . . .	78
3.13	Estimation of Error Variance . . . . .	79
3.14	Analysis of Variance . . . . .	79
3.15	Projections . . . . .	80
3.16	Regression Components . . . . .	80
3.17	Regression Components (Alternative Derivation)*	83
3.18	Residual Regression . . . . .	84
3.19	Leverage Values . . . . .	85
3.20	Leave-One-Out Regression . . . . .	86
3.21	Influential Observations . . . . .	88
3.22	CPS Data Set . . . . .	90
3.23	Numerical Computation . . . . .	91
3.24	Collinearity Errors . . . . .	91
3.25	Programming . . . . .	93
	Exercises . . . . .	97
<b>4</b>	<b>Least Squares Regression</b>	<b>101</b>
4.1	Introduction . . . . .	101
4.2	Random Sampling . . . . .	101
4.3	Sample Mean . . . . .	102
4.4	Linear Regression Model . . . . .	102
4.5	Mean of Least-Squares Estimator . . . . .	103
4.6	Variance of Least Squares Estimator . . . . .	105
4.7	Unconditional Moments . . . . .	106

4.8	Gauss-Markov Theorem . . . . .	107
4.9	Generalized Least Squares . . . . .	108
4.10	Residuals . . . . .	109
4.11	Estimation of Error Variance . . . . .	111
4.12	Mean-Square Forecast Error . . . . .	112
4.13	Covariance Matrix Estimation Under Homoskedasticity . . . . .	113
4.14	Covariance Matrix Estimation Under Heteroskedasticity . . . . .	114
4.15	Standard Errors . . . . .	117
4.16	Covariance Matrix Estimation with Sparse Dummy Variables . . . . .	118
4.17	Computation . . . . .	119
4.18	Measures of Fit . . . . .	121
4.19	Empirical Example . . . . .	122
4.20	Multicollinearity . . . . .	122
4.21	Clustered Sampling . . . . .	126
4.22	Inference with Clustered Samples . . . . .	132
4.23	At What Level to Cluster? . . . . .	133
	Exercises . . . . .	135
<b>5</b>	<b>Normal Regression and Maximum Likelihood</b>	<b>139</b>
5.1	Introduction . . . . .	139
5.2	The Normal Distribution . . . . .	139
5.3	Chi-Square Distribution . . . . .	142
5.4	Student t Distribution . . . . .	143
5.5	F Distribution . . . . .	144
5.6	Non-Central Chi-Square and F Distributions . . . . .	146
5.7	Joint Normality and Linear Regression . . . . .	147
5.8	Normal Regression Model . . . . .	147
5.9	Distribution of OLS Coefficient Vector . . . . .	149
5.10	Distribution of OLS Residual Vector . . . . .	150
5.11	Distribution of Variance Estimator . . . . .	151
5.12	t-statistic . . . . .	151
5.13	Confidence Intervals for Regression Coefficients . . . . .	152
5.14	Confidence Intervals for Error Variance . . . . .	154
5.15	t Test . . . . .	154
5.16	Likelihood Ratio Test . . . . .	156
5.17	Likelihood Properties . . . . .	157
5.18	Information Bound for Normal Regression . . . . .	159
5.19	Gamma Function* . . . . .	160
5.20	Technical Proofs* . . . . .	160
	Exercises . . . . .	168
<b>II</b>	<b>Large Sample Methods</b>	<b>170</b>
<b>6</b>	<b>An Introduction to Large Sample Asymptotics</b>	<b>171</b>
6.1	Introduction . . . . .	171
6.2	Asymptotic Limits . . . . .	172
6.3	Convergence in Probability . . . . .	173
6.4	Weak Law of Large Numbers . . . . .	174
6.5	Almost Sure Convergence and the Strong Law* . . . . .	175
6.6	Vector-Valued Moments . . . . .	176
6.7	Convergence in Distribution . . . . .	177

6.8	Central Limit Theorem . . . . .	178
6.9	Higher Moments . . . . .	181
6.10	Multivariate Central Limit Theorem . . . . .	182
6.11	Moments of Transformations . . . . .	183
6.12	Smooth Function Model . . . . .	184
6.13	Continuous Mapping Theorem . . . . .	186
6.14	Delta Method . . . . .	186
6.15	Asymptotic Distribution for Smooth Function Model . . . . .	187
6.16	Covariance Matrix Estimation . . . . .	188
6.17	t-ratios . . . . .	188
6.18	Stochastic Order Symbols . . . . .	189
6.19	Uniform WLLN* . . . . .	190
6.20	Uniform CLT* . . . . .	191
6.21	Convergence of Moments* . . . . .	192
6.22	Edgeworth Expansion for the Sample Mean* . . . . .	195
6.23	Edgeworth Expansion for Smooth Function Model* . . . . .	197
6.24	Cornish-Fisher Expansions* . . . . .	199
6.25	Uniform Stochastic Bounds* . . . . .	200
6.26	Marcinkiewicz Weak Law of Large Numbers* . . . . .	201
6.27	Semiparametric Efficiency* . . . . .	201
6.28	Technical Proofs* . . . . .	204
	Exercises . . . . .	210
<b>7</b>	<b>Asymptotic Theory for Least Squares</b>	<b>212</b>
7.1	Introduction . . . . .	212
7.2	Consistency of Least-Squares Estimator . . . . .	212
7.3	Asymptotic Normality . . . . .	214
7.4	Joint Distribution . . . . .	217
7.5	Consistency of Error Variance Estimators . . . . .	220
7.6	Homoskedastic Covariance Matrix Estimation . . . . .	222
7.7	Heteroskedastic Covariance Matrix Estimation . . . . .	222
7.8	Summary of Covariance Matrix Notation . . . . .	224
7.9	Alternative Covariance Matrix Estimators* . . . . .	225
7.10	Functions of Parameters . . . . .	226
7.11	Asymptotic Standard Errors . . . . .	228
7.12	t-statistic . . . . .	230
7.13	Confidence Intervals . . . . .	231
7.14	Regression Intervals . . . . .	233
7.15	Forecast Intervals . . . . .	234
7.16	Wald Statistic . . . . .	236
7.17	Homoskedastic Wald Statistic . . . . .	236
7.18	Confidence Regions . . . . .	237
7.19	Edgeworth Expansion* . . . . .	238
7.20	Semiparametric Efficiency in the Projection Model* . . . . .	239
7.21	Semiparametric Efficiency in the Homoskedastic Regression Model* . . . . .	240
7.22	Uniformly Consistent Residuals* . . . . .	242
7.23	Asymptotic Leverage* . . . . .	243
	Exercises . . . . .	244
<b>8</b>	<b>Restricted Estimation</b>	<b>251</b>
8.1	Introduction . . . . .	251
8.2	Constrained Least Squares . . . . .	252

8.3	Exclusion Restriction . . . . .	253
8.4	Finite Sample Properties . . . . .	254
8.5	Minimum Distance . . . . .	257
8.6	Asymptotic Distribution . . . . .	258
8.7	Variance Estimation and Standard Errors . . . . .	260
8.8	Efficient Minimum Distance Estimator . . . . .	260
8.9	Exclusion Restriction Revisited . . . . .	261
8.10	Variance and Standard Error Estimation . . . . .	263
8.11	Hausman Equality . . . . .	263
8.12	Example: Mankiw, Romer and Weil (1992) . . . . .	264
8.13	Misspecification . . . . .	268
8.14	Nonlinear Constraints . . . . .	270
8.15	Inequality Restrictions . . . . .	271
8.16	Technical Proofs* . . . . .	271
	Exercises . . . . .	273
<b>9</b>	<b>Hypothesis Testing</b>	<b>276</b>
9.1	Hypotheses . . . . .	276
9.2	Acceptance and Rejection . . . . .	277
9.3	Type I Error . . . . .	279
9.4	t tests . . . . .	279
9.5	Type II Error and Power . . . . .	281
9.6	Statistical Significance . . . . .	281
9.7	P-Values . . . . .	282
9.8	t-ratios and the Abuse of Testing . . . . .	284
9.9	Wald Tests . . . . .	285
9.10	Homoskedastic Wald Tests . . . . .	287
9.11	Criterion-Based Tests . . . . .	287
9.12	Minimum Distance Tests . . . . .	288
9.13	Minimum Distance Tests Under Homoskedasticity . . . . .	289
9.14	F Tests . . . . .	290
9.15	Hausman Tests . . . . .	291
9.16	Score Tests . . . . .	292
9.17	Problems with Tests of Nonlinear Hypotheses . . . . .	293
9.18	Monte Carlo Simulation . . . . .	297
9.19	Confidence Intervals by Test Inversion . . . . .	299
9.20	Multiple Tests and Bonferroni Corrections . . . . .	300
9.21	Power and Test Consistency . . . . .	301
9.22	Asymptotic Local Power . . . . .	302
9.23	Asymptotic Local Power, Vector Case . . . . .	305
	Exercises . . . . .	307
<b>10</b>	<b>Resampling Methods</b>	<b>314</b>
10.1	Introduction . . . . .	314
10.2	Example . . . . .	314
10.3	Jackknife Estimation of Variance . . . . .	315
10.4	Example . . . . .	318
10.5	Jackknife for Clustered Observations . . . . .	319
10.6	Empirical Distribution Function . . . . .	320
10.7	Quantiles . . . . .	321
10.8	The Bootstrap Algorithm . . . . .	323
10.9	Bootstrap Variance and Standard Errors . . . . .	325

10.10 Percentile Interval . . . . .	326
10.11 The Bootstrap Distribution . . . . .	327
10.12 The Distribution of the Bootstrap Observations . . . . .	328
10.13 The Distribution of the Bootstrap Sample Mean . . . . .	329
10.14 Bootstrap Asymptotics . . . . .	330
10.15 Consistency of the Bootstrap Estimate of Variance . . . . .	333
10.16 Trimmed Estimator of Bootstrap Variance . . . . .	334
10.17 Unreliability of Untrimmed Bootstrap Standard Errors . . . . .	336
10.18 Consistency of the Percentile Interval . . . . .	336
10.19 Bias-Corrected Percentile Interval . . . . .	338
10.20 BC <sub>a</sub> Percentile Interval . . . . .	340
10.21 Percentile-t Interval . . . . .	342
10.22 Percentile-t Asymptotic Refinement . . . . .	343
10.23 Bootstrap Hypothesis Tests . . . . .	345
10.24 Wald-Type Bootstrap Tests . . . . .	347
10.25 Criterion-Based Bootstrap Tests . . . . .	348
10.26 Parametric Bootstrap . . . . .	349
10.27 How Many Bootstrap Replications? . . . . .	350
10.28 Setting the Bootstrap Seed . . . . .	351
10.29 Bootstrap Regression . . . . .	351
10.30 Bootstrap Regression Asymptotic Theory . . . . .	352
10.31 Wild Bootstrap . . . . .	354
10.32 Bootstrap for Clustered Observations . . . . .	355
10.33 Technical Proofs* . . . . .	357
Exercises . . . . .	362
<b>III Multiple Equation Models</b>	<b>367</b>
<b>11 Multivariate Regression</b>	<b>368</b>
11.1 Introduction . . . . .	368
11.2 Regression Systems . . . . .	368
11.3 Least-Squares Estimator . . . . .	369
11.4 Mean and Variance of Systems Least-Squares . . . . .	371
11.5 Asymptotic Distribution . . . . .	372
11.6 Covariance Matrix Estimation . . . . .	373
11.7 Seemingly Unrelated Regression . . . . .	374
11.8 Equivalence of SUR and Least-Squares . . . . .	376
11.9 Maximum Likelihood Estimator . . . . .	377
11.10 Restricted Estimation . . . . .	378
11.11 Reduced Rank Regression . . . . .	378
11.12 Principal Component Analysis . . . . .	381
11.13 PCA with Additional Regressors . . . . .	383
11.14 Factor-Augmented Regression . . . . .	384
Exercises . . . . .	386
<b>12 Instrumental Variables</b>	<b>388</b>
12.1 Introduction . . . . .	388
12.2 Overview . . . . .	388
12.3 Examples . . . . .	389
12.4 Instruments . . . . .	391
12.5 Example: College Proximity . . . . .	392

12.6	Reduced Form . . . . .	393
12.7	Reduced Form Estimation . . . . .	395
12.8	Identification . . . . .	396
12.9	Instrumental Variables Estimator . . . . .	396
12.10	Demeaned Representation . . . . .	399
12.11	Wald Estimator . . . . .	399
12.12	Two-Stage Least Squares . . . . .	401
12.13	Limited Information Maximum Likelihood . . . . .	403
12.14	JIVE . . . . .	406
12.15	Consistency of 2SLS . . . . .	407
12.16	Asymptotic Distribution of 2SLS . . . . .	408
12.17	Determinants of 2SLS Variance . . . . .	409
12.18	Covariance Matrix Estimation . . . . .	410
12.19	LIML Asymptotic Distribution . . . . .	412
12.20	Functions of Parameters . . . . .	413
12.21	Hypothesis Tests . . . . .	414
12.22	Finite Sample Theory . . . . .	415
12.23	Bootstrap for 2SLS . . . . .	416
12.24	The Peril of Bootstrap 2SLS Standard Errors . . . . .	418
12.25	Clustered Dependence . . . . .	419
12.26	Generated Regressors . . . . .	420
12.27	Regression with Expectation Errors . . . . .	423
12.28	Control Function Regression . . . . .	425
12.29	Endogeneity Tests . . . . .	428
12.30	Subset Endogeneity Tests . . . . .	431
12.31	OverIdentification Tests . . . . .	432
12.32	Subset OverIdentification Tests . . . . .	435
12.33	Bootstrap Overidentification Tests . . . . .	437
12.34	Local Average Treatment Effects . . . . .	438
12.35	Identification Failure . . . . .	441
12.36	Weak Instruments . . . . .	442
12.37	Many Instruments . . . . .	445
12.38	Testing for Weak Instruments . . . . .	448
12.39	Weak Instruments with $k_2 > 1$ . . . . .	455
12.40	Example: Acemoglu, Johnson and Robinson (2001) . . . . .	456
12.41	Example: Angrist and Krueger (1991) . . . . .	458
12.42	Programming . . . . .	460
	Exercises . . . . .	463
<b>13</b>	<b>Generalized Method of Moments</b>	<b>471</b>
13.1	Introduction . . . . .	471
13.2	Moment Equation Models . . . . .	471
13.3	Method of Moments Estimators . . . . .	471
13.4	Overidentified Moment Equations . . . . .	473
13.5	Linear Moment Models . . . . .	474
13.6	GMM Estimator . . . . .	474
13.7	Distribution of GMM Estimator . . . . .	475
13.8	Efficient GMM . . . . .	475
13.9	Efficient GMM versus 2SLS . . . . .	476
13.10	Estimation of the Efficient Weight Matrix . . . . .	477
13.11	Iterated GMM . . . . .	478

13.12 Covariance Matrix Estimation . . . . .	478
13.13 Clustered Dependence . . . . .	479
13.14 Wald Test . . . . .	479
13.15 Restricted GMM . . . . .	480
13.16 Nonlinear Restricted GMM . . . . .	482
13.17 Constrained Regression . . . . .	483
13.18 Multivariate Regression . . . . .	483
13.19 Distance Test . . . . .	484
13.20 Continuously-Updated GMM . . . . .	486
13.21 OverIdentification Test . . . . .	486
13.22 Subset OverIdentification Tests . . . . .	487
13.23 Endogeneity Test . . . . .	488
13.24 Subset Endogeneity Test . . . . .	489
13.25 Nonlinear GMM . . . . .	490
13.26 Bootstrap for GMM . . . . .	491
13.27 Conditional Moment Equation Models . . . . .	492
13.28 Technical Proofs* . . . . .	493
Exercises . . . . .	496
<b>IV Dependent and Panel Data</b>	<b>503</b>
<b>14 Time Series</b>	<b>504</b>
14.1 Introduction . . . . .	504
14.2 Examples . . . . .	505
14.3 Differences and Growth Rates . . . . .	506
14.4 Stationarity . . . . .	507
14.5 Transformations of Stationary Processes . . . . .	509
14.6 Convergent Series . . . . .	510
14.7 Ergodicity . . . . .	511
14.8 Ergodic Theorem . . . . .	513
14.9 Conditioning on Information Sets . . . . .	514
14.10 Martingale Difference Sequences . . . . .	515
14.11 CLT for Martingale Differences . . . . .	517
14.12 Mixing . . . . .	517
14.13 CLT for Correlated Observations . . . . .	519
14.14 Linear Projection . . . . .	520
14.15 White Noise . . . . .	522
14.16 The Wold Decomposition . . . . .	522
14.17 Linear Models . . . . .	523
14.18 Moving Average Processes . . . . .	523
14.19 Infinite-Order Moving Average Process . . . . .	524
14.20 Lag Operator . . . . .	525
14.21 First-Order Autoregressive Process . . . . .	526
14.22 Unit Root and Explosive AR(1) Processes . . . . .	529
14.23 Second-Order Autoregressive Process . . . . .	530
14.24 AR(p) Processes . . . . .	532
14.25 Impulse Response Function . . . . .	533
14.26 ARMA and ARIMA Processes . . . . .	535
14.27 Mixing Properties of Linear Processes . . . . .	535
14.28 Identification . . . . .	536
14.29 Estimation of Autoregressive Models . . . . .	539

14.30	Asymptotic Distribution of Least Squares Estimator . . . . .	540
14.31	Distribution Under Homoskedasticity . . . . .	540
14.32	Asymptotic Distribution Under General Dependence . . . . .	541
14.33	Covariance Matrix Estimation . . . . .	542
14.34	Covariance Matrix Estimation Under General Dependence . . . . .	542
14.35	Testing the Hypothesis of No Serial Correlation . . . . .	544
14.36	Testing for Omitted Serial Correlation . . . . .	544
14.37	Model Selection . . . . .	546
14.38	Illustrations . . . . .	546
14.39	Time Series Regression Models . . . . .	547
14.40	Static, Distributed Lag, and Autoregressive Distributed Lag Models . . . . .	549
14.41	Time Trends . . . . .	550
14.42	Illustration . . . . .	552
14.43	Granger Causality . . . . .	552
14.44	Testing for Serial Correlation in Regression Models . . . . .	555
14.45	Bootstrap for Time Series . . . . .	555
14.46	Technical Proofs* . . . . .	557
14.47	Exercises . . . . .	566
	Exercises . . . . .	566
<b>15</b>	<b>Multivariate Time Series</b>	<b>570</b>
15.1	Introduction . . . . .	570
15.2	Multiple Equation Time Series Models . . . . .	570
15.3	Linear Projection . . . . .	571
15.4	Multivariate Wold Decomposition . . . . .	572
15.5	Impulse Response . . . . .	573
15.6	VAR(1) Model . . . . .	575
15.7	VAR(p) Model . . . . .	575
15.8	Regression Notation . . . . .	576
15.9	Estimation . . . . .	576
15.10	Asymptotic Distribution . . . . .	577
15.11	Covariance Matrix Estimation . . . . .	578
15.12	Selection of Lag Length in an VAR . . . . .	579
15.13	Illustration . . . . .	579
15.14	Predictive Regressions . . . . .	581
15.15	Impulse Response Estimation . . . . .	582
15.16	Local Projection Estimator . . . . .	583
15.17	Regression on Residuals . . . . .	583
15.18	Orthogonalized Shocks . . . . .	584
15.19	Orthogonalized Impulse Response Function . . . . .	585
15.20	Orthogonalized Impulse Response Estimation . . . . .	586
15.21	Illustration . . . . .	586
15.22	Forecast Error Decomposition . . . . .	587
15.23	Identification of Recursive VARs . . . . .	588
15.24	Oil Price Shocks . . . . .	590
15.25	Structural VARs . . . . .	591
15.26	Identification of Structural VARs . . . . .	594
15.27	Long-Run Restrictions . . . . .	595
15.28	Blanchard and Quah (1989) Illustration . . . . .	597
15.29	External Instruments . . . . .	599
15.30	Dynamic Factor Models . . . . .	600

15.31 Technical Proofs*	601
15.32 Exercises	604
Exercises	604
<b>16 Non Stationary Time Series</b>	<b>608</b>
16.1 Introduction	608
16.2 Trend Stationarity	608
16.3 Autoregressive Unit Roots	608
16.4 Cointegration	610
16.5 Cointegrated VARs	611
<b>17 Panel Data</b>	<b>612</b>
17.1 Introduction	612
17.2 Time Indexing and Unbalanced Panels	613
17.3 Notation	614
17.4 Pooled Regression	614
17.5 One-Way Error Component Model	616
17.6 Random Effects	616
17.7 Fixed Effect Model	618
17.8 Within Transformation	620
17.9 Fixed Effects Estimator	622
17.10 Differenced Estimator	623
17.11 Dummy Variables Regression	624
17.12 Fixed Effects Covariance Matrix Estimation	626
17.13 Fixed Effects Estimation in Stata	627
17.14 Between Estimator	628
17.15 Feasible GLS	629
17.16 Intercept in Fixed Effects Regression	630
17.17 Estimation of Fixed Effects	631
17.18 GMM Interpretation of Fixed Effects	631
17.19 Identification in the Fixed Effects Model	633
17.20 Asymptotic Distribution of Fixed Effects Estimator	633
17.21 Asymptotic Distribution for Unbalanced Panels	635
17.22 Heteroskedasticity-Robust Covariance Matrix Estimation	637
17.23 Heteroskedasticity-Robust Estimation – Unbalanced Case	638
17.24 Hausman Test for Random vs Fixed Effects	638
17.25 Random Effects or Fixed Effects?	639
17.26 Time Trends	639
17.27 Two-Way Error Components	640
17.28 Instrumental Variables	642
17.29 Identification with Instrumental Variables	643
17.30 Asymptotic Distribution of Fixed Effects 2SLS Estimator	643
17.31 Linear GMM	645
17.32 Estimation with Time-Invariant Regressors	645
17.33 Hausman-Taylor Model	647
17.34 Jackknife Covariance Matrix Estimation	649
17.35 Panel Bootstrap	650
17.36 Dynamic Panel Models	650
17.37 The Bias of Fixed Effects Estimation	651
17.38 Anderson-Hsiao Estimator	652
17.39 Arellano-Bond Estimator	653
17.40 Weak Instruments	655

17.41	Dynamic Panels with Predetermined Regressors . . . . .	656
17.42	Blundell-Bond Estimator . . . . .	657
17.43	Forward Orthogonal Transformation . . . . .	660
17.44	Empirical Illustration . . . . .	661
	Exercises . . . . .	663
<b>18</b>	<b>Difference in Differences</b>	<b>666</b>
18.1	Introduction . . . . .	666
18.2	Minimum Wage in New Jersey . . . . .	666
18.3	Identification . . . . .	669
18.4	Multiple Units . . . . .	670
18.5	Do Police Reduce Crime? . . . . .	671
18.6	Trend Specification . . . . .	673
18.7	Do Blue Laws Affect Liquor Sales? . . . . .	674
18.8	Check Your Code: Does Abortion Impact Crime? . . . . .	676
18.9	Inference . . . . .	676
	Exercises . . . . .	678
<b>V</b>	<b>Nonparametric and Nonlinear Methods</b>	<b>680</b>
<b>19</b>	<b>Density Estimation</b>	<b>681</b>
19.1	Introduction . . . . .	681
19.2	Histogram Density Estimation . . . . .	681
19.3	Kernel Density Estimator . . . . .	682
19.4	Bias of Density Estimator . . . . .	684
19.5	Variance of Density Estimator . . . . .	687
19.6	Variance Estimation and Standard Errors . . . . .	688
19.7	Clustered Observations . . . . .	688
19.8	IMSE of Density Estimator . . . . .	688
19.9	Optimal Kernel . . . . .	689
19.10	Reference Bandwidth . . . . .	690
19.11	Sheather-Jones Bandwidth* . . . . .	692
19.12	Recommendations for Bandwidth Selection . . . . .	693
19.13	Practical Issues in Density Estimation . . . . .	695
19.14	Computation . . . . .	695
19.15	Asymptotic Distribution . . . . .	696
19.16	Undersmoothing . . . . .	696
19.17	Application . . . . .	697
19.18	Technical Proofs* . . . . .	697
	Exercises . . . . .	701
<b>20</b>	<b>Nonparametric Regression</b>	<b>702</b>
20.1	Introduction . . . . .	702
20.2	Binned Means Estimator . . . . .	702
20.3	Kernel Regression . . . . .	704
20.4	Local Linear Estimator . . . . .	704
20.5	Local Polynomial Estimator . . . . .	705
20.6	Asymptotic Bias . . . . .	707
20.7	Asymptotic Variance . . . . .	708
20.8	AIMSE . . . . .	709
20.9	Boundary Bias . . . . .	711

20.10 Reference Bandwidth . . . . .	712
20.11 Nonparametric Residuals and Prediction Errors . . . . .	713
20.12 Cross-Validation Bandwidth Selection . . . . .	714
20.13 Asymptotic Distribution . . . . .	715
20.14 Undersmoothing . . . . .	718
20.15 Conditional Variance Estimation . . . . .	719
20.16 Variance Estimation and Standard Errors . . . . .	719
20.17 Confidence Bands . . . . .	720
20.18 The Local Nature of Kernel Regression . . . . .	721
20.19 Application to Wage Regression . . . . .	721
20.20 Clustered Observations . . . . .	723
20.21 Application to Testscores . . . . .	725
20.22 Multiple Regressors . . . . .	727
20.23 Curse of Dimensionality . . . . .	729
20.24 Computation . . . . .	729
20.25 Technical Proofs* . . . . .	730
Exercises . . . . .	735
<b>21 Series Regression</b>	<b>737</b>
21.1 Introduction . . . . .	737
21.2 Polynomial Regression . . . . .	738
21.3 Illustrating Polynomial Regression . . . . .	738
21.4 Orthogonal Polynomials . . . . .	739
21.5 Splines . . . . .	741
21.6 Illustrating Spline Regression . . . . .	742
21.7 The Global/Local Nature of Series Regression . . . . .	743
21.8 Stone-Weierstrass and Jackson Approximation Theory . . . . .	745
21.9 Regressor Bounds . . . . .	747
21.10 Matrix Convergence . . . . .	747
21.11 Consistent Estimation . . . . .	749
21.12 Convergence Rate . . . . .	750
21.13 Asymptotic Normality . . . . .	751
21.14 Regression Estimation . . . . .	753
21.15 Undersmoothing . . . . .	753
21.16 Residuals and Regression Fit . . . . .	754
21.17 Cross-Validation Model Selection . . . . .	754
21.18 Variance and Standard Error Estimation . . . . .	755
21.19 Clustered Observations . . . . .	756
21.20 Confidence Bands . . . . .	757
21.21 Uniform Approximations . . . . .	757
21.22 Partially Linear Model . . . . .	758
21.23 Panel Fixed Effects . . . . .	759
21.24 Multiple Regressors . . . . .	759
21.25 Additively Separable Models . . . . .	760
21.26 Nonparametric Instrumental Variables Regression . . . . .	760
21.27 NPIV Identification . . . . .	761
21.28 NPIV Convergence Rate . . . . .	763
21.29 Nonparametric vs Parametric Identification . . . . .	763
21.30 Example: Angrist and Lavy (1999) . . . . .	764
21.31 Technical Proofs* . . . . .	767
Exercises . . . . .	772

<b>22 Regression Discontinuity</b>	<b>775</b>
<b>23 Nonlinear Econometric Models</b>	<b>776</b>
23.1 Introduction . . . . .	776
23.2 Nonlinear Least Squares . . . . .	776
23.3 Least Absolute Deviations . . . . .	779
23.4 Quantile Regression . . . . .	781
23.5 Limited Dependent Variables . . . . .	783
23.6 Binary Choice . . . . .	783
23.7 Count Data . . . . .	784
23.8 Censored Data . . . . .	785
23.9 Sample Selection . . . . .	786
Exercises . . . . .	788
<b>24 Machine Learning</b>	<b>790</b>
24.1 Introduction . . . . .	790
24.2 Model Selection . . . . .	790
24.3 Bayesian Information Criterion . . . . .	793
24.4 Akaike Information Criterion for Regression . . . . .	794
24.5 Akaike Information Criterion for Likelihood . . . . .	797
24.6 Mallows Criterion . . . . .	798
24.7 Cross-Validation Criterion . . . . .	799
24.8 K-Fold Cross-Validation . . . . .	800
24.9 Many Selection Criteria are Similar . . . . .	801
24.10 Relation with Likelihood Ratio Testing . . . . .	802
24.11 Consistent Selection . . . . .	803
24.12 Asymptotic Selection Optimality . . . . .	805
24.13 Focused Information Criterion . . . . .	807
24.14 Best Subset and Stepwise Regression . . . . .	809
24.15 The MSE of Model Selection Estimators . . . . .	810
24.16 Inference After Model Selection . . . . .	812
24.17 Empirical Illustration . . . . .	814
24.18 Shrinkage Methods . . . . .	815
24.19 James-Stein Shrinkage Estimator . . . . .	816
24.20 Derivation of James-Stein Theorem* . . . . .	818
24.21 Interpretation of the Stein Effect . . . . .	820
24.22 Positive Part Estimator . . . . .	820
24.23 Shrinkage Towards Restrictions . . . . .	822
24.24 Group James-Stein . . . . .	823
24.25 Empirical Illustrations . . . . .	824
24.26 Model Averaging . . . . .	827
24.27 Smoothed BIC and AIC . . . . .	829
24.28 Mallows Model Averaging . . . . .	831
24.29 Jackknife (CV) Model Averaging . . . . .	833
24.30 Empirical Illustration . . . . .	834
24.31 Ridge Regression . . . . .	834
24.32 LASSO . . . . .	839
24.33 Computation of the LASSO Estimator . . . . .	841
24.34 Elastic Net . . . . .	842
24.35 Regression Sample Splitting . . . . .	842
24.36 Regression Trees . . . . .	844
24.37 Bagging . . . . .	845

24.38 Random Forests . . . . .	846
24.39 Ensembling . . . . .	846
24.40 Technical Proofs* . . . . .	847
Exercises . . . . .	856
<b>Appendices</b>	<b>859</b>
<b>A Matrix Algebra</b>	<b>859</b>
A.1 Notation . . . . .	859
A.2 Complex Matrices* . . . . .	860
A.3 Matrix Addition . . . . .	860
A.4 Matrix Multiplication . . . . .	861
A.5 Trace . . . . .	862
A.6 Rank and Inverse . . . . .	862
A.7 Orthogonal and Orthonormal Matrices . . . . .	863
A.8 Determinant . . . . .	864
A.9 Eigenvalues . . . . .	865
A.10 Positive Definite Matrices . . . . .	866
A.11 Idempotent Matrices . . . . .	866
A.12 Singular Values . . . . .	867
A.13 Matrix Decompositions . . . . .	867
A.14 Generalized Eigenvalues . . . . .	868
A.15 Extrema of Quadratic Forms . . . . .	869
A.16 Cholesky Decomposition . . . . .	871
A.17 QR Decomposition . . . . .	872
A.18 Solving Linear Systems . . . . .	872
A.19 Algorithmic Matrix Inversion . . . . .	874
A.20 Matrix Calculus . . . . .	875
A.21 Kronecker Products and the Vec Operator . . . . .	876
A.22 Vector Norms . . . . .	877
A.23 Matrix Norms . . . . .	878
<b>B Useful Inequalities</b>	<b>880</b>
B.1 Inequalities for Real Numbers . . . . .	880
B.2 Inequalities for Vectors . . . . .	881
B.3 Inequalities for Matrices . . . . .	881
B.4 Probability Inequalities . . . . .	882
B.5 Proofs* . . . . .	885
<b>References</b>	<b>901</b>

# Preface

This book is intended to serve as the textbook for a first-year graduate course in econometrics.

Students are assumed to have an understanding of multivariate calculus, probability theory, linear algebra, and mathematical statistics. A prior course in undergraduate econometrics would be helpful, but not required. Two excellent undergraduate textbooks are Wooldridge (2015) and Stock and Watson (2014).

For reference, the basic tools of matrix algebra and probability inequalities are reviewed in the Appendix.

For students wishing to deepen their knowledge of matrix algebra in relation to their study of econometrics, I recommend *Matrix Algebra* by Abadir and Magnus (2005).

An excellent introduction to probability and statistics is *Statistical Inference* by Casella and Berger (2002). For those wanting a deeper foundation in probability, I recommend Ash (1972) or Billingsley (1995). For more advanced statistical theory, I recommend Lehmann and Casella (1998), van der Vaart (1998), Shao (2003), and Lehmann and Romano (2005). Probability and statistics textbooks written by econometricians include Ramanathan (1993), Amemiya (1994), Gallant (1997), and Linton (2017).

For further study in econometrics beyond this text, I recommend White (1984) and Davidson (1994) for asymptotic theory, Hamilton (1994) and Kilian and Lütkepohl (2017) for time series methods, Cameron and Trivedi (2005) and Wooldridge (2010) for panel data and discrete response models, and Li and Racine (2007) for nonparametrics and semiparametric econometrics. Beyond these texts, the *Handbook of Econometrics* series provides advanced summaries of contemporary econometric methods and theory.

Alternative PhD-level econometrics textbooks include Theil (1971), Amemiya (1985), Judge, Griffiths, Hill, Lütkepohl, and Lee (1985), Goldberger (1991), Davidson and MacKinnon (1993), Johnston and DiNardo (1997), Davidson (2000), Hayashi (2000), Ruud (2000), Davidson and MacKinnon (2004), Greene (2017) and Magnus (2017). For a focus on applied methods see Angrist and Pischke (2009).

The end-of-chapter exercises are important parts of the text and are meant to help teach students of econometrics. Answers are not provided, and this is intentional.

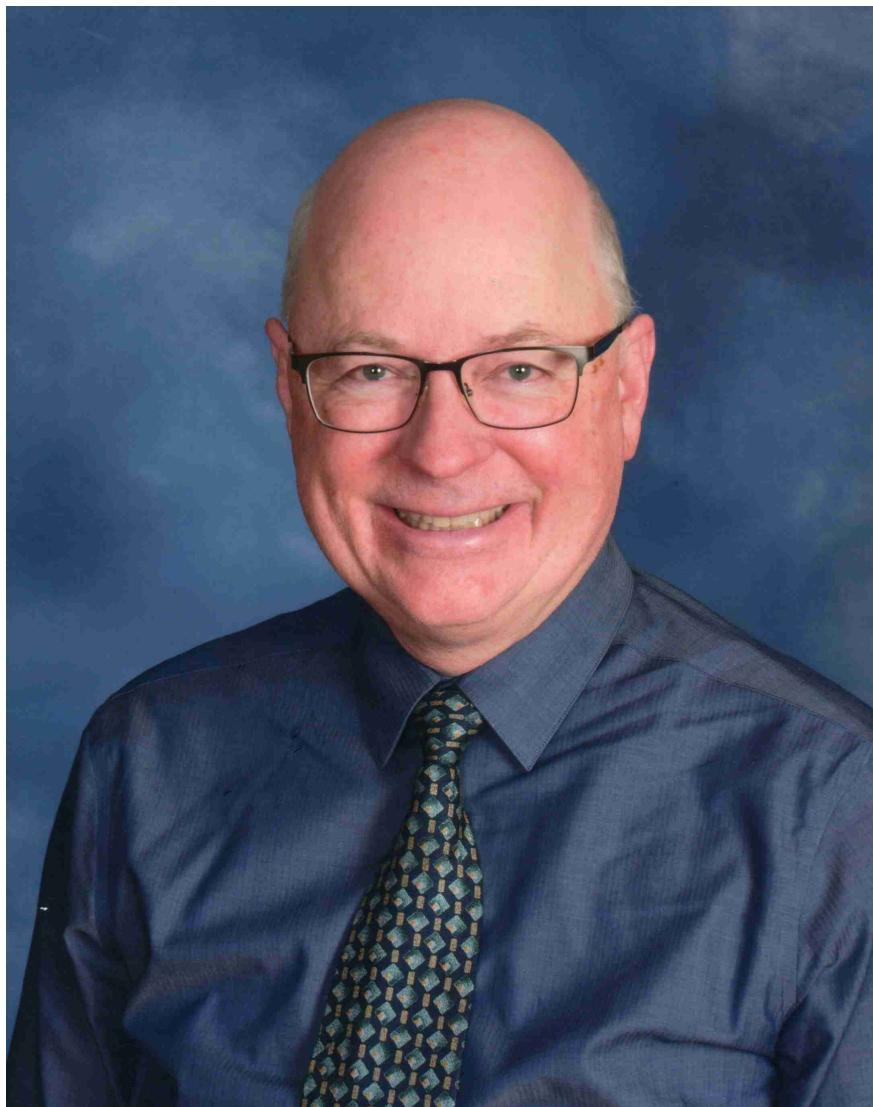
I would like to thank Ying-Ying Lee and Wooyoung Kim for providing research assistance in preparing some of the numerical analysis, graphics, and empirical examples presented in the text.

This is a manuscript in progress. Parts I-III are near complete. Parts IV and V are incomplete, in particular Chapters 16, 22, 23 and 24.

# About the Author

Bruce E. Hansen is the Mary Claire Aschenbrenner Phipps Distinguished Chair of Economics at the University of Wisconsin-Madison. Bruce is originally from Los Angeles, California, has an undergraduate degree in economics from Occidental College, and a Ph.D. in economics from Yale University. He previously taught at the University of Rochester and Boston College.

Bruce is a Fellow of the Econometric Society, the Journal of Econometrics, and the International Association of Applied Econometrics. He has served as Co-Editor of *Econometric Theory* and as Associate Editor of *Econometrica*. He has published 62 papers in refereed journals which have received over 30,000 citations.



# Chapter 1

## Introduction

### 1.1 What is Econometrics?

The term “econometrics” is believed to have been crafted by Ragnar Frisch (1895-1973) of Norway, one of the three principal founders of the Econometric Society, first editor of the journal *Econometrica*, and co-winner of the first Nobel Memorial Prize in Economic Sciences in 1969. It is therefore fitting that we turn to Frisch’s own words in the introduction to the first issue of *Econometrica* to describe the discipline.

A word of explanation regarding the term econometrics may be in order. Its definition is implied in the statement of the scope of the [Econometric] Society, in Section I of the Constitution, which reads: “The Econometric Society is an international society for the advancement of economic theory in its relation to statistics and mathematics.... Its main object shall be to promote studies that aim at a unification of the theoretical-quantitative and the empirical-quantitative approach to economic problems....”

But there are several aspects of the quantitative approach to economics, and no single one of these aspects, taken by itself, should be confounded with econometrics. Thus, econometrics is by no means the same as economic statistics. Nor is it identical with what we call general economic theory, although a considerable portion of this theory has a definitely quantitative character. Nor should econometrics be taken as synonymous with the application of mathematics to economics. Experience has shown that each of these three viewpoints, that of statistics, economic theory, and mathematics, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the *unification* of all three that is powerful. And it is this unification that constitutes econometrics.

Ragnar Frisch, *Econometrica*, (1933), 1, pp. 1-2.

This definition remains valid today, although some terms have evolved somewhat in their usage. Today, we would say that econometrics is the unified study of economic models, mathematical statistics, and economic data.

Within the field of econometrics there are sub-divisions and specializations. **Econometric theory** concerns the development of tools and methods, and the study of the properties of econometric methods. **Applied econometrics** is a term describing the development of quantitative economic models and the application of econometric methods to these models using economic data.

### 1.2 The Probability Approach to Econometrics

The unifying methodology of modern econometrics was articulated by Trygve Haavelmo (1911-1999) of Norway, winner of the 1989 Nobel Memorial Prize in Economic Sciences, in his seminal paper “The

probability approach in econometrics" (1944). Haavelmo argued that quantitative economic models must necessarily be *probability models* (by which today we would mean *stochastic*). Deterministic models are blatantly inconsistent with observed economic quantities, and it is incoherent to apply deterministic models to non-deterministic data. Economic models should be explicitly designed to incorporate randomness; stochastic errors should not be simply added to deterministic models to make them random. Once we acknowledge that an economic model is a probability model, it follows naturally that an appropriate tool way to quantify, estimate, and conduct inferences about the economy is through the powerful theory of mathematical statistics. The appropriate method for a quantitative economic analysis follows from the probabilistic construction of the economic model.

Haavelmo's probability approach was quickly embraced by the economics profession. Today no quantitative work in economics shuns its fundamental vision.

While all economists embrace the probability approach, there has been some evolution in its implementation.

The **structural approach** is the closest to Haavelmo's original idea. A probabilistic economic model is specified, and the quantitative analysis performed under the assumption that the economic model is correctly specified. Researchers often describe this as "taking their model seriously." The structural approach typically leads to likelihood-based analysis, including maximum likelihood and Bayesian estimation.

A criticism of the structural approach is that it is misleading to treat an economic model as correctly specified. Rather, it is more accurate to view a model as a useful abstraction or approximation. In this case, how should we interpret structural econometric analysis? The **quasi-structural approach** to inference views a structural economic model as an approximation rather than the truth. This theory has led to the concepts of the pseudo-true value (the parameter value defined by the estimation problem), the quasi-likelihood function, quasi-MLE, and quasi-likelihood inference.

Closely related is the **semiparametric approach**. A probabilistic economic model is partially specified but some features are left unspecified. This approach typically leads to estimation methods such as least-squares and the Generalized Method of Moments. The semiparametric approach dominates contemporary econometrics, and is the main focus of this textbook.

Another branch of quantitative structural economics is the **calibration approach**. Similar to the quasi-structural approach, the calibration approach interprets structural models as approximations and hence inherently false. The difference is that the calibrationist literature rejects mathematical statistics (deeming classical theory as inappropriate for approximate models) and instead selects parameters by matching model and data moments using non-statistical *ad hoc*<sup>1</sup> methods.

### Trygve Haavelmo

The founding ideas of the field of econometrics are largely due to the Norwegian econometrician Trygve Haavelmo (1911-1999). His advocacy of probability models revolutionized the field, and his use of formal mathematical reasoning laid the foundation for subsequent generations. He was awarded the Nobel Memorial Prize in Economic Sciences in 1989.

## 1.3 Econometric Terms and Notation

In a typical application, an econometrician has a set of repeated measurements on a set of variables. For example, in a labor application the variables could include weekly earnings, educational attainment, age, and other descriptive characteristics. We call this information the **data**, **dataset**, or **sample**.

<sup>1</sup>Ad hoc means "for this purpose" – a method designed for a specific problem – and not based on a generalizable principle.

We use the term **observations** to refer to the distinct repeated measurements on the variables. An individual observation often corresponds to a specific economic unit, such as a person, household, corporation, firm, organization, country, state, city or other geographical region. An individual observation could also be a measurement at a point in time, such as quarterly GDP or a daily interest rate.

Economists typically denote variables by the italicized roman characters  $y$ ,  $x$ , and/or  $z$ . The convention in econometrics is to use the character  $y$  to denote the variable to be explained, while the characters  $x$  and  $z$  are used to denote the conditioning (explaining) variables.

Following mathematical convention, real numbers (elements of the real line  $\mathbb{R}$ , also called **scalars**) are written using lower case italics such as  $y$ , and vectors (elements of  $\mathbb{R}^k$ ) by lower case bold italics such as  $\mathbf{x}$ , e.g.

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}.$$

Upper case bold italics such as  $\mathbf{X}$  are used for matrices.

We denote the number of observations by the natural number  $n$ , and subscript the variables by the index  $i$  to denote the individual observation, e.g.  $y_i$ ,  $\mathbf{x}_i$  and  $\mathbf{z}_i$ . In some contexts we use indices other than  $i$ , such as in time series applications where the index  $t$  is common. In panel studies we typically use the double index  $it$  to refer to individual  $i$  at a time period  $t$

The  $i^{th}$  **observation** is the set  $(y_i, \mathbf{x}_i, \mathbf{z}_i)$ . The **sample** is the set  $\{(y_i, \mathbf{x}_i, \mathbf{z}_i) : i = 1, \dots, n\}$ .

It is proper mathematical practice to use upper case  $X$  for random variables and lower case  $x$  for realizations or specific values. Since we use upper case to denote matrices, the distinction between random variables and their realizations is not rigorously followed in econometric notation. Thus the notation  $y_i$  will in some places refer to a random variable, and in other places a specific realization. This is undesirable but there is little to be done about it without terribly complicating the notation. Hopefully there will be no confusion as the use should be evident from the context.

We typically use Greek letters such as  $\beta$ ,  $\theta$  and  $\sigma^2$  to denote unknown parameters of an econometric model, and will use boldface, e.g.  $\boldsymbol{\beta}$  or  $\boldsymbol{\theta}$ , when these are vector-valued. Estimators are typically denoted by putting a hat “ $\hat{}$ ”, tilde “ $\tilde{}$ ” or bar “ $\bar{}$ ” over the corresponding letter, e.g.  $\hat{\beta}$  and  $\tilde{\beta}$  are estimators of  $\beta$ .

The covariance matrix of an econometric estimator will typically be written using the capital bold-face  $V$ , often with a subscript to denote the estimator, e.g.  $V_{\hat{\beta}} = \text{var}(\hat{\beta})$  as the covariance matrix for  $\hat{\beta}$ . Hopefully without causing confusion, we will use the notation  $V_{\beta} = \text{avar}(\hat{\beta})$  to denote the asymptotic covariance matrix of  $\sqrt{n}(\hat{\beta} - \beta)$  (the variance of the asymptotic distribution). Estimators will be denoted by appending hats or tildes, e.g.  $\hat{V}_{\beta}$  is an estimator of  $V_{\beta}$ .

## 1.4 Observational Data

A common econometric question is to quantify the causal impact of one set of variables on another variable. For example, a concern in labor economics is the returns to schooling – the change in earnings induced by increasing a worker’s education, holding other variables constant. Another issue of interest is the earnings gap between men and women.

Ideally, we would use **experimental** data to answer these questions. To measure the returns to schooling, an experiment might randomly divide children into groups, mandate different levels of education to the different groups, and then follow the children’s wage path after they mature and enter the

labor force. The differences between the groups would be direct measurements of the effects of different levels of education. However, experiments such as this would be widely condemned as immoral! Consequently, in economics non-laboratory experimental data sets are typically narrow in scope.

Instead, most economic data is **observational**. To continue the above example, through data collection we can record the level of a person's education and their wage. With such data we can measure the joint distribution of these variables, and assess the joint dependence. But from observational data it is difficult to infer **causality**, as we are not able to manipulate one variable to see the direct effect on the other. For example, a person's level of education is (at least partially) determined by that person's choices. These factors are likely to be affected by their personal abilities and attitudes towards work. The fact that a person is highly educated suggests a high level of ability, which suggests a high relative wage. This is an alternative explanation for an observed positive correlation between educational levels and wages. High ability individuals do better in school, and therefore choose to attain higher levels of education, and their high ability is the fundamental reason for their high wages. The point is that multiple explanations are consistent with a positive correlation between schooling levels and education. Knowledge of the joint distribution alone may not be able to distinguish between these explanations.

Most economic data sets are observational, not experimental. This means that all variables must be treated as random and possibly jointly determined.

This discussion means that it is difficult to infer causality from observational data alone. Causal inference requires identification, and this is based on strong assumptions. We will discuss these issues on occasion throughout the text.

## 1.5 Standard Data Structures

There are five major types of economic data sets: cross-sectional, time series, panel, clustered, and spatial. They are distinguished by the dependence structure across observations.

Cross-sectional data sets have one observation per individual. Surveys and administrative records are a typical source for cross-sectional data. In typical applications, the individuals surveyed are persons, households, firms or other economic agents. In many contemporary econometric cross-section studies the sample size  $n$  is quite large. It is conventional to assume that cross-sectional observations are mutually independent. Most of this text is devoted to the study of cross-section data.

Time series data are indexed by time. Typical examples include macroeconomic aggregates, prices and interest rates. This type of data is characterized by serial dependence. Most aggregate economic data is only available at a low frequency (annual, quarterly or perhaps monthly) so the sample size is typically much smaller than in cross-section studies. An exception is financial data where data are available at a high frequency (weekly, daily, hourly, or by transaction) so sample sizes can be quite large.

Panel data combines elements of cross-section and time series. These data sets consist of a set of individuals (typically persons, households, or corporations) measured repeatedly over time. The common modeling assumption is that the individuals are mutually independent of one another, but a given individual's observations are mutually dependent. In some panel data contexts, the number of time series observations  $T$  per individual is small while the number of individuals  $n$  is large. In other panel data contexts (for example when countries or states are taken as the unit of measurement) the number of individuals  $n$  can be small while the number of time series observations  $T$  can be moderately large. An important issue in econometric panel data is the treatment of error components.

Clustered samples are increasing popular in applied economics, and is related to panel data. In clustered sampling, the observations are grouped into "clusters" which are treated as mutually independent, yet allowed to be dependent within the cluster. The major difference with panel data is that clustered sampling typically does not explicitly model error component structures, nor the dependence within

clusters, but rather is concerned with inference which is robust to arbitrary forms of within-cluster correlation.

Spatial dependence is another model of interdependence. The observations are treated as mutually dependent according to a spatial measure (for example, geographic proximity). Unlike clustering, spatial models allow all observations to be mutually dependent, and typically rely on explicit modeling of the dependence relationships. Spatial dependence can also be viewed as a generalization of time series dependence.

### Data Structures

- Cross-section
- Time-series
- Panel
- Clustered
- Spatial

As we mentioned above, most of this text will be devoted to cross-sectional data under the assumption of mutually independent observations. By mutual independence we mean that the  $i^{th}$  observation  $(y_i, \mathbf{x}_i, \mathbf{z}_i)$  is independent of the  $j^{th}$  observation  $(y_j, \mathbf{x}_j, \mathbf{z}_j)$  for  $i \neq j$ . (Sometimes the label “independent” is misconstrued. It is a statement about the relationship between observations  $i$  and  $j$ , not a statement about the relationship between  $y_i$  and  $\mathbf{x}_i$  and/or  $\mathbf{z}_i$ .) In this case we say that the data are **independently distributed**.

Furthermore, if the data is randomly gathered, it is reasonable to model each observation as a draw from the same probability distribution. In this case we say that the data are **identically distributed**. If the observations are mutually independent and identically distributed, we say that the observations are **independent and identically distributed, i.i.d.**, or a **random sample**. For most of this text we will assume that our observations come from a random sample.

**Definition 1.1** The observations  $(y_i, \mathbf{x}_i, \mathbf{z}_i)$  are a **sample** from the distribution  $F$  if they are identically distributed across  $i = 1, \dots, n$  with joint distribution  $F$ .

**Definition 1.2** The observations  $(y_i, \mathbf{x}_i, \mathbf{z}_i)$  are a **random sample** if they are mutually independent and identically distributed (**i.i.d.**) across  $i = 1, \dots, n$ .

In the random sampling framework, we think of an individual observation  $(y_i, \mathbf{x}_i, \mathbf{z}_i)$  as a realization from a joint probability distribution  $F(\mathbf{y}, \mathbf{x}, \mathbf{z})$  which we can call the **population**. This “population” is infinitely large. This abstraction can be a source of confusion as it does not correspond to a physical population in the real world. It is an abstraction since the distribution  $F$  is unknown, and the goal of statistical inference is to learn about features of  $F$  from the sample. The *assumption* of random sampling

provides the mathematical foundation for treating economic statistics with the tools of mathematical statistics.

The random sampling framework was a major intellectual breakthrough of the late 19th century, allowing the application of mathematical statistics to the social sciences. Before this conceptual development, methods from mathematical statistics had not been applied to economic data as the latter was viewed as non-random. The random sampling framework enabled economic samples to be treated as random, a necessary precondition for the application of statistical methods.

## 1.6 Econometric Software

Economists use a variety of econometric, statistical, and programming software.

Stata ([www.stata.com](http://www.stata.com)) is a powerful statistical program with a broad set of pre-programmed econometric and statistical tools. It is quite popular among economists, and is continuously being updated with new methods. It is an excellent package for most econometric analysis, but is limited when you want to use new or less-common econometric methods which have not yet been programmed. At many points in this textbook specific Stata estimation methods and commands are described. These commands are valid for Stata version 15.

MATLAB ([www.mathworks.com](http://www.mathworks.com)), GAUSS ([www.aptech.com](http://www.aptech.com)), and OxMetrics ([www.oxmetrics.net](http://www.oxmetrics.net)) are high-level matrix programming languages with a wide variety of built-in statistical functions. Many econometric methods have been programmed in these languages and are available on the web. The advantage of these packages is that you are in complete control of your analysis, and it is easier to program new methods than in Stata. Some disadvantages are that you have to do much of the programming yourself, programming complicated procedures takes significant time, and programming errors are hard to prevent and difficult to detect and eliminate. Of these languages, GAUSS used to be quite popular among econometricians, but currently MATLAB is more popular.

An intermediate choice is R ([www.r-project.org](http://www.r-project.org)). R has the capabilities of the above high-level matrix programming languages, but also has many built-in statistical environments which can replicate much of the functionality of Stata. R is the dominate programming language in the statistic field, so methods developed in that arena are most commonly available in R. Uniquely, R is open-source, user-contributed, and best of all, completely free! A smaller but growing group of econometricians are enthusiastic fans of R.

For highly-intensive computational tasks, some economists write their programs in a standard programming language such as Fortran or C. This can lead to major gains in computational speed, at the cost of increased time in programming and debugging.

There are many other packages which are used by econometricians, include Eviews, Gretl, PcGive, Python, RATS, SAS.

As the packages described above have distinct advantages, many empirical economists end up using more than one package. As a student of econometrics, you will learn at least one of these packages, and probably more than one. My advice is that all students of econometrics should develop a basic level of familiarity with Stata, and either Matlab or R (or all three).

## 1.7 Replication

Scientific research needs to be documented and replicable. For social science research using observational data, this requires careful documentation and archiving of the research methods, data manipulations, and coding.

The best practice is as follows. Accompanying each published paper an author should create a complete replication package (set of data files, documentation, and program code files). This package should contain the source (raw) data used for analysis, and code which executes the empirical analysis and other numerical work reported in the paper. In most cases this is a set of programs, which may need to be executed sequentially. (For example, there may be an initial program which “cleans” and manipulates

the data, and then a second set of programs which estimate the reported models.) The ideal is full documentation and clarity. This package should be posted on the author(s) website, and posted at the journal website when that is an option.

A complicating factor is that many current economic data sets have restricted access and cannot be shared without permission. In these cases the data cannot be posted nor shared. The computed code, however, can and should be posted.

Most journals in economics require authors of published papers to make their datasets generally available. For example:

*Econometrica* states:

*Econometrica* has the policy that all empirical, experimental and simulation results must be replicable. Therefore, authors of accepted papers must submit data sets, programs, and information on empirical analysis, experiments and simulations that are needed for replication and some limited sensitivity analysis.

The *American Economic Review* states:

All data used in analysis must be made available to any researcher for purposes of replication.

The *Journal of Political Economy* states:

It is the policy of the *Journal of Political Economy* to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication.

If you are interested in using the data from a published paper, first check the journal's website, as many journals archive data and replication programs online. Second, check the website(s) of the paper's author(s). Most academic economists maintain webpages, and some make available replication files complete with data and programs. If these investigations fail, email the author(s), politely requesting the data. You may need to be persistent.

As a matter of professional etiquette, all authors absolutely have the obligation to make their data and programs available. Unfortunately, many fail to do so, and typically for poor reasons. The irony of the situation is that it is typically in the best interests of a scholar to make as much of their work (including all data and programs) freely available, as this only increases the likelihood of their work being cited and having an impact.

Keep this in mind as you start your own empirical project. Remember that as part of your end product, you will need (and want) to provide all data and programs to the community of scholars. The greatest form of flattery is to learn that another scholar has read your paper, wants to extend your work, or wants to use your empirical methods. In addition, public openness provides a healthy incentive for transparency and integrity in empirical analysis.

## 1.8 Data Files for Textbook

On the textbook webpage <http://www.ssc.wisc.edu/~bhansen/econometrics/> there are posted a number of files containing data sets which are used in this textbook both for illustration and for end-of-chapter empirical exercises. For each data sets there are four files: (1) Description (pdf format); (2) Excel data file; (3) Text data file; (4) Stata data file. The three data files are identical in content, the observations and variables are listed in the same order in each, all have variable labels.

For example, the text makes frequent reference to a wage data set extracted from the Current Population Survey. This data set is named `cps09mar`, and is represented by the files `cps09mar_description.pdf`, `cps09mar.xlsx`, `cps09mar.txt`, and `cps09mar.dta`.

The data sets currently included are

- AB1991
  - Data file from Arellano and Bond (1991)
- AJR2001
  - Data file from Acemoglu, Johnson and Robinson (2001)
- AK1991
  - Data file from Angrist and Krueger (1991)
- AL1999
  - Data file from Angrist and Lavy (1999)
- BMN2016
  - Data file from Bernheim, Meer and Novarro (2016)
- cps09mar
  - household survey data extracted from the March 2009 Current Population Survey
- Card1995
  - Data file from Card (1995)
- CHJ2004
  - Data file from Cox, Hansen and Jimenez (2004)
- CK1994
  - Data file from Card and Krueger (1994)
- DDK2011
  - Data file from Duflo, Dupas and Kremer (2011)
- DS2004
  - Data file from DiTella and Schargrodsky (2004)
- FRED-MD and FRED-QD
  - U.S. monthly and quarterly macroeconomic databases from McCracken and Ng (2015)
- Invest1993
  - Data file from Hall and Hall (1993)
- Kilian2009
  - Data file from Kilian (2009)

- MRW1992
  - Data file from Mankiw, Romer and Weil (1992)
- Nerlove1963
  - Data file from Nerlov (1963)
- RR2010
  - Data file from Reinhard and Rogoff (2010)

## 1.9 Reading the Manuscript

I have endeavored to use a unified notation and nomenclature. The development of the material is cumulative, with later chapters building on the earlier ones. Nevertheless, every attempt has been made to make each chapter self-contained, so readers can pick and choose topics according to their interests.

To fully understand econometric methods, it is necessary to have a mathematical understanding of its mechanics, and this includes the mathematical proofs of the main results. Consequently, this text is self-contained, with nearly all results proved with full mathematical rigor. The mathematical development and proofs aim at brevity and conciseness (sometimes described as mathematical elegance), but also at pedagogy. To understand a mathematical proof, it is not sufficient to simply *read* the proof, you need to follow it, and re-create it for yourself.

Nevertheless, many readers will not be interested in each mathematical detail, explanation, or proof. This is okay. To use a method it may not be necessary to understand the mathematical details. Accordingly I have placed the more technical mathematical proofs and details in chapter appendices. These appendices and other technical sections are marked with an asterisk (\*). These sections can be skipped without any loss in exposition.

The key concepts of matrix algebra and probability inequalities are reviewed in Appendices A & B. It may be useful to read or review Appendix A.1-A.11 before starting Chapter 3, and review Appendix B before Chapter 6. It is not necessary to understand all the material in the appendices. They are intended to be reference material and many list results are not used in this textbook.

## 1.10 Common Symbols

$y$	scalar
$\mathbf{x}$	vector
$X$	matrix
$\mathbb{R}$	real line
$\mathbb{R}^k$	Euclidean $k$ space
$\mathbb{E}(y)$	mathematical expectation
$\text{var}(y)$	variance
$\text{cov}(x, y)$	covariance
$\text{var}(\mathbf{x})$	covariance matrix
$\text{corr}(x, y)$	correlation
$\mathbb{P}$	probability
$\longrightarrow$	limit
$\xrightarrow{p}$	convergence in probability
$\xrightarrow{d}$	convergence in distribution
$\text{plim}_{n \rightarrow \infty}$	probability limit
$N(0, 1)$	standard normal distribution
$N(\mu, \sigma^2)$	normal distribution with mean $\mu$ and variance $\sigma^2$
$\chi_k^2$	chi-square distribution with $k$ degrees of freedom
$I_n$	$n \times n$ identity matrix
$\mathbf{1}_n$	$n \times 1$ vector of ones
$\text{tr } A$	trace
$A'$	matrix transpose
$A^{-1}$	matrix inverse
$A > 0$	positive definite
$A \geq 0$	positive semi-definite
$\ \mathbf{a}\ $	Euclidean norm
$\ A\ $	matrix (Frobenius or spectral) norm
$\mathbf{1}(a)$	indicator function (1 if $a$ is true, else 0)
$\simeq$	approximate equality
$\stackrel{\text{def}}{=}$	definitional equality
$\sim$	is distributed as
$\log$	natural logarithm

## **Part I**

# **Regression**

## Chapter 2

# Conditional Expectation and Projection

### 2.1 Introduction

The most commonly applied econometric tool is least-squares estimation, also known as **regression**. As we will see, least-squares is a tool to estimate an approximate conditional mean of one variable (the **dependent variable**) given another set of variables (the **regressors**, **conditioning variables**, or **covariates**).

In this chapter we abstract from estimation, and focus on the probabilistic foundation of the conditional expectation model and its projection approximation.

### 2.2 The Distribution of Wages

Suppose that we are interested in wage rates in the United States. Since wage rates vary across workers, we cannot describe wage rates by a single number. Instead, we can describe wages using a probability distribution. Formally, we view the wage of an individual worker as a random variable *wage* with the **probability distribution**

$$F(u) = \mathbb{P}(\text{wage} \leq u).$$

When we say that a person's wage is random we mean that we do not know their wage before it is measured, and we treat observed wage rates as realizations from the distribution *F*. Treating unobserved wages as random variables and observed wages as realizations is a powerful mathematical abstraction which allows us to use the tools of mathematical probability.

A useful thought experiment is to imagine dialing a telephone number selected at random, and then asking the person who responds to tell us their wage rate. (Assume for simplicity that all workers have equal access to telephones, and that the person who answers your call will respond honestly.) In this thought experiment, the wage of the person you have called is a single draw from the distribution *F* of wages in the population. By making many such phone calls we can learn the distribution *F* of the entire population.

When a distribution function *F* is differentiable we define the probability density function

$$f(u) = \frac{d}{du} F(u).$$

The density contains the same information as the distribution function, but the density is typically easier to visually interpret.

In Figure 2.1 we display estimates<sup>1</sup> of the probability distribution function (on the left) and density function (on the right) of U.S. wage rates in 2009. We see that the density is peaked around \$15, and most

---

<sup>1</sup>The distribution and density are estimated nonparametrically from the sample of 50,742 full-time non-military wage-earners reported in the March 2009 Current Population Survey. The wage rate is constructed as annual individual wage and salary earnings divided by hours worked.

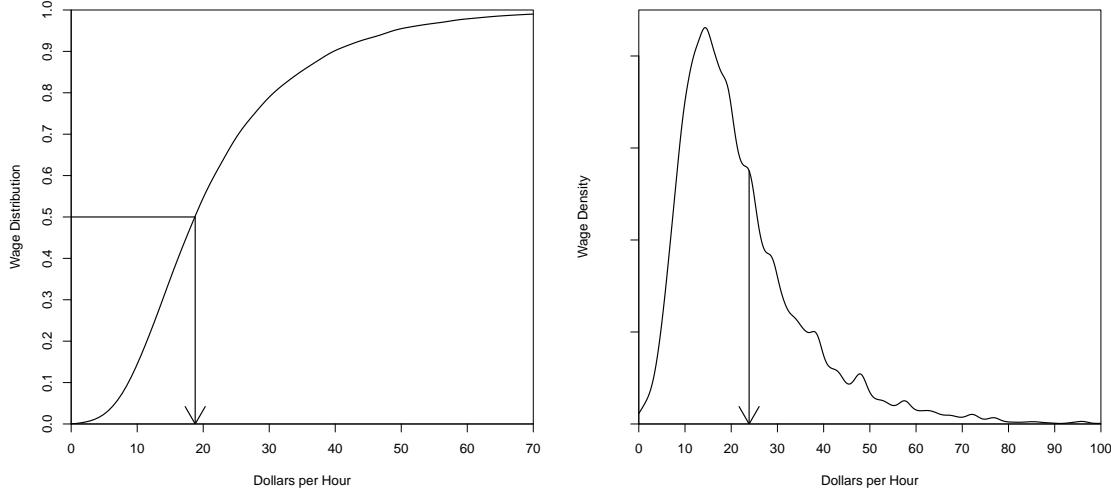


Figure 2.1: Wage Distribution and Density. All Full-time U.S. Workers

of the probability mass appears to lie between \$10 and \$40. These are ranges for typical wage rates in the U.S. population.

Important measures of central tendency are the median and the mean. The **median**  $m$  of a continuous<sup>2</sup> distribution  $F$  is the unique solution to

$$F(m) = \frac{1}{2}.$$

The median U.S. wage (\$19.23) is indicated in the left panel of Figure 2.1 by the arrow. The median is a robust<sup>3</sup> measure of central tendency, but it is tricky to use for many calculations as it is not a linear operator.

The **expectation** or **mean** of a random variable  $y$  with density  $f$  is

$$\mu = \mathbb{E}(y) = \int_{-\infty}^{\infty} u f(u) du.$$

Here we have used the common and convenient convention of using the single character  $y$  to denote a random variable, rather than the more cumbersome label *wage*. A general definition of the mean is presented in Section 2.31. The mean U.S. wage (\$23.90) is indicated in the right panel of Figure 2.1 by the arrow.

We sometimes use the notation  $\mathbb{E}y$  instead of  $\mathbb{E}(y)$  when the variable whose expectation is being taken is clear from the context. There is no distinction in meaning.

The mean is a convenient measure of central tendency because it is a linear operator and arises naturally in many economic models. A disadvantage of the mean is that it is not robust<sup>4</sup> especially in the presence of substantial skewness or thick tails, which are both features of the wage distribution as can be seen easily in the right panel of Figure 2.1. Another way of viewing this is that 64% of workers earn less than the mean wage of \$23.90, suggesting that it is incorrect to describe the mean as a “typical” wage rate.

In this context it is useful to transform the data by taking the natural logarithm<sup>5</sup>. Figure 2.2 shows the density of log hourly wages  $\log(wage)$  for the same population, with its mean 2.95 drawn in with the

<sup>2</sup>If  $F$  is not continuous the definition is  $m = \inf\{u : F(u) \geq \frac{1}{2}\}$

<sup>3</sup>The median is not sensitive to perturbations in the tails of the distribution.

<sup>4</sup>The mean is sensitive to perturbations in the tails of the distribution.

<sup>5</sup>Throughout the text, we will use  $\log(y)$  or  $\log y$  to denote the natural logarithm of  $y$ .

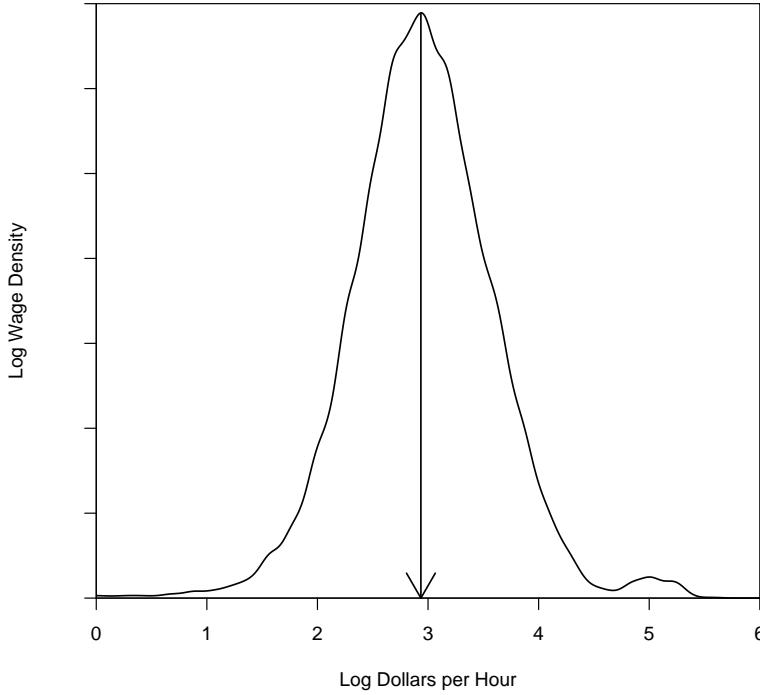


Figure 2.2: Log Wage Density

arrow. The density of log wages is much less skewed and fat-tailed than the density of the level of wages, so its mean

$$\mathbb{E}(\log(wage)) = 2.95$$

is a much better (more robust) measure<sup>6</sup> of central tendency of the distribution. For this reason, wage regressions typically use log wages as a dependent variable rather than the level of wages.

Another useful way to summarize the probability distribution  $F(u)$  is in terms of its quantiles. For any  $\alpha \in (0, 1)$ , the  $\alpha^{th}$  quantile of the continuous<sup>7</sup> distribution  $F$  is the real number  $q_\alpha$  which satisfies

$$F(q_\alpha) = \alpha.$$

The quantile function  $q_\alpha$ , viewed as a function of  $\alpha$ , is the inverse of the distribution function  $F$ . The most commonly used quantile is the median, that is,  $q_{0.5} = m$ . We sometimes refer to quantiles by the percentile representation of  $\alpha$ , and in this case they are often called percentiles, e.g. the median is the 50<sup>th</sup> percentile.

## 2.3 Conditional Expectation

We saw in Figure 2.2 the density of log wages. Is this distribution the same for all workers, or does the wage distribution vary across subpopulations? To answer this question, we can compare wage distributions for different groups – for example, men and women. The plot on the left in Figure 2.3 displays the densities of log wages for U.S. men and women with their means (3.05 and 2.81) indicated by the arrows. We can see that the two wage densities take similar shapes but the density for men is somewhat shifted to the right with a higher mean.

<sup>6</sup>More precisely, the geometric mean  $\exp(\mathbb{E}(\log w)) = \$19.11$  is a robust measure of central tendency.

<sup>7</sup>If  $F$  is not continuous the definition is  $q_\alpha = \inf\{u : F(u) \geq \alpha\}$

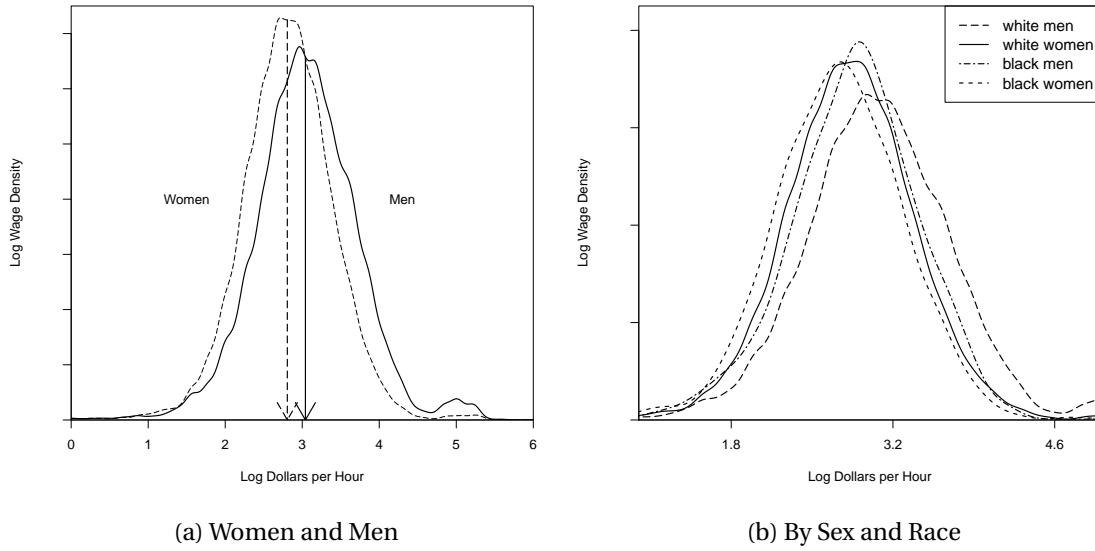


Figure 2.3: Log Wage Density by Sex and Race

The values 3.05 and 2.81 are the mean log wages in the subpopulations of men and women workers. They are called the **conditional means** (or **conditional expectations**) of log wages given sex. We can write their specific values as

$$\mathbb{E}(\log(wage) | sex = man) = 3.05$$

$$\mathbb{E}(\log(wage) | sex = woman) = 2.81.$$

We call these means *conditional* as they are conditioning on a fixed value of the variable *sex*. While you might not think of a person's sex as a random variable, it is random from the viewpoint of econometric analysis. If you randomly select an individual, the sex of the individual is unknown and thus random. (In the population of U.S. workers, the probability that a worker is a woman happens to be 43%.) In observational data, it is most appropriate to view all measurements as random variables, and the means of subpopulations are then conditional means.

As the two densities in Figure 2.3 appear similar, a hasty inference might be that there is not a meaningful difference between the wage distributions of men and women. Before jumping to this conclusion let us examine the differences in the distributions more carefully. As we mentioned above, the primary difference between the two densities appears to be their means. This difference equals

$$\mathbb{E}(\log(wage) | sex = man) - \mathbb{E}(\log(wage) | sex = woman) = 3.05 - 2.81 \\ = 0.24.$$

A difference in expected log wages of 0.24 implies an average 24% difference between the wages of men and women, which is quite substantial. (For an explanation of logarithmic and percentage differences see Section 2.4.)

Consider further splitting the men and women subpopulations by race, dividing the population into whites, blacks, and other races. We display the log wage density functions of four of these groups on the right in Figure 2.3. Again we see that the primary difference between the four density functions is their central tendency.

Focusing on the means of these distributions, Table 2.1 reports the mean log wage for each of the six sub-populations.

The entries in Table 2.1 are the conditional means of  $\log(wage)$  given *sex* and *race*. For example

$$\mathbb{E}(\log(wage) | sex = man, race = white) = 3.07$$

Table 2.1: Mean Log Wages by Sex and Race

	men	women
white	3.07	2.82
black	2.86	2.73
other	3.03	2.86

and

$$\mathbb{E}(\log(wage) | sex = woman, race = black) = 2.73.$$

One benefit of focusing on conditional means is that they reduce complicated distributions to a single summary measure, and thereby facilitate comparisons across groups. Because of this simplifying property, conditional means are the primary interest of regression analysis and are a major focus in econometrics.

Table 2.1 allows us to easily calculate average wage differences between groups. For example, we can see that the wage gap between men and women continues after disaggregation by race, as the average gap between white men and white women is 25%, and that between black men and black women is 13%. We also can see that there is a race gap, as the average wages of blacks are substantially less than the other race categories. In particular, the average wage gap between white men and black men is 21%, and that between white women and black women is 9%.

## 2.4 Log Differences\*

A useful approximation for the natural logarithm for small  $x$  is

$$\log(1 + x) \approx x. \quad (2.1)$$

This can be derived from the infinite series expansion of  $\log(1 + x)$ :

$$\begin{aligned} \log(1 + x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots \\ &= x + O(x^2). \end{aligned}$$

The symbol  $O(x^2)$  means that the remainder is bounded by  $Ax^2$  as  $x \rightarrow 0$  for some  $A < \infty$ . A plot of  $\log(1 + x)$  and the linear approximation  $x$  is shown in Figure 2.4. We can see that  $\log(1 + x)$  and the linear approximation  $x$  are very close for  $|x| \leq 0.1$ , and reasonably close for  $|x| \leq 0.2$ , but the difference increases with  $|x|$ .

Now, if  $y^*$  is  $c\%$  greater than  $y$ , then

$$y^* = (1 + c/100)y.$$

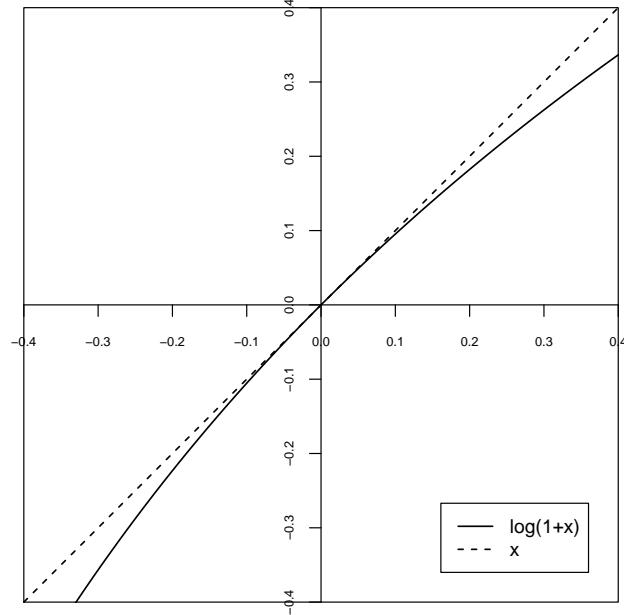
Taking natural logarithms,

$$\log y^* = \log y + \log(1 + c/100)$$

or

$$\log y^* - \log y = \log(1 + c/100) \approx \frac{c}{100}$$

where the approximation is (2.1). This shows that 100 multiplied by the difference in logarithms is approximately the percentage difference between  $y$  and  $y^*$ , and this approximation is quite good for  $|c| \leq 10$ .

Figure 2.4:  $\log(1 + x)$ 

## 2.5 Conditional Expectation Function

An important determinant of wage levels is education. In many empirical studies economists measure educational attainment by the number of years<sup>8</sup> of schooling, and we will write this variable as *education*.

The conditional mean of log wages given *sex*, *race*, and *education* is a single number for each category. For example

$$\mathbb{E}(\log(wage) | sex = man, race = white, education = 12) = 2.84.$$

We display in Figure 2.5 the conditional means of  $\log(wage)$  for white men and white women as a function of *education*. The plot is quite revealing. We see that the conditional mean is increasing in years of education, but at a different rate for schooling levels above and below nine years. Another striking feature of Figure 2.5 is that the gap between men and women is roughly constant for all education levels. As the variables are measured in logs this implies a constant average percentage gap between men and women regardless of educational attainment.

In many cases it is convenient to simplify the notation by writing variables using single characters, typically  $y$ ,  $x$  and/or  $z$ . It is conventional in econometrics to denote the dependent variable (e.g.  $\log(wage)$ ) by the letter  $y$ , a conditioning variable (such as *sex*) by the letter  $x$ , and multiple conditioning variables (such as *race*, *education* and *sex*) by the subscripted letters  $x_1, x_2, \dots, x_k$ .

Conditional expectations can be written with the generic notation

$$\mathbb{E}(y | x_1, x_2, \dots, x_k) = m(x_1, x_2, \dots, x_k).$$

We call this the **conditional expectation function** (CEF). The CEF is a function of  $(x_1, x_2, \dots, x_k)$  as it varies

---

<sup>8</sup>Here, *education* is defined as years of schooling beyond kindergarten. A high school graduate has *education*=12, a college graduate has *education*=16, a Master's degree has *education*=18, and a professional degree (medical, law or PhD) has *education*=20.

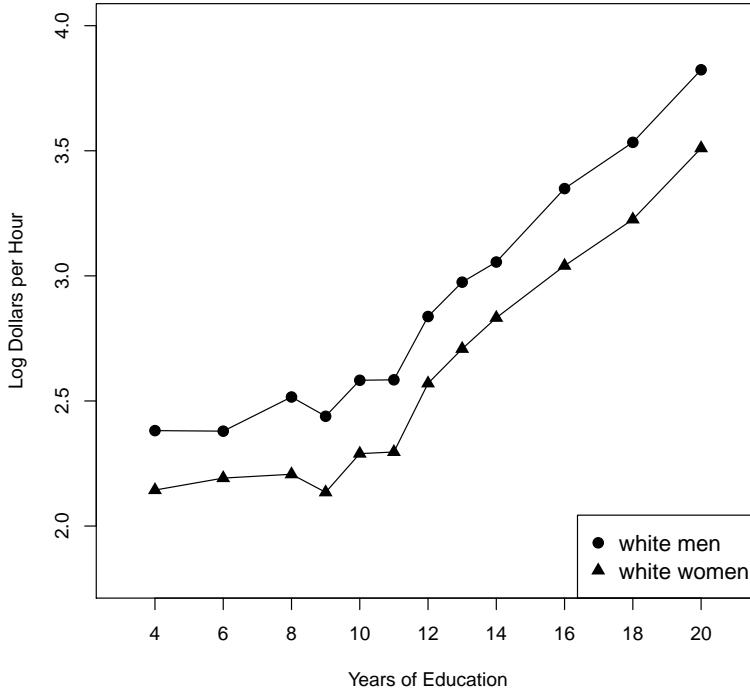


Figure 2.5: Mean Log Wage as a Function of Years of Education

with the variables. For example, the conditional expectation of  $y = \log(wage)$  given  $(x_1, x_2) = (sex, race)$  is given by the six entries of Table 2.1. The CEF is a function of  $(sex, race)$  as it varies across the entries.

For greater compactness, we will typically write the conditioning variables as a vector in  $\mathbb{R}^k$ :

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}. \quad (2.2)$$

Here we follow the convention of using lower case bold italics  $\boldsymbol{x}$  to denote a vector. Given this notation, the CEF can be compactly written as

$$\mathbb{E}(y | \boldsymbol{x}) = m(\boldsymbol{x}).$$

The CEF  $\mathbb{E}(y | \boldsymbol{x})$  is a random variable as it is a function of the random variable  $\boldsymbol{x}$ . It is also sometimes useful to view the CEF as a function of  $\boldsymbol{x}$ . In this case we can write  $m(\boldsymbol{u}) = \mathbb{E}(y | \boldsymbol{x} = \boldsymbol{u})$ , which is a function of the argument  $\boldsymbol{u}$ . The expression  $\mathbb{E}(y | \boldsymbol{x} = \boldsymbol{u})$  is the conditional expectation of  $y$ , given that we know that the random variable  $\boldsymbol{x}$  equals the specific value  $\boldsymbol{u}$ . However, sometimes in econometrics we take a notational shortcut and use  $\mathbb{E}(y | \boldsymbol{x})$  to refer to this function. Hopefully, the use of  $\mathbb{E}(y | \boldsymbol{x})$  should be apparent from the context.

## 2.6 Continuous Variables

In the previous sections, we implicitly assumed that the conditioning variables are discrete. However, many conditioning variables are continuous. In this section, we take up this case and assume that the variables  $(y, \boldsymbol{x})$  are continuously distributed with a joint density function  $f(y, \boldsymbol{x})$ .

As an example, take  $y = \log(wage)$  and  $x = experience$ , the number of years of potential labor market experience<sup>9</sup>. The contours of their joint density are plotted on the left side of Figure 2.6 for the population of white men with 12 years of education.

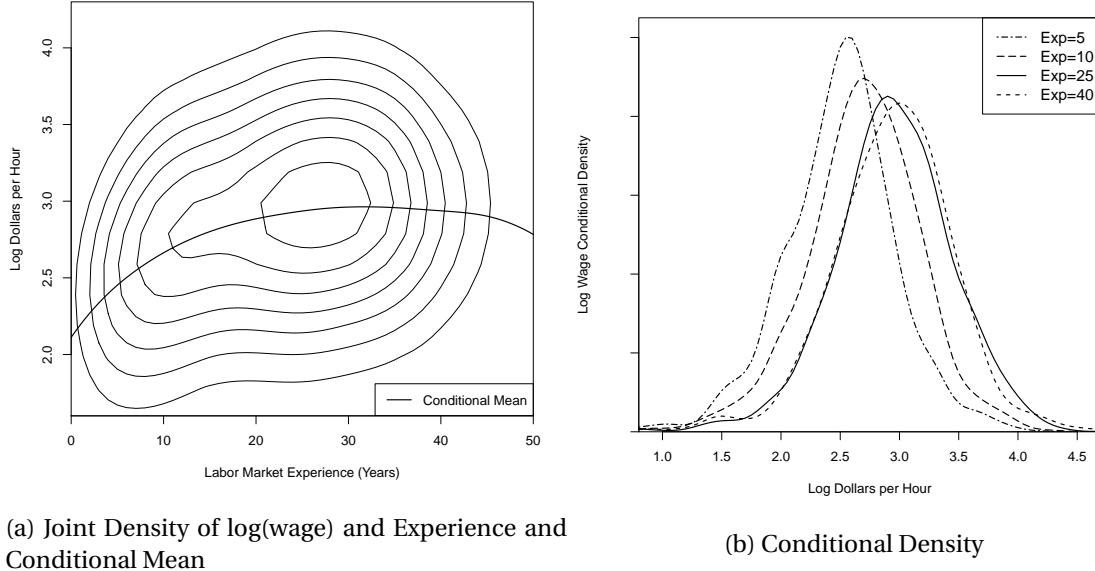


Figure 2.6: White Men with High School Degree

Given the joint density  $f(y, \mathbf{x})$  the variable  $\mathbf{x}$  has the marginal density

$$f_{\mathbf{x}}(\mathbf{x}) = \int_{-\infty}^{\infty} f(y, \mathbf{x}) dy.$$

For any  $\mathbf{x}$  such that  $f_{\mathbf{x}}(\mathbf{x}) > 0$  the conditional density of  $y$  given  $\mathbf{x}$  is defined as

$$f_{y|\mathbf{x}}(y | \mathbf{x}) = \frac{f(y, \mathbf{x})}{f_{\mathbf{x}}(\mathbf{x})}. \quad (2.3)$$

The conditional density is a (renormalized) slice of the joint density  $f(y, \mathbf{x})$  holding  $\mathbf{x}$  fixed. The slice is renormalized (divided by  $f_{\mathbf{x}}(\mathbf{x})$  so that it integrates to one and is thus a density.) We can visualize this by slicing the joint density function at a specific value of  $\mathbf{x}$  parallel with the  $y$ -axis. For example, take the density contours on the left side of Figure 2.6 and slice through the contour plot at a specific value of  $experience$ , and then renormalize the slice so that it is a proper density. This gives us the conditional density of  $\log(wage)$  for white men with 12 years of education and this level of experience. We do this for four levels of  $experience$  (5, 10, 25, and 40 years), and plot these densities on the right side of Figure 2.6. We can see that the distribution of wages shifts to the right and becomes more diffuse as experience increases from 5 to 10 years, and from 10 to 25 years, but there is little change from 25 to 40 years experience.

The CEF of  $y$  given  $\mathbf{x}$  is the mean of the conditional density (2.3)

$$m(\mathbf{x}) = \mathbb{E}(y | \mathbf{x}) = \int_{-\infty}^{\infty} y f_{y|\mathbf{x}}(y | \mathbf{x}) dy. \quad (2.4)$$

Intuitively,  $m(\mathbf{x})$  is the mean of  $y$  for the idealized subpopulation where the conditioning variables are fixed at  $\mathbf{x}$ . This is idealized since  $\mathbf{x}$  is continuously distributed so this subpopulation is infinitely small.

This definition (2.4) is appropriate when the conditional density (2.3) is well defined. However, the conditional mean  $m(\mathbf{x})$  exists quite generally. In Theorem 2.13 in Section 2.34 we show that  $m(\mathbf{x})$  exists so long as  $\mathbb{E}|y| < \infty$ .

<sup>9</sup>Here,  $experience$  is defined as potential labor market experience, equal to  $age - education - 6$

In Figure 2.6 the CEF of  $\log(wage)$  given  $experience$  is plotted as the solid line. We can see that the CEF is a smooth but nonlinear function. The CEF is initially increasing in  $experience$ , flattens out around  $experience = 30$ , and then decreases for high levels of experience.

## 2.7 Law of Iterated Expectations

An extremely useful tool from probability theory is the **law of iterated expectations**. An important special case is the known as the Simple Law.

**Theorem 2.1 Simple Law of Iterated Expectations**

If  $\mathbb{E}|y| < \infty$  then for any random vector  $\mathbf{x}$ ,

$$\mathbb{E}(\mathbb{E}(y | \mathbf{x})) = \mathbb{E}(y).$$

The simple law states that the expectation of the conditional expectation is the unconditional expectation. In other words, the average of the conditional averages is the unconditional average. When  $\mathbf{x}$  is discrete

$$\mathbb{E}(\mathbb{E}(y | \mathbf{x})) = \sum_{j=1}^{\infty} \mathbb{E}(y | \mathbf{x}_j) \mathbb{P}(\mathbf{x} = \mathbf{x}_j)$$

and when  $\mathbf{x}$  is continuous

$$\mathbb{E}(\mathbb{E}(y | \mathbf{x})) = \int_{\mathbb{R}^k} \mathbb{E}(y | \mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}.$$

Going back to our investigation of average log wages for men and women, the simple law states that

$$\begin{aligned} & \mathbb{E}(\log(wage) | sex = man) \mathbb{P}(sex = man) \\ & + \mathbb{E}(\log(wage) | sex = woman) \mathbb{P}(sex = woman) \\ & = \mathbb{E}(\log(wage)). \end{aligned}$$

Or numerically,

$$3.05 \times 0.57 + 2.81 \times 0.43 = 2.95.$$

The general law of iterated expectations allows two sets of conditioning variables.

**Theorem 2.2 Law of Iterated Expectations**

If  $\mathbb{E}|y| < \infty$  then for any random vectors  $\mathbf{x}_1$  and  $\mathbf{x}_2$ ,

$$\mathbb{E}(\mathbb{E}(y | \mathbf{x}_1, \mathbf{x}_2) | \mathbf{x}_1) = \mathbb{E}(y | \mathbf{x}_1).$$

Notice the way the law is applied. The inner expectation conditions on  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , while the outer expectation conditions only on  $\mathbf{x}_1$ . The iterated expectation yields the simple answer  $\mathbb{E}(y | \mathbf{x}_1)$ , the expectation conditional on  $\mathbf{x}_1$  alone. Sometimes we phrase this as: “The smaller information set wins.”

As an example

$$\begin{aligned} & \mathbb{E}(\log(wage) | sex = man, race = white) \mathbb{P}(race = white | sex = man) \\ & + \mathbb{E}(\log(wage) | sex = man, race = black) \mathbb{P}(race = black | sex = man) \\ & + \mathbb{E}(\log(wage) | sex = man, race = other) \mathbb{P}(race = other | sex = man) \\ & = \mathbb{E}(\log(wage) | sex = man) \end{aligned}$$

or numerically

$$3.07 \times 0.84 + 2.86 \times 0.08 + 3.03 \times 0.08 = 3.05.$$

A property of conditional expectations is that when you condition on a random vector  $\mathbf{x}$  you can effectively treat it as if it is constant. For example,  $\mathbb{E}(\mathbf{x} | \mathbf{x}) = \mathbf{x}$  and  $\mathbb{E}(g(\mathbf{x}) | \mathbf{x}) = g(\mathbf{x})$  for any function  $g(\cdot)$ . The general property is known as the Conditioning Theorem.

**Theorem 2.3 Conditioning Theorem**

If  $\mathbb{E}|y| < \infty$  then

$$\mathbb{E}(g(\mathbf{x})y | \mathbf{x}) = g(\mathbf{x})\mathbb{E}(y | \mathbf{x}). \quad (2.5)$$

If in addition  $\mathbb{E}|g(\mathbf{x})y| < \infty$  then

$$\mathbb{E}(g(\mathbf{x})y) = \mathbb{E}(g(\mathbf{x})\mathbb{E}(y | \mathbf{x})). \quad (2.6)$$

The proofs of Theorems 2.1, 2.2 and 2.3 are given in Section 2.36.

## 2.8 CEF Error

The CEF error  $e$  is defined as the difference between  $y$  and the CEF evaluated at the random vector  $\mathbf{x}$ :

$$e = y - m(\mathbf{x}).$$

By construction, this yields the formula

$$y = m(\mathbf{x}) + e. \quad (2.7)$$

In (2.7) it is useful to understand that the error  $e$  is derived from the joint distribution of  $(y, \mathbf{x})$ , and so its properties are derived from this construction.

Many authors in econometrics denote the CEF error using the Greek letter  $\varepsilon$  (epsilon). I do not follow this convention since the error  $e$  is a random variable similar to  $y$  and  $\mathbf{x}$ , and typically use Latin characters for random variables.

A key property of the CEF error is that it has a conditional mean of zero. To see this, by the linearity of expectations, the definition  $m(\mathbf{x}) = \mathbb{E}(y | \mathbf{x})$  and the Conditioning Theorem

$$\begin{aligned} \mathbb{E}(e | \mathbf{x}) &= \mathbb{E}((y - m(\mathbf{x})) | \mathbf{x}) \\ &= \mathbb{E}(y | \mathbf{x}) - \mathbb{E}(m(\mathbf{x}) | \mathbf{x}) \\ &= m(\mathbf{x}) - m(\mathbf{x}) \\ &= 0. \end{aligned}$$

This fact can be combined with the law of iterated expectations to show that the unconditional mean is also zero.

$$\mathbb{E}(e) = \mathbb{E}(\mathbb{E}(e | \mathbf{x})) = \mathbb{E}(0) = 0.$$

We state this and some other results formally.

**Theorem 2.4 Properties of the CEF error**

If  $\mathbb{E}|y| < \infty$  then

1.  $\mathbb{E}(e | \mathbf{x}) = 0$ .

2.  $\mathbb{E}(e) = 0$ .

3. If  $\mathbb{E}|y|^r < \infty$  for  $r \geq 1$  then  $\mathbb{E}|e|^r < \infty$ .

4. For any function  $h(\mathbf{x})$  such that  $\mathbb{E}|h(\mathbf{x})e| < \infty$  then  $\mathbb{E}(h(\mathbf{x})e) = 0$ .

The proof of the third result is deferred to Section 2.36.

The fourth result, whose proof is left to Exercise 2.3, implies that  $e$  is uncorrelated with any function of the regressors.

The equations

$$y = m(\mathbf{x}) + e$$

$$\mathbb{E}(e | \mathbf{x}) = 0$$

together imply that  $m(\mathbf{x})$  is the CEF of  $y$  given  $\mathbf{x}$ . It is important to understand that this is not a restriction. These equations hold true by definition.

The condition  $\mathbb{E}(e | \mathbf{x}) = 0$  is implied by the definition of  $e$  as the difference between  $y$  and the CEF  $m(\mathbf{x})$ . The equation  $\mathbb{E}(e | \mathbf{x}) = 0$  is sometimes called a conditional mean restriction, since the conditional mean of the error  $e$  is restricted to equal zero. The property is also sometimes called **mean independence**, for the conditional mean of  $e$  is 0 and thus independent of  $\mathbf{x}$ . However, it does not imply that the distribution of  $e$  is independent of  $\mathbf{x}$ . Sometimes the assumption “ $e$  is independent of  $\mathbf{x}$ ” is added as a convenient simplification, but it is not generic feature of the conditional mean. Typically and generally,  $e$  and  $\mathbf{x}$  are jointly dependent, even though the conditional mean of  $e$  is zero.

As an example, the contours of the joint density of  $e$  and *experience* are plotted in Figure 2.7 for the same population as Figure 2.6. Notice that the shape of the conditional distribution varies with the level of *experience*.

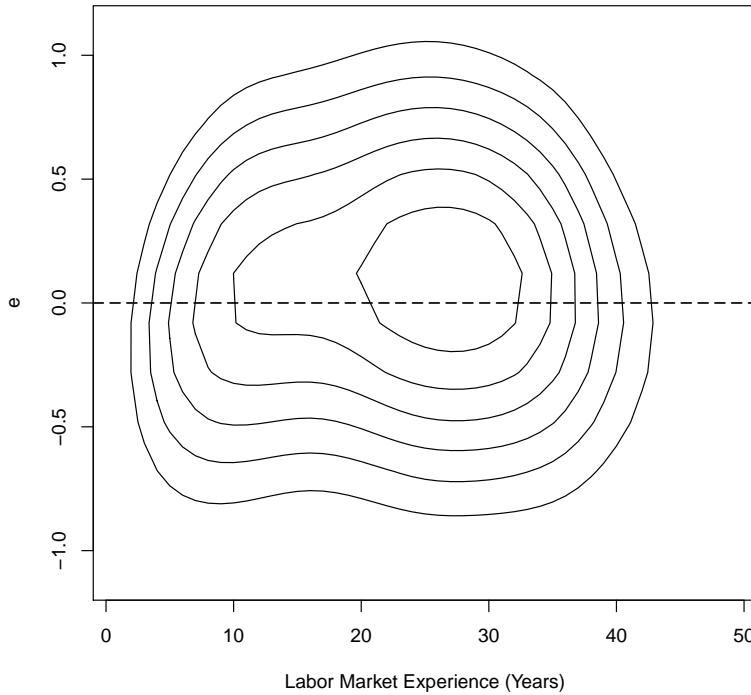


Figure 2.7: Joint Density of Error  $e$  and Experience for white men with High School Education

As a simple example of a case where  $x$  and  $e$  are mean independent yet dependent, let  $e = xe$  where  $x$  and  $e$  are independent  $N(0, 1)$ . Then conditional on  $x$ , the error  $e$  has the distribution  $N(0, x^2)$ . Thus  $\mathbb{E}(e | x) = 0$  and  $e$  is mean independent of  $x$ , yet  $e$  is not fully independent of  $x$ . Mean independence does not imply full independence.

## 2.9 Intercept-Only Model

A special case of the regression model is when there are no regressors  $\mathbf{x}$ . In this case  $m(\mathbf{x}) = \mathbb{E}(y) = \mu$ , the unconditional mean of  $y$ . We can still write an equation for  $y$  in the regression format:

$$\begin{aligned} y &= \mu + e \\ \mathbb{E}(e) &= 0. \end{aligned}$$

This is useful for it unifies the notation.

## 2.10 Regression Variance

An important measure of the dispersion about the CEF function is the unconditional variance of the CEF error  $e$ . We write this as

$$\sigma^2 = \text{var}(e) = \mathbb{E}((e - \mathbb{E}e)^2) = \mathbb{E}(e^2).$$

Theorem 2.4.3 implies the following simple but useful result.

**Theorem 2.5** If  $\mathbb{E}(y^2) < \infty$  then  $\sigma^2 < \infty$ .

We can call  $\sigma^2$  the regression variance or the variance of the regression error. The magnitude of  $\sigma^2$  measures the amount of variation in  $y$  which is not “explained” or accounted for in the conditional mean  $\mathbb{E}(y | \mathbf{x})$ .

The regression variance depends on the regressors  $\mathbf{x}$ . Consider two regressions

$$\begin{aligned} y &= \mathbb{E}(y | \mathbf{x}_1) + e_1 \\ y &= \mathbb{E}(y | \mathbf{x}_1, \mathbf{x}_2) + e_2. \end{aligned}$$

We write the two errors distinctly as  $e_1$  and  $e_2$  as they are different – changing the conditioning information changes the conditional mean and therefore the regression error as well.

In our discussion of iterated expectations, we have seen that by increasing the conditioning set, the conditional expectation reveals greater detail about the distribution of  $y$ . What is the implication for the regression error?

It turns out that there is a simple relationship. We can think of the conditional mean  $\mathbb{E}(y | \mathbf{x})$  as the “explained portion” of  $y$ . The remainder  $e = y - \mathbb{E}(y | \mathbf{x})$  is the “unexplained portion”. The simple relationship we now derive shows that the variance of this unexplained portion decreases when we condition on more variables. This relationship is monotonic in the sense that increasing the amount of information always decreases the variance of the unexplained portion.

**Theorem 2.6** If  $\mathbb{E}(y^2) < \infty$  then

$$\text{var}(y) \geq \text{var}(y - \mathbb{E}(y | \mathbf{x}_1)) \geq \text{var}(y - \mathbb{E}(y | \mathbf{x}_1, \mathbf{x}_2)).$$

Theorem 2.6 says that the variance of the difference between  $y$  and its conditional mean (weakly) decreases whenever an additional variable is added to the conditioning information.

The proof of Theorem 2.6 is given in Section 2.36.

## 2.11 Best Predictor

Suppose that given a realized value of  $\mathbf{x}$ , we want to create a prediction or forecast of  $y$ . We can write any predictor as a function  $g(\mathbf{x})$  of  $\mathbf{x}$ . The prediction error is the realized difference  $y - g(\mathbf{x})$ . A non-stochastic measure of the magnitude of the prediction error is the expectation of its square

$$\mathbb{E}((y - g(\mathbf{x}))^2). \quad (2.8)$$

We can define the best predictor as the function  $g(\mathbf{x})$  which minimizes (2.8). What function is the best predictor? It turns out that the answer is the CEF  $m(\mathbf{x})$ . This holds regardless of the joint distribution of  $(y, \mathbf{x})$ .

To see this, note that the mean squared error of a predictor  $g(\mathbf{x})$  is

$$\begin{aligned}\mathbb{E}((y - g(\mathbf{x}))^2) &= \mathbb{E}((e + m(\mathbf{x}) - g(\mathbf{x}))^2) \\ &= \mathbb{E}(e^2) + 2\mathbb{E}(e(m(\mathbf{x}) - g(\mathbf{x}))) + \mathbb{E}((m(\mathbf{x}) - g(\mathbf{x}))^2) \\ &= \mathbb{E}(e^2) + \mathbb{E}((m(\mathbf{x}) - g(\mathbf{x}))^2) \\ &\geq \mathbb{E}(e^2) \\ &= \mathbb{E}((y - m(\mathbf{x}))^2)\end{aligned}$$

where the first equality makes the substitution  $y = m(\mathbf{x}) + e$  and the third equality uses Theorem 2.4.4. The right-hand-side after the third equality is minimized by setting  $g(\mathbf{x}) = m(\mathbf{x})$ , yielding the inequality in the fourth line. The minimum is finite under the assumption  $\mathbb{E}(y^2) < \infty$  as shown by Theorem 2.5.

We state this formally in the following result.

**Theorem 2.7 Conditional Mean as Best Predictor**

If  $\mathbb{E}(y^2) < \infty$ , then for any predictor  $g(\mathbf{x})$ ,

$$\mathbb{E}((y - g(\mathbf{x}))^2) \geq \mathbb{E}((y - m(\mathbf{x}))^2)$$

where  $m(\mathbf{x}) = \mathbb{E}(y | \mathbf{x})$ .

It may be helpful to consider this result in the context of the intercept-only model

$$y = \mu + e$$

$$\mathbb{E}(e) = 0.$$

Theorem 2.7 shows that the best predictor for  $y$  (in the class of constants) is the unconditional mean  $\mu = \mathbb{E}(y)$ , in the sense that the mean minimizes the mean squared prediction error.

## 2.12 Conditional Variance

While the conditional mean is a good measure of the location of a conditional distribution, it does not provide information about the spread of the distribution. A common measure of the dispersion is the **conditional variance**. We first give the general definition of the conditional variance of a random variable  $w$ .

**Definition 2.1** If  $\mathbb{E}(w^2) < \infty$ , the **conditional variance** of  $w$  given  $\mathbf{x}$  is

$$\text{var}(w | \mathbf{x}) = \mathbb{E}((w - \mathbb{E}(w | \mathbf{x}))^2 | \mathbf{x}).$$

Notice that the conditional variance is the conditional second moment, centered around the conditional first moment. Given this definition, we define the conditional variance of the regression error.

**Definition 2.2** If  $\mathbb{E}(e^2) < \infty$ , the **conditional variance** of the regression error  $e$  is

$$\sigma^2(\mathbf{x}) = \text{var}(e | \mathbf{x}) = \mathbb{E}(e^2 | \mathbf{x}).$$

Generally,  $\sigma^2(\mathbf{x})$  is a non-trivial function of  $\mathbf{x}$  and can take any form subject to the restriction that it is non-negative. One way to think about  $\sigma^2(\mathbf{x})$  is that it is the conditional mean of  $e^2$  given  $\mathbf{x}$ . Notice as well that  $\sigma^2(\mathbf{x}) = \text{var}(y | \mathbf{x})$  so it is equivalently the conditional variance of the dependent variable.

The variance is in a different unit of measurement than the original variable. To convert the variance back to the same unit of measure we define the **conditional standard deviation** as its square root  $\sigma(\mathbf{x}) = \sqrt{\sigma^2(\mathbf{x})}$ .

As an example of how the conditional variance depends on observables, compare the conditional log wage densities for men and women displayed in Figure 2.3. The difference between the densities is not purely a location shift, but is also a difference in spread. Specifically, we can see that the density for men's log wages is somewhat more spread out than that for women, while the density for women's wages is somewhat more peaked. Indeed, the conditional standard deviation for men's wages is 3.05 and that for women is 2.81. So while men have higher average wages, they are also somewhat more dispersed.

The unconditional error variance and the conditional variance are related by the law of iterated expectations

$$\sigma^2 = \mathbb{E}(e^2) = \mathbb{E}(\mathbb{E}(e^2 | \mathbf{x})) = \mathbb{E}(\sigma^2(\mathbf{x})).$$

That is, the unconditional error variance is the average conditional variance.

Given the conditional variance, we can define a rescaled error

$$\varepsilon = \frac{e}{\sigma(\mathbf{x})}. \quad (2.9)$$

We can calculate that since  $\sigma(\mathbf{x})$  is a function of  $\mathbf{x}$

$$\mathbb{E}(\varepsilon | \mathbf{x}) = \mathbb{E}\left(\frac{e}{\sigma(\mathbf{x})} \mid \mathbf{x}\right) = \frac{1}{\sigma(\mathbf{x})} \mathbb{E}(e | \mathbf{x}) = 0$$

and

$$\text{var}(\varepsilon | \mathbf{x}) = \mathbb{E}(\varepsilon^2 | \mathbf{x}) = \mathbb{E}\left(\frac{e^2}{\sigma^2(\mathbf{x})} \mid \mathbf{x}\right) = \frac{1}{\sigma^2(\mathbf{x})} \mathbb{E}(e^2 | \mathbf{x}) = \frac{\sigma^2(\mathbf{x})}{\sigma^2(\mathbf{x})} = 1.$$

Thus  $\varepsilon$  has a conditional mean of zero, and a conditional variance of 1.

Notice that (2.9) can be rewritten as

$$e = \sigma(\mathbf{x})\varepsilon.$$

and substituting this for  $e$  in the CEF equation (2.7), we find that

$$y = m(\mathbf{x}) + \sigma(\mathbf{x})\varepsilon.$$

This is an alternative (mean-variance) representation of the CEF equation.

Many econometric studies focus on the conditional mean  $m(\mathbf{x})$  and either ignore the conditional variance  $\sigma^2(\mathbf{x})$ , treat it as a constant  $\sigma^2(\mathbf{x}) = \sigma^2$ , or treat it as a nuisance parameter (a parameter not of primary interest). This is appropriate when the primary variation in the conditional distribution is in the mean, but can be short-sighted in other cases. Dispersion is relevant to many economic topics, including income and wealth distribution, economic inequality, and price dispersion. Conditional dispersion (variance) can be a fruitful subject for investigation.

The perverse consequences of a narrow-minded focus on the mean has been parodied in a classic joke:

An economist was standing with one foot in a bucket of boiling water and the other foot in a bucket of ice. When asked how he felt, he replied, “On average I feel just fine.”

Clearly, the economist in question ignored variance!

## 2.13 Homoskedasticity and Heteroskedasticity

An important special case obtains when the conditional variance  $\sigma^2(\mathbf{x})$  is a constant and independent of  $\mathbf{x}$ . This is called **homoskedasticity**.

**Definition 2.3** The error is **homoskedastic** if  $\mathbb{E}(e^2 | \mathbf{x}) = \sigma^2$  does not depend on  $\mathbf{x}$ .

In the general case where  $\sigma^2(\mathbf{x})$  depends on  $\mathbf{x}$  we say that the error  $e$  is **heteroskedastic**.

**Definition 2.4** The error is **heteroskedastic** if  $\mathbb{E}(e^2 | \mathbf{x}) = \sigma^2(\mathbf{x})$  depends on  $\mathbf{x}$ .

It is helpful to understand that the concepts homoskedasticity and heteroskedasticity concern the conditional variance, not the unconditional variance. By definition, the unconditional variance  $\sigma^2$  is a constant and independent of the regressors  $\mathbf{x}$ . So when we talk about the variance as a function of the regressors, we are talking about the conditional variance  $\sigma^2(\mathbf{x})$ .

Some older or introductory textbooks describe heteroskedasticity as the case where “the variance of  $e$  varies across observations”. This is a poor and confusing definition. It is more constructive to understand that heteroskedasticity means that the conditional variance  $\sigma^2(\mathbf{x})$  depends on observables.

Older textbooks also tend to describe homoskedasticity as a component of a correct regression specification, and describe heteroskedasticity as an exception or deviance. This description has influenced many generations of economists, but it is unfortunately backwards. The correct view is that heteroskedasticity is generic and “standard”, while homoskedasticity is unusual and exceptional. The default in empirical work should be to assume that the errors are heteroskedastic, not the converse.

In apparent contradiction to the above statement, we will still frequently impose the homoskedasticity assumption when making theoretical investigations into the properties of estimation and inference methods. The reason is that in many cases homoskedasticity greatly simplifies the theoretical calculations, and it is therefore quite advantageous for teaching and learning. It should always be remembered, however, that homoskedasticity is never imposed because it is believed to be a correct feature of an empirical model, but rather because of its simplicity.

### Heteroskedastic or Heteroscedastic?

The spelling of the words *homoskedastic* and *heteroskedastic* have been somewhat controversial. Early econometrics textbooks were split, with some using a “c” as in *heteroscedastic* and some “k” as in *heteroskedastic*. McCulloch (1985) pointed out that the word is derived from Greek roots. *ομοιος* means “same”. *ετερο* means “other” or “different”. *σκεδαννυμι* means “to scatter”. Since the proper transliteration of the Greek letter  $\kappa$  in *σκεδαννυμι* is “k”, this implies that the correct English spelling of the two words is with a “k” as in *homoskedastic* and *heteroskedastic*.

## 2.14 Regression Derivative

One way to interpret the CEF  $m(\mathbf{x}) = \mathbb{E}(y | \mathbf{x})$  is in terms of how marginal changes in the regressors  $\mathbf{x}$  imply changes in the conditional mean of the response variable  $y$ . It is typical to consider marginal changes in a single regressor, say  $x_1$ , holding the remainder fixed. When a regressor  $x_1$  is continuously distributed, we define the marginal effect of a change in  $x_1$ , holding the variables  $x_2, \dots, x_k$  fixed, as the partial derivative of the CEF

$$\frac{\partial}{\partial x_1} m(x_1, \dots, x_k).$$

When  $x_1$  is discrete we define the marginal effect as a discrete difference. For example, if  $x_1$  is binary, then the marginal effect of  $x_1$  on the CEF is

$$m(1, x_2, \dots, x_k) - m(0, x_2, \dots, x_k).$$

We can unify the continuous and discrete cases with the notation

$$\nabla_1 m(\mathbf{x}) = \begin{cases} \frac{\partial}{\partial x_1} m(x_1, \dots, x_k), & \text{if } x_1 \text{ is continuous} \\ m(1, x_2, \dots, x_k) - m(0, x_2, \dots, x_k), & \text{if } x_1 \text{ is binary.} \end{cases}$$

Collecting the  $k$  effects into one  $k \times 1$  vector, we define the **regression derivative** with respect to  $\mathbf{x}$ :

$$\nabla m(\mathbf{x}) = \begin{bmatrix} \nabla_1 m(\mathbf{x}) \\ \nabla_2 m(\mathbf{x}) \\ \vdots \\ \nabla_k m(\mathbf{x}) \end{bmatrix}.$$

When all elements of  $\mathbf{x}$  are continuous, then we have the simplification  $\nabla m(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} m(\mathbf{x})$ , the vector of partial derivatives.

There are two important points to remember concerning our definition of the regression derivative.

First, the effect of each variable is calculated holding the other variables constant. This is the **ceteris paribus** concept commonly used in economics. But in the case of a regression derivative, the conditional mean does not literally hold *all else* constant. It only holds constant the variables included in the conditional mean. This means that the regression derivative depends on which regressors are included. For example, in a regression of wages on education, experience, race and sex, the regression derivative with respect to education shows the marginal effect of education on mean wages, holding constant experience, race and sex. But it does not hold constant an individual's unobservable characteristics (such as ability), nor variables not included in the regression (such as the quality of education).

Second, the regression derivative is the change in the conditional expectation of  $y$ , not the change in the actual value of  $y$  for an individual. It is tempting to think of the regression derivative as the change in the actual value of  $y$ , but this is not a correct interpretation. The regression derivative  $\nabla m(\mathbf{x})$  is the change in the actual value of  $y$  only if the error  $e$  is unaffected by the change in the regressor  $\mathbf{x}$ . We return to a discussion of causal effects in Section 2.30.

## 2.15 Linear CEF

An important special case is when the CEF  $m(\mathbf{x}) = \mathbb{E}(y | \mathbf{x})$  is linear in  $\mathbf{x}$ . In this case we can write the mean equation as

$$m(\mathbf{x}) = x_1\beta_1 + x_2\beta_2 + \cdots + x_k\beta_k + \beta_{k+1}.$$

Notationally it is convenient to write this as a simple function of the vector  $\mathbf{x}$ . An easy way to do so is to augment the regressor vector  $\mathbf{x}$  by listing the number “1” as an element. We call this the “constant” and the corresponding coefficient is called the “intercept”. Equivalently, specify that the final element<sup>10</sup> of the vector  $\mathbf{x}$  is  $x_k = 1$ . Thus (2.2) has been redefined as the  $k \times 1$  vector

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{k-1} \\ 1 \end{pmatrix}. \quad (2.10)$$

With this redefinition, the CEF is

$$m(\mathbf{x}) = x_1\beta_1 + x_2\beta_2 + \cdots + \beta_k = \mathbf{x}'\boldsymbol{\beta} \quad (2.11)$$

where

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

is a  $k \times 1$  coefficient vector. This is the **linear CEF model**. It is also often called the **linear regression model**, or the regression of  $y$  on  $\mathbf{x}$ .

In the linear CEF model, the regression derivative is simply the coefficient vector. That is

$$\nabla m(\mathbf{x}) = \boldsymbol{\beta}.$$

This is one of the appealing features of the linear CEF model. The coefficients have simple and natural interpretations as the marginal effects of changing one variable, holding the others constant.

### Linear CEF Model

$$y = \mathbf{x}'\boldsymbol{\beta} + e$$

$$\mathbb{E}(e | \mathbf{x}) = 0$$

If in addition the error is homoskedastic, we call this the homoskedastic linear CEF model.

---

<sup>10</sup>The order doesn't matter. It could be any element.

### Homoskedastic Linear CEF Model

$$\begin{aligned}y &= \mathbf{x}'\boldsymbol{\beta} + e \\ \mathbb{E}(e | \mathbf{x}) &= 0 \\ \mathbb{E}(e^2 | \mathbf{x}) &= \sigma^2\end{aligned}$$

## 2.16 Linear CEF with Nonlinear Effects

The linear CEF model of the previous section is less restrictive than it might appear, as we can include as regressors nonlinear transformations of the original variables. In this sense, the linear CEF framework is flexible and can capture many nonlinear effects.

For example, suppose we have two scalar variables  $x_1$  and  $x_2$ . The CEF could take the quadratic form

$$m(x_1, x_2) = x_1\beta_1 + x_2\beta_2 + x_1^2\beta_3 + x_2^2\beta_4 + x_1x_2\beta_5 + \beta_6. \quad (2.12)$$

This equation is quadratic in the regressors  $(x_1, x_2)$  yet linear in the coefficients  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_6)'$ . We will descriptively call (2.12) a quadratic CEF, and yet (2.12) is also a linear CEF in the sense of being linear in the coefficients. The key is to understand that (2.12) is quadratic in the variables  $(x_1, x_2)$  yet linear in the coefficients  $\boldsymbol{\beta}$ .

To simplify the expression, we define the transformations  $x_3 = x_1^2$ ,  $x_4 = x_2^2$ ,  $x_5 = x_1x_2$ , and  $x_6 = 1$ , and redefine the regressor vector as  $\mathbf{x} = (x_1, \dots, x_6)'$ . With this redefinition,

$$m(x_1, x_2) = \mathbf{x}'\boldsymbol{\beta}$$

which is linear in  $\boldsymbol{\beta}$ . For most econometric purposes (estimation and inference on  $\boldsymbol{\beta}$ ) the linearity in  $\boldsymbol{\beta}$  is all that is important.

An exception is in the analysis of regression derivatives. In nonlinear equations such as (2.12), the regression derivative should be defined with respect to the original variables, not with respect to the transformed variables. Thus

$$\begin{aligned}\frac{\partial}{\partial x_1} m(x_1, x_2) &= \beta_1 + 2x_1\beta_3 + x_2\beta_5 \\ \frac{\partial}{\partial x_2} m(x_1, x_2) &= \beta_2 + 2x_2\beta_4 + x_1\beta_5.\end{aligned}$$

We see that in the model (2.12), the regression derivatives are not a simple coefficient, but are functions of several coefficients plus the levels of  $(x_1, x_2)$ . Consequently it is difficult to interpret the coefficients individually. It is more useful to interpret them as a group.

We typically call  $\beta_5$  the **interaction effect**. Notice that it appears in both regression derivative equations, and has a symmetric interpretation in each. If  $\beta_5 > 0$  then the regression derivative with respect to  $x_1$  is increasing in the level of  $x_2$  (and the regression derivative with respect to  $x_2$  is increasing in the level of  $x_1$ ), while if  $\beta_5 < 0$  the reverse is true.

## 2.17 Linear CEF with Dummy Variables

When all regressors take a finite set of values, it turns out the CEF can be written as a linear function of regressors.

This simplest example is a **binary** variable, which takes only two distinct values. For example, in most data sets the variable *sex* takes only the values *man* and *woman* (or male and female). Binary variables

are extremely common in econometric applications, and are alternatively called **dummy variables** or **indicator variables**.

Consider the simple case of a single binary regressor. In this case, the conditional mean can only take two distinct values. For example,

$$\mathbb{E}(y | \text{sex}) = \begin{cases} \mu_0 & \text{if } \text{sex}=\text{man} \\ \mu_1 & \text{if } \text{sex}=\text{woman} \end{cases}.$$

To facilitate a mathematical treatment, we typically record dummy variables with the values {0, 1}. For example

$$x_1 = \begin{cases} 0 & \text{if } \text{sex}=\text{man} \\ 1 & \text{if } \text{sex}=\text{woman} \end{cases}. \quad (2.13)$$

Given this notation we can write the conditional mean as a linear function of the dummy variable  $x_1$ , that is

$$\mathbb{E}(y | x_1) = \beta_1 x_1 + \beta_2$$

where  $\beta_1 = \mu_1 - \mu_0$  and  $\beta_2 = \mu_0$ . In this simple regression equation the intercept  $\beta_2$  is equal to the conditional mean of  $y$  for the  $x_1 = 0$  subpopulation (men) and the slope  $\beta_1$  is equal to the difference in the conditional means between the two subpopulations.

Equivalently, we could have defined  $x_1$  as

$$x_1 = \begin{cases} 1 & \text{if } \text{sex}=\text{man} \\ 0 & \text{if } \text{sex}=\text{woman} \end{cases}. \quad (2.14)$$

In this case, the regression intercept is the mean for women (rather than for men) and the regression slope has switched signs. The two regressions are equivalent but the interpretation of the coefficients has changed. Therefore it is always important to understand the precise definitions of the variables, and illuminating labels are helpful. For example, labelling  $x_1$  as “sex” does not help distinguish between definitions (2.13) and (2.14). Instead, it is better to label  $x_1$  as “women” or “female” if definition (2.13) is used, or as “men” or “male” if (2.14) is used.

Now suppose we have two dummy variables  $x_1$  and  $x_2$ . For example,  $x_2 = 1$  if the person is married, else  $x_2 = 0$ . The conditional mean given  $x_1$  and  $x_2$  takes at most four possible values:

$$\mathbb{E}(y | x_1, x_2) = \begin{cases} \mu_{00} & \text{if } x_1 = 0 \text{ and } x_2 = 0 \quad (\text{unmarried men}) \\ \mu_{01} & \text{if } x_1 = 0 \text{ and } x_2 = 1 \quad (\text{married men}) \\ \mu_{10} & \text{if } x_1 = 1 \text{ and } x_2 = 0 \quad (\text{unmarried women}) \\ \mu_{11} & \text{if } x_1 = 1 \text{ and } x_2 = 1 \quad (\text{married women}) \end{cases}.$$

In this case we can write the conditional mean as a linear function of  $x_1$ ,  $x_2$  and their product  $x_1 x_2$ :

$$\mathbb{E}(y | x_1, x_2) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4$$

where  $\beta_1 = \mu_{10} - \mu_{00}$ ,  $\beta_2 = \mu_{01} - \mu_{00}$ ,  $\beta_3 = \mu_{11} - \mu_{10} - \mu_{01} + \mu_{00}$ , and  $\beta_4 = \mu_{00}$ .

We can view the coefficient  $\beta_1$  as the effect of sex on expected log wages for unmarried wage earners, the coefficient  $\beta_2$  as the effect of marriage on expected log wages for men wage earners, and the coefficient  $\beta_3$  as the difference between the effects of marriage on expected log wages among women and among men. Alternatively, it can also be interpreted as the difference between the effects of sex on expected log wages among married and non-married wage earners. Both interpretations are equally valid. We often describe  $\beta_3$  as measuring the **interaction** between the two dummy variables, or the **interaction effect**, and describe  $\beta_3 = 0$  as the case when the interaction effect is zero.

In this setting we can see that the CEF is linear in the three variables  $(x_1, x_2, x_1 x_2)$ . Thus to put the model in the framework of Section 2.15, we would define the regressor  $x_3 = x_1 x_2$  and the regressor vector

as

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix}.$$

So even though we started with only 2 dummy variables, the number of regressors (including the intercept) is 4.

If there are 3 dummy variables  $x_1, x_2, x_3$ , then  $\mathbb{E}(y | x_1, x_2, x_3)$  takes at most  $2^3 = 8$  distinct values and can be written as the linear function

$$\mathbb{E}(y | x_1, x_2, x_3) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1 x_2 x_3 + \beta_8$$

which has eight regressors including the intercept.

In general, if there are  $p$  dummy variables  $x_1, \dots, x_p$  then the CEF  $\mathbb{E}(y | x_1, x_2, \dots, x_p)$  takes at most  $2^p$  distinct values, and can be written as a linear function of the  $2^p$  regressors including  $x_1, x_2, \dots, x_p$  and all cross-products. This might be excessive in practice if  $p$  is modestly large. In the next section we will discuss projection approximations which yield more parsimonious parameterizations.

We started this section by saying that the conditional mean is linear whenever all regressors take only a finite number of possible values. How can we see this? Take a **categorical** variable, such as *race*. For example, we earlier divided race into three categories. We can record categorical variables using numbers to indicate each category, for example

$$x_3 = \begin{cases} 1 & \text{if } \textit{white} \\ 2 & \text{if } \textit{black} \\ 3 & \text{if } \textit{other} \end{cases}.$$

When doing so, the values of  $x_3$  have no meaning in terms of magnitude, they simply indicate the relevant category.

When the regressor is categorical the conditional mean of  $y$  given  $x_3$  takes a distinct value for each possibility:

$$\mathbb{E}(y | x_3) = \begin{cases} \mu_1 & \text{if } x_3 = 1 \\ \mu_2 & \text{if } x_3 = 2 \\ \mu_3 & \text{if } x_3 = 3 \end{cases}.$$

This is not a linear function of  $x_3$  itself, but it can be made a linear function by constructing dummy variables for two of the three categories. For example

$$x_4 = \begin{cases} 1 & \text{if } \textit{black} \\ 0 & \text{if } \textit{not black} \end{cases}$$

$$x_5 = \begin{cases} 1 & \text{if } \textit{other} \\ 0 & \text{if } \textit{not other} \end{cases}.$$

In this case, the categorical variable  $x_3$  is equivalent to the pair of dummy variables  $(x_4, x_5)$ . The explicit relationship is

$$x_3 = \begin{cases} 1 & \text{if } x_4 = 0 \text{ and } x_5 = 0 \\ 2 & \text{if } x_4 = 1 \text{ and } x_5 = 0 \\ 3 & \text{if } x_4 = 0 \text{ and } x_5 = 1 \end{cases}.$$

Given these transformations, we can write the conditional mean of  $y$  as a linear function of  $x_4$  and  $x_5$

$$\mathbb{E}(y | x_3) = \mathbb{E}(y | x_4, x_5) = \beta_1 x_4 + \beta_2 x_5 + \beta_3.$$

We can write the CEF as either  $\mathbb{E}(y | x_3)$  or  $\mathbb{E}(y | x_4, x_5)$  (they are equivalent), but it is only linear as a function of  $x_4$  and  $x_5$ .

This setting is similar to the case of two dummy variables, with the difference that we have not included the interaction term  $x_4 x_5$ . This is because the event  $\{x_4 = 1 \text{ and } x_5 = 1\}$  is empty by construction, so  $x_4 x_5 = 0$  by definition.

## 2.18 Best Linear Predictor

While the conditional mean  $m(\mathbf{x}) = \mathbb{E}(y | \mathbf{x})$  is the best predictor of  $y$  among all functions of  $\mathbf{x}$ , its functional form is typically unknown. In particular, the linear CEF model is empirically unlikely to be accurate unless  $\mathbf{x}$  is discrete and low-dimensional so all interactions are included. Consequently in most cases it is more realistic to view the linear specification (2.11) as an approximation. In this section we derive a specific approximation with a simple interpretation.

Theorem 2.7 showed that the conditional mean  $m(\mathbf{x})$  is the best predictor in the sense that it has the lowest mean squared error among all predictors. By extension, we can define an approximation to the CEF by the linear function with the lowest mean squared error among all linear predictors.

For this derivation we require the following regularity condition.

**Assumption 2.1**

1.  $\mathbb{E}(y^2) < \infty$ .
2.  $\mathbb{E}\|\mathbf{x}\|^2 < \infty$ .
3.  $\mathbf{Q}_{\mathbf{x}\mathbf{x}} = \mathbb{E}(\mathbf{x}\mathbf{x}')$  is positive definite.

In Assumption 2.1.2 we use the notation  $\|\mathbf{x}\| = (\mathbf{x}'\mathbf{x})^{1/2}$  to denote the Euclidean length of the vector  $\mathbf{x}$ .

The first two parts of Assumption 2.1 imply that the variables  $y$  and  $\mathbf{x}$  have finite means, variances, and covariances. The third part of the assumption is more technical, and its role will become apparent shortly. It is equivalent to imposing that the columns of the matrix  $\mathbf{Q}_{\mathbf{x}\mathbf{x}} = \mathbb{E}(\mathbf{x}\mathbf{x}')$  are linearly independent, or that the matrix is invertible.

A linear predictor for  $y$  is a function of the form  $\mathbf{x}'\boldsymbol{\beta}$  for some  $\boldsymbol{\beta} \in \mathbb{R}^k$ . The mean squared prediction error is

$$S(\boldsymbol{\beta}) = \mathbb{E}((y - \mathbf{x}'\boldsymbol{\beta})^2). \quad (2.15)$$

The **best linear predictor** of  $y$  given  $\mathbf{x}$ , written  $\mathcal{P}(y | \mathbf{x})$ , is found by selecting the vector  $\boldsymbol{\beta}$  to minimize  $S(\boldsymbol{\beta})$ .

**Definition 2.5** The **Best Linear Predictor** of  $y$  given  $\mathbf{x}$  is

$$\mathcal{P}(y | \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$$

where  $\boldsymbol{\beta}$  minimizes the mean squared prediction error

$$S(\boldsymbol{\beta}) = \mathbb{E}((y - \mathbf{x}'\boldsymbol{\beta})^2).$$

The minimizer

$$\boldsymbol{\beta} = \underset{\boldsymbol{b} \in \mathbb{R}^k}{\operatorname{argmin}} S(\boldsymbol{b}) \quad (2.16)$$

is called the **Linear Projection Coefficient**.

We now calculate an explicit expression for its value. The mean squared prediction error (2.15) can be written out as a quadratic function of  $\boldsymbol{\beta}$ :

$$S(\boldsymbol{\beta}) = \mathbb{E}(y^2) - 2\boldsymbol{\beta}'\mathbb{E}(\mathbf{x}y) + \boldsymbol{\beta}'\mathbb{E}(\mathbf{x}\mathbf{x}')\boldsymbol{\beta}. \quad (2.17)$$

The quadratic structure of  $S(\beta)$  means that we can solve explicitly for the minimizer. The first-order condition for minimization (from Appendix A.20) is

$$\mathbf{0} = \frac{\partial}{\partial \beta} S(\beta) = -2\mathbb{E}(\mathbf{x}\mathbf{y}) + 2\mathbb{E}(\mathbf{x}\mathbf{x}')\beta. \quad (2.18)$$

Rewriting (2.18) as

$$2\mathbb{E}(\mathbf{x}\mathbf{y}) = 2\mathbb{E}(\mathbf{x}\mathbf{x}')\beta$$

and dividing by 2, this equation takes the form

$$\mathbf{Q}_{xy} = \mathbf{Q}_{xx}\beta \quad (2.19)$$

where  $\mathbf{Q}_{xy} = \mathbb{E}(\mathbf{x}\mathbf{y})$  is  $k \times 1$  and  $\mathbf{Q}_{xx} = \mathbb{E}(\mathbf{x}\mathbf{x}')$  is  $k \times k$ . The solution is found by inverting the matrix  $\mathbf{Q}_{xx}$ , and is written

$$\beta = \mathbf{Q}_{xx}^{-1}\mathbf{Q}_{xy}$$

or

$$\beta = (\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-1}\mathbb{E}(\mathbf{x}\mathbf{y}). \quad (2.20)$$

It is worth taking the time to understand the notation involved in the expression (2.20).  $\mathbf{Q}_{xx}$  is a  $k \times k$  matrix and  $\mathbf{Q}_{xy}$  is a  $k \times 1$  column vector. Therefore, alternative expressions such as  $\frac{\mathbb{E}(\mathbf{x}\mathbf{y})}{\mathbb{E}(\mathbf{x}\mathbf{x}')}$  or  $\mathbb{E}(\mathbf{x}\mathbf{y})(\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-1}$  are incoherent and incorrect. We also can now see the role of Assumption 2.1.3. It is equivalent to assuming that  $\mathbf{Q}_{xx}$  has an inverse  $\mathbf{Q}_{xx}^{-1}$  which is necessary for the normal equations (2.19) to have a solution or equivalently for (2.20) to be uniquely defined. In the absence of Assumption 2.1.3 there could be multiple solutions to the equation (2.19).

We now have an explicit expression for the best linear predictor:

$$\mathcal{P}(y | \mathbf{x}) = \mathbf{x}'(\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-1}\mathbb{E}(\mathbf{x}\mathbf{y}).$$

This expression is also referred to as the **linear projection** of  $y$  on  $\mathbf{x}$ .

The **projection error** is

$$e = y - \mathbf{x}'\beta. \quad (2.21)$$

This equals the error (2.7) from the regression equation when (and only when) the conditional mean is linear in  $\mathbf{x}$ , otherwise they are distinct.

Rewriting, we obtain a decomposition of  $y$  into linear predictor and error

$$y = \mathbf{x}'\beta + e. \quad (2.22)$$

In general we call equation (2.22) or  $\mathbf{x}'\beta$  the best linear predictor of  $y$  given  $\mathbf{x}$ , or the linear projection of  $y$  on  $\mathbf{x}$ . Equation (2.22) is also often called the **regression** of  $y$  on  $\mathbf{x}$  but this can sometimes be confusing as economists use the term *regression* in many contexts. (Recall that we said in Section 2.15 that the linear CEF model is also called the linear regression model.)

An important property of the projection error  $e$  is

$$\mathbb{E}(\mathbf{x}e) = \mathbf{0}. \quad (2.23)$$

To see this, using the definitions (2.21) and (2.20) and the matrix properties  $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$  and  $\mathbf{I}\mathbf{a} = \mathbf{a}$ ,

$$\begin{aligned} \mathbb{E}(\mathbf{x}e) &= \mathbb{E}(\mathbf{x}(y - \mathbf{x}'\beta)) \\ &= \mathbb{E}(\mathbf{x}\mathbf{y}) - \mathbb{E}(\mathbf{x}\mathbf{x}')(\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-1}\mathbb{E}(\mathbf{x}\mathbf{y}) \\ &= \mathbf{0} \end{aligned} \quad (2.24)$$

as claimed.

Equation (2.23) is a set of  $k$  equations, one for each regressor. In other words, (2.23) is equivalent to

$$\mathbb{E}(x_j e) = 0 \quad (2.25)$$

for  $j = 1, \dots, k$ . As in (2.10), the regressor vector  $\mathbf{x}$  typically contains a constant, e.g.  $x_k = 1$ . In this case (2.25) for  $j = k$  is the same as

$$\mathbb{E}(e) = 0. \quad (2.26)$$

Thus the projection error has a mean of zero when the regressor vector contains a constant. (When  $\mathbf{x}$  does not have a constant, (2.26) is not guaranteed. As it is desirable for  $e$  to have a zero mean, this is a good reason to always include a constant in any regression model.)

It is also useful to observe that since  $\text{cov}(x_j, e) = \mathbb{E}(x_j e) - \mathbb{E}(x_j)\mathbb{E}(e)$ , then (2.25)-(2.26) together imply that the variables  $x_j$  and  $e$  are uncorrelated.

This completes the derivation of the model. We summarize some of the most important properties.

**Theorem 2.8 Properties of Linear Projection Model**

Under Assumption 2.1,

1. The moments  $\mathbb{E}(\mathbf{x}\mathbf{x}')$  and  $\mathbb{E}(\mathbf{x}y)$  exist with finite elements.
2. The Linear Projection Coefficient (2.16) exists, is unique, and equals

$$\boldsymbol{\beta} = (\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-1} \mathbb{E}(\mathbf{x}y).$$

3. The best linear predictor of  $y$  given  $\mathbf{x}$  is

$$\mathcal{P}(y | \mathbf{x}) = \mathbf{x}' (\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-1} \mathbb{E}(\mathbf{x}y).$$

4. The projection error  $e = y - \mathbf{x}'\boldsymbol{\beta}$  exists and satisfies

$$\mathbb{E}(e^2) < \infty$$

and

$$\mathbb{E}(\mathbf{x}e) = \mathbf{0}.$$

5. If  $\mathbf{x}$  contains a constant, then

$$\mathbb{E}(e) = 0.$$

6. If  $\mathbb{E}|y|^r < \infty$  and  $\mathbb{E}\|\mathbf{x}\|^r < \infty$  for  $r \geq 2$  then  $\mathbb{E}|e|^r < \infty$ .

A complete proof of Theorem 2.8 is given in Section 2.36.

It is useful to reflect on the generality of Theorem 2.8. The only restriction is Assumption 2.1. Thus for any random variables  $(y, \mathbf{x})$  with finite variances we can define a linear equation (2.22) with the properties listed in Theorem 2.8. Stronger assumptions (such as the linear CEF model) are not necessary. In this sense the linear model (2.22) exists quite generally. However, it is important not to misinterpret the generality of this statement. The linear equation (2.22) is defined as the best linear predictor. It is not necessarily a conditional mean, nor a parameter of a structural or causal economic model.

### Linear Projection Model

$$\begin{aligned}
 y &= \mathbf{x}' \boldsymbol{\beta} + e. \\
 \mathbb{E}(\mathbf{x}e) &= \mathbf{0} \\
 \boldsymbol{\beta} &= (\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-1} \mathbb{E}(\mathbf{x}y)
 \end{aligned}$$

### Invertibility and Identification

The linear projection coefficient  $\boldsymbol{\beta} = (\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-1} \mathbb{E}(\mathbf{x}y)$  exists and is unique as long as the  $k \times k$  matrix  $\mathbf{Q}_{xx} = \mathbb{E}(\mathbf{x}\mathbf{x}')$  is invertible. The matrix  $\mathbf{Q}_{xx}$  is sometimes called the **design matrix**, as in experimental settings the researcher is able to control  $\mathbf{Q}_{xx}$  by manipulating the distribution of the regressors  $\mathbf{x}$ .

Observe that for any non-zero  $\boldsymbol{\alpha} \in \mathbb{R}^k$ ,

$$\boldsymbol{\alpha}' \mathbf{Q}_{xx} \boldsymbol{\alpha} = \mathbb{E}(\boldsymbol{\alpha}' \mathbf{x} \mathbf{x}' \boldsymbol{\alpha}) = \mathbb{E}(\boldsymbol{\alpha}' \mathbf{x})^2 \geq 0$$

so  $\mathbf{Q}_{xx}$  by construction is positive semi-definite, conventionally written as  $\mathbf{Q}_{xx} \geq 0$ . The assumption that it is positive definite means that this is a strict inequality,  $\mathbb{E}(\boldsymbol{\alpha}' \mathbf{x})^2 > 0$ . This is conventionally written as  $\mathbf{Q}_{xx} > 0$ . This condition means that there is no non-zero vector  $\boldsymbol{\alpha}$  such that  $\boldsymbol{\alpha}' \mathbf{x} = 0$  identically. Positive definite matrices are invertible. Thus when  $\mathbf{Q}_{xx} > 0$  then  $\boldsymbol{\beta} = (\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-1} \mathbb{E}(\mathbf{x}y)$  exists and is uniquely defined. In other words, if we can exclude the possibility that a linear function of  $\mathbf{x}$  is degenerate, then  $\boldsymbol{\beta}$  is uniquely defined.

Theorem 2.5 shows that the linear projection coefficient  $\boldsymbol{\beta}$  is **identified** (uniquely determined) under Assumption 2.1. The key is invertibility of  $\mathbf{Q}_{xx}$ . Otherwise, there is no unique solution to the equation

$$\mathbf{Q}_{xx} \boldsymbol{\beta} = \mathbf{Q}_{xy}. \quad (2.27)$$

When  $\mathbf{Q}_{xx}$  is not invertible there are multiple solutions to (2.27). In this case the coefficient  $\boldsymbol{\beta}$  is **not identified** as it does not have a unique value.

### Minimization

The mean squared prediction error (2.17) is a function with vector argument of the form

$$f(\mathbf{x}) = a - 2\mathbf{b}'\mathbf{x} + \mathbf{x}'\mathbf{C}\mathbf{x}$$

where  $\mathbf{C} > 0$ . For any function of this form, the unique minimizer is

$$\mathbf{x} = \mathbf{C}^{-1}\mathbf{b}. \quad (2.28)$$

To see that this is the unique minimizer we present two proofs. The first uses matrix calculus. From Appendix A.20

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{b}'\mathbf{x}) = \mathbf{b} \quad (2.29)$$

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{x}'\mathbf{C}\mathbf{x}) = 2\mathbf{C}\mathbf{x} \quad (2.30)$$

$$\frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}'} (\mathbf{x}'\mathbf{C}\mathbf{x}) = 2\mathbf{C}. \quad (2.31)$$

Using (2.29) and (2.30), we find

$$\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) = -2\mathbf{b} + 2\mathbf{C}\mathbf{x}.$$

The first-order condition for minimization sets this derivative equal to zero. Thus the solution satisfies  $-2\mathbf{b} + 2\mathbf{C}\mathbf{x} = \mathbf{0}$ . Solving for  $\mathbf{x}$  we find (2.28). Using (2.31) we also find

$$\frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}'} f(\mathbf{x}) = 2\mathbf{C} > 0$$

which is the second-order condition for minimization. This shows that (2.28) is the unique minimizer of  $f(\mathbf{x})$ .

Our second proof is algebraic. Re-write  $f(\mathbf{x})$  as

$$f(\mathbf{x}) = (a - \mathbf{b}'\mathbf{C}^{-1}\mathbf{b}) + (\mathbf{x} - \mathbf{C}^{-1}\mathbf{b})' \mathbf{C} (\mathbf{x} - \mathbf{C}^{-1}\mathbf{b}).$$

The first term does not depend on  $\mathbf{x}$  so does not affect the minimizer. The second term is a quadratic form in a positive definite matrix. This means that for any non-zero  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\alpha}'\mathbf{C}\boldsymbol{\alpha} > 0$ . Thus for  $\mathbf{x} \neq \mathbf{C}^{-1}\mathbf{b}$ , the second-term is strictly positive, yet for  $\mathbf{x} = \mathbf{C}^{-1}\mathbf{b}$  this term equals zero. It is therefore minimized at  $\mathbf{x} = \mathbf{C}^{-1}\mathbf{b}$  as claimed.

## 2.19 Illustrations of Best Linear Predictor

We illustrate the best linear predictor (projection) using three log wage equations introduced in earlier sections.

For our first example, we consider a model with the two dummy variables for sex and race similar to Table 2.1. As we learned in Section 2.17, the entries in this table can be equivalently expressed by a linear CEF. For simplicity, let's consider the CEF of  $\log(wage)$  as a function of *Black* and *Female*.

$$\mathbb{E}(\log(wage) | Black, Female) = -0.20Black - 0.24Female + 0.10Black \times Female + 3.06. \quad (2.32)$$

This is a CEF as the variables are binary and all interactions are included.

Now consider a simpler model omitting the interaction effect. This is the linear projection on the variables *Black* and *Female*

$$\mathcal{P}(\log(wage) | Black, Female) = -0.15Black - 0.23Female + 3.06. \quad (2.33)$$

What is the difference? The full CEF (2.32) shows that the race gap is differentiated by sex: it is 20% for black men (relative to non-black men) and 10% for black women (relative to non-black women). The projection model (2.33) simplifies this analysis, calculating an average 15% wage gap for blacks, ignoring the role of sex. Notice that this is despite the fact that the sex variable is included in (2.33).

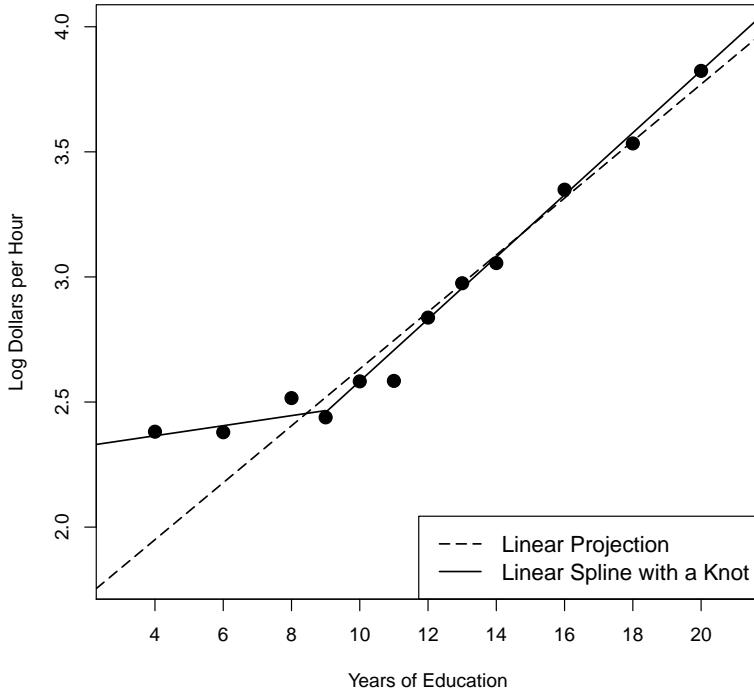


Figure 2.8: Projections of  $\log(wage)$  onto Education

For our second example we consider the CEF of log wages as a function of years of education for white men which was illustrated in Figure 2.5 and is repeated in Figure 2.8. Superimposed on the figure are two projections. The first (given by the dashed line) is the linear projection of log wages on years of education

$$\mathcal{P}(\log(wage) | Education) = 0.11Education + 1.5.$$

This simple equation indicates an average 11% increase in wages for every year of education. An inspection of the Figure shows that this approximation works well for  $education \geq 9$ , but under-predicts for individuals with lower levels of education. To correct this imbalance we use a linear spline equation which allows different rates of return above and below 9 years of education:

$$\begin{aligned} \mathcal{P}(\log(wage) | Education, (Education - 9) \times \mathbf{1}(Education > 9)) \\ = 0.02Education + 0.10 \times (Education - 9) \times \mathbf{1}(Education > 9) + 2.3. \end{aligned}$$

This equation is displayed in Figure 2.8 using the solid line, and appears to fit much better. It indicates a 2% increase in mean wages for every year of education below 9, and a 12% increase in mean wages for every year of education above 9. It is still an approximation to the conditional mean but it appears to be fairly reasonable.

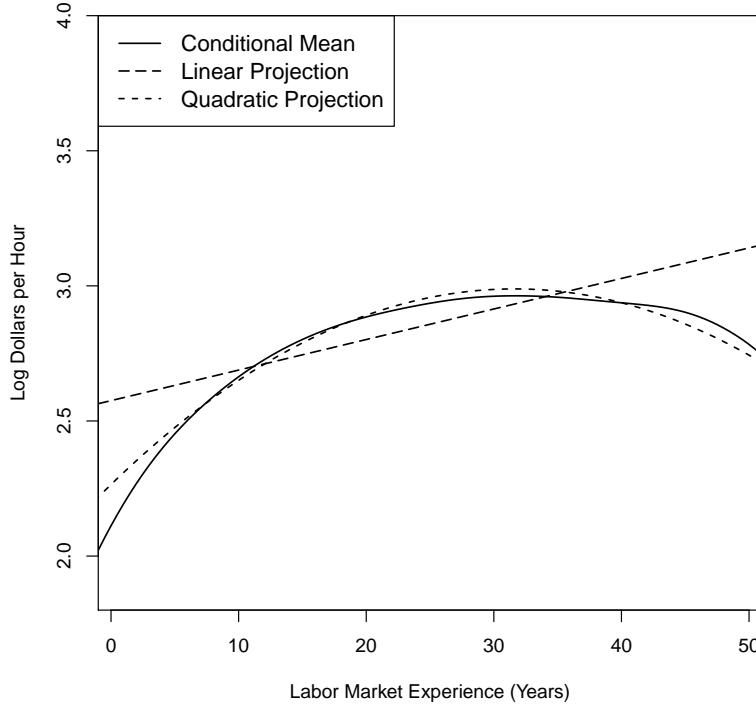


Figure 2.9: Linear and Quadratic Projections of  $\log(wage)$  onto Experience

For our third example we take the CEF of log wages as a function of years of experience for white men with 12 years of education, which was illustrated in Figure 2.6 and is repeated as the solid line in Figure 2.9. Superimposed on the figure are two projections. The first (given by the dot-dashed line) is the linear projection on experience

$$\mathcal{P}(\log(wage) | Experience) = 0.011Experience + 2.5$$

and the second (given by the dashed line) is the linear projection on experience and its square

$$\mathcal{P}(\log(wage) | Experience) = 0.046Experience - 0.0007Experience^2 + 2.3.$$

It is fairly clear from an examination of Figure 2.9 that the first linear projection is a poor approximation. It over-predicts wages for young and old workers, and under-predicts for the rest. Most importantly, it misses the strong downturn in expected wages for older wage-earners. The second projection fits much better. We can call this equation a **quadratic projection** since the function is quadratic in *experience*.

## 2.20 Linear Predictor Error Variance

As in the CEF model, we define the error variance as

$$\sigma^2 = \mathbb{E}(e^2).$$

Setting  $Q_{yy} = \mathbb{E}(y^2)$  and  $\mathbf{Q}_{yx} = \mathbb{E}(yx')$  we can write  $\sigma^2$  as

$$\begin{aligned}\sigma^2 &= \mathbb{E}((y - \mathbf{x}'\boldsymbol{\beta})^2) \\ &= \mathbb{E}(y^2) - 2\mathbb{E}(yx')\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbb{E}(\mathbf{x}\mathbf{x}')\boldsymbol{\beta} \\ &= Q_{yy} - 2\mathbf{Q}_{yx}\mathbf{Q}_{xx}^{-1}\mathbf{Q}_{xy} + \mathbf{Q}_{yx}\mathbf{Q}_{xx}^{-1}\mathbf{Q}_{xx}\mathbf{Q}_{xx}^{-1}\mathbf{Q}_{xy} \\ &= Q_{yy} - \mathbf{Q}_{yx}\mathbf{Q}_{xx}^{-1}\mathbf{Q}_{xy} \\ &\stackrel{\text{def}}{=} Q_{yy|\mathbf{x}}.\end{aligned}\tag{2.34}$$

One useful feature of this formula is that it shows that  $Q_{yy|\mathbf{x}} = Q_{yy} - \mathbf{Q}_{yx}\mathbf{Q}_{xx}^{-1}\mathbf{Q}_{xy}$  equals the variance of the error from the linear projection of  $y$  on  $\mathbf{x}$ .

## 2.21 Regression Coefficients

Sometimes it is useful to separate the constant from the other regressors, and write the linear projection equation in the format

$$y = \mathbf{x}'\boldsymbol{\beta} + \alpha + e\tag{2.35}$$

where  $\alpha$  is the intercept and  $\mathbf{x}$  does not contain a constant.

Taking expectations of this equation, we find

$$\mathbb{E}(y) = \mathbb{E}(\mathbf{x}'\boldsymbol{\beta}) + \mathbb{E}(\alpha) + \mathbb{E}(e)$$

or

$$\mu_y = \boldsymbol{\mu}_x'\boldsymbol{\beta} + \alpha$$

where  $\mu_y = \mathbb{E}(y)$  and  $\boldsymbol{\mu}_x = \mathbb{E}(\mathbf{x})$ , since  $\mathbb{E}(e) = 0$  from (2.26). (While  $\mathbf{x}$  does not contain a constant, the equation does so (2.26) still applies.) Rearranging, we find

$$\alpha = \mu_y - \boldsymbol{\mu}_x'\boldsymbol{\beta}.$$

Subtracting this equation from (2.35) we find

$$y - \mu_y = (\mathbf{x} - \boldsymbol{\mu}_x)'\boldsymbol{\beta} + e,\tag{2.36}$$

a linear equation between the centered variables  $y - \mu_y$  and  $\mathbf{x} - \boldsymbol{\mu}_x$ . (They are centered at their means, so are mean-zero random variables.) Because  $\mathbf{x} - \boldsymbol{\mu}_x$  is uncorrelated with  $e$ , (2.36) is also a linear projection, thus by the formula for the linear projection model,

$$\begin{aligned}\boldsymbol{\beta} &= \left( \mathbb{E}((\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{x} - \boldsymbol{\mu}_x)') \right)^{-1} \mathbb{E}((\mathbf{x} - \boldsymbol{\mu}_x)(y - \mu_y)) \\ &= \text{var}(\mathbf{x})^{-1} \text{cov}(\mathbf{x}, y)\end{aligned}$$

a function only of the covariances<sup>11</sup> of  $\mathbf{x}$  and  $y$ .

**Theorem 2.9** In the linear projection model

$$y = \mathbf{x}'\boldsymbol{\beta} + \alpha + e,$$

then

$$\alpha = \mu_y - \boldsymbol{\mu}_x'\boldsymbol{\beta}\tag{2.37}$$

and

$$\boldsymbol{\beta} = \text{var}(\mathbf{x})^{-1} \text{cov}(\mathbf{x}, y).\tag{2.38}$$

<sup>11</sup>The covariance matrix between vectors  $\mathbf{x}$  and  $\mathbf{z}$  is  $\text{cov}(\mathbf{x}, \mathbf{z}) = \mathbb{E}((\mathbf{x} - \mathbb{E}\mathbf{x})(\mathbf{z} - \mathbb{E}\mathbf{z})')$ . The (co)variance matrix of the vector  $\mathbf{x}$  is  $\text{var}(\mathbf{x}) = \text{cov}(\mathbf{x}, \mathbf{x}) = \mathbb{E}((\mathbf{x} - \mathbb{E}\mathbf{x})(\mathbf{x} - \mathbb{E}\mathbf{x})')$ .

## 2.22 Regression Sub-Vectors

Let the regressors be partitioned as

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}. \quad (2.39)$$

We can write the projection of  $y$  on  $\mathbf{x}$  as

$$\begin{aligned} y &= \mathbf{x}' \boldsymbol{\beta} + e \\ &= \mathbf{x}'_1 \boldsymbol{\beta}_1 + \mathbf{x}'_2 \boldsymbol{\beta}_2 + e \\ \mathbb{E}(\mathbf{x}e) &= \mathbf{0}. \end{aligned} \quad (2.40)$$

In this section we derive formula for the sub-vectors  $\boldsymbol{\beta}_1$  and  $\boldsymbol{\beta}_2$ .

Partition  $\mathbf{Q}_{xx}$  conformably with  $\mathbf{x}$

$$\mathbf{Q}_{xx} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix} = \begin{bmatrix} \mathbb{E}(\mathbf{x}_1 \mathbf{x}'_1) & \mathbb{E}(\mathbf{x}_1 \mathbf{x}'_2) \\ \mathbb{E}(\mathbf{x}_2 \mathbf{x}'_1) & \mathbb{E}(\mathbf{x}_2 \mathbf{x}'_2) \end{bmatrix}$$

and similarly  $\mathbf{Q}_{xy}$

$$\mathbf{Q}_{xy} = \begin{bmatrix} \mathbf{Q}_{1y} \\ \mathbf{Q}_{2y} \end{bmatrix} = \begin{bmatrix} \mathbb{E}(\mathbf{x}_1 y) \\ \mathbb{E}(\mathbf{x}_2 y) \end{bmatrix}.$$

By the partitioned matrix inversion formula (A.3)

$$\mathbf{Q}_{xx}^{-1} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix}^{-1} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{Q}^{11} & \mathbf{Q}^{12} \\ \mathbf{Q}^{21} & \mathbf{Q}^{22} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_{11\cdot 2}^{-1} & -\mathbf{Q}_{11\cdot 2}^{-1} \mathbf{Q}_{12} \mathbf{Q}_{22}^{-1} \\ -\mathbf{Q}_{22\cdot 1}^{-1} \mathbf{Q}_{21} \mathbf{Q}_{11}^{-1} & \mathbf{Q}_{22\cdot 1}^{-1} \end{bmatrix} \quad (2.41)$$

where  $\mathbf{Q}_{11\cdot 2} \stackrel{\text{def}}{=} \mathbf{Q}_{11} - \mathbf{Q}_{12} \mathbf{Q}_{22}^{-1} \mathbf{Q}_{21}$  and  $\mathbf{Q}_{22\cdot 1} \stackrel{\text{def}}{=} \mathbf{Q}_{22} - \mathbf{Q}_{21} \mathbf{Q}_{11}^{-1} \mathbf{Q}_{12}$ . Thus

$$\begin{aligned} \boldsymbol{\beta} &= \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \\ &= \begin{bmatrix} \mathbf{Q}_{11\cdot 2}^{-1} & -\mathbf{Q}_{11\cdot 2}^{-1} \mathbf{Q}_{12} \mathbf{Q}_{22}^{-1} \\ -\mathbf{Q}_{22\cdot 1}^{-1} \mathbf{Q}_{21} \mathbf{Q}_{11}^{-1} & \mathbf{Q}_{22\cdot 1}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{Q}_{1y} \\ \mathbf{Q}_{2y} \end{bmatrix} \\ &= \begin{pmatrix} \mathbf{Q}_{11\cdot 2}^{-1} (\mathbf{Q}_{1y} - \mathbf{Q}_{12} \mathbf{Q}_{22}^{-1} \mathbf{Q}_{2y}) \\ \mathbf{Q}_{22\cdot 1}^{-1} (\mathbf{Q}_{2y} - \mathbf{Q}_{21} \mathbf{Q}_{11}^{-1} \mathbf{Q}_{1y}) \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{Q}_{11\cdot 2}^{-1} \mathbf{Q}_{1y\cdot 2} \\ \mathbf{Q}_{22\cdot 1}^{-1} \mathbf{Q}_{2y\cdot 1} \end{pmatrix}. \end{aligned}$$

We have shown that

$$\begin{aligned} \boldsymbol{\beta}_1 &= \mathbf{Q}_{11\cdot 2}^{-1} \mathbf{Q}_{1y\cdot 2} \\ \boldsymbol{\beta}_2 &= \mathbf{Q}_{22\cdot 1}^{-1} \mathbf{Q}_{2y\cdot 1}. \end{aligned}$$

## 2.23 Coefficient Decomposition

In the previous section we derived formulae for the coefficient sub-vectors  $\boldsymbol{\beta}_1$  and  $\boldsymbol{\beta}_2$ . We now use these formulae to give a useful interpretation of the coefficients in terms of an iterated projection.

Take equation (2.40) for the case  $\dim(x_1) = 1$  so that  $\beta_1 \in \mathbb{R}$ .

$$y = x_1 \beta_1 + \mathbf{x}'_2 \boldsymbol{\beta}_2 + e. \quad (2.42)$$

Now consider the projection of  $x_1$  on  $\mathbf{x}_2$ :

$$\begin{aligned} x_1 &= \mathbf{x}'_2 \boldsymbol{\gamma}_2 + u_1 \\ \mathbb{E}(\mathbf{x}_2 u_1) &= \mathbf{0}. \end{aligned}$$

From (2.20) and (2.34),  $\boldsymbol{\gamma}_2 = \mathbf{Q}_{22}^{-1} \mathbf{Q}_{21}$  and  $\mathbb{E} u_1^2 = \mathbf{Q}_{11 \cdot 2} = \mathbf{Q}_{11} - \mathbf{Q}_{12} \mathbf{Q}_{22}^{-1} \mathbf{Q}_{21}$ . We can also calculate that

$$\mathbb{E}(u_1 y) = \mathbb{E}((x_1 - \boldsymbol{\gamma}'_2 \mathbf{x}_2) y) = \mathbb{E}(x_1 y) - \boldsymbol{\gamma}'_2 \mathbb{E}(\mathbf{x}_2 y) = \mathbf{Q}_{1y} - \mathbf{Q}_{12} \mathbf{Q}_{22}^{-1} \mathbf{Q}_{2y} = \mathbf{Q}_{1y \cdot 2}.$$

We have found that

$$\beta_1 = \mathbf{Q}_{11 \cdot 2}^{-1} \mathbf{Q}_{1y \cdot 2} = \frac{\mathbb{E}(u_1 y)}{\mathbb{E}(u_1^2)}$$

the coefficient from the simple regression of  $y$  on  $u_1$ .

What this means is that in the multivariate projection equation (2.42), the coefficient  $\beta_1$  equals the projection coefficient from a regression of  $y$  on  $u_1$ , the error from a projection of  $x_1$  on the other regressors  $\mathbf{x}_2$ . The error  $u_1$  can be thought of as the component of  $x_1$  which is not linearly explained by the other regressors. Thus the coefficient  $\beta_1$  equals the linear effect of  $x_1$  on  $y$ , after stripping out the effects of the other variables.

There was nothing special in the choice of the variable  $x_1$ . This derivation applies symmetrically to all coefficients in a linear projection. Each coefficient equals the simple regression of  $y$  on the error from a projection of that regressor on all the other regressors. Each coefficient equals the linear effect of that variable on  $y$ , after linearly controlling for all the other regressors.

## 2.24 Omitted Variable Bias

Again, let the regressors be partitioned as in (2.39). Consider the projection of  $y$  on  $\mathbf{x}_1$  only. Perhaps this is done because the variables  $\mathbf{x}_2$  are not observed. This is the equation

$$\begin{aligned} y &= \mathbf{x}'_1 \boldsymbol{\gamma}_1 + u \\ \mathbb{E}(\mathbf{x}_1 u) &= \mathbf{0}. \end{aligned} \tag{2.43}$$

Notice that we have written the coefficient on  $\mathbf{x}_1$  as  $\boldsymbol{\gamma}_1$  rather than  $\boldsymbol{\beta}_1$  and the error as  $u$  rather than  $e$ . This is because (2.43) is different than (2.40). Goldberger (1991) introduced the catchy labels **long regression** for (2.40) and **short regression** for (2.43) to emphasize the distinction.

Typically,  $\boldsymbol{\beta}_1 \neq \boldsymbol{\gamma}_1$ , except in special cases. To see this, we calculate

$$\begin{aligned} \boldsymbol{\gamma}_1 &= (\mathbb{E}(\mathbf{x}_1 \mathbf{x}'_1))^{-1} \mathbb{E}(\mathbf{x}_1 y) \\ &= (\mathbb{E}(\mathbf{x}_1 \mathbf{x}'_1))^{-1} \mathbb{E}(\mathbf{x}_1 (\mathbf{x}'_1 \boldsymbol{\beta}_1 + \mathbf{x}'_2 \boldsymbol{\beta}_2 + e)) \\ &= \boldsymbol{\beta}_1 + (\mathbb{E}(\mathbf{x}_1 \mathbf{x}'_1))^{-1} \mathbb{E}(\mathbf{x}_1 \mathbf{x}'_2) \boldsymbol{\beta}_2 \\ &= \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}_{12} \boldsymbol{\beta}_2 \end{aligned}$$

where  $\boldsymbol{\Gamma}_{12} = \mathbf{Q}_{11}^{-1} \mathbf{Q}_{12}$  is the coefficient matrix from a projection of  $\mathbf{x}_2$  on  $\mathbf{x}_1$ , where we use the notation from Section 2.22.

Observe that  $\boldsymbol{\gamma}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}_{12} \boldsymbol{\beta}_2 \neq \boldsymbol{\beta}_1$  unless  $\boldsymbol{\Gamma}_{12} = \mathbf{0}$  or  $\boldsymbol{\beta}_2 = \mathbf{0}$ . Thus the short and long regressions have different coefficients on  $\mathbf{x}_1$ . They are the same only under one of two conditions. First, if the projection of  $\mathbf{x}_2$  on  $\mathbf{x}_1$  yields a set of zero coefficients (they are uncorrelated), or second, if the coefficient on  $\mathbf{x}_2$  in (2.40) is zero. In general, the coefficient in (2.43) is  $\boldsymbol{\gamma}_1$  rather than  $\boldsymbol{\beta}_1$ . The difference  $\boldsymbol{\Gamma}_{12} \boldsymbol{\beta}_2$  between  $\boldsymbol{\gamma}_1$  and  $\boldsymbol{\beta}_1$  is known as **omitted variable bias**. It is the consequence of omission of a relevant correlated variable.

To avoid omitted variables bias the standard advice is to include all potentially relevant variables in estimated models. By construction, the general model will be free of such bias. Unfortunately in many cases it is not feasible to completely follow this advice as many desired variables are not observed. In this case, the possibility of omitted variables bias should be acknowledged and discussed in the course of an empirical investigation.

For example, suppose  $y$  is log wages,  $x_1$  is education, and  $x_2$  is intellectual ability. It seems reasonable to suppose that education and intellectual ability are positively correlated (highly able individuals attain

higher levels of education) which means  $\Gamma_{12} > 0$ . It also seems reasonable to suppose that conditional on education, individuals with higher intelligence will earn higher wages on average, so that  $\beta_2 > 0$ . This implies that  $\Gamma_{12}\beta_2 > 0$  and  $\gamma_1 = \beta_1 + \Gamma_{12}\beta_2 > \beta_1$ . Therefore, it seems reasonable to expect that in a regression of wages on education with ability omitted, the coefficient on education is higher than in a regression where ability is included. In other words, in this context the omitted variable biases the regression coefficient upwards. It is possible, for example, that  $\beta_1 = 0$  so that education has no direct effect on wages yet  $\gamma_1 = \Gamma_{12}\beta_2 > 0$  meaning that the regression coefficient on education alone is positive, but is a consequence of the unmodeled correlation between education and intellectual ability.

Unfortunately the above simple characterization of omitted variable bias does not immediately carry over to more complicated settings, as discovered by Luca, Magnus, and Peracchi (2018). For example, suppose we compare three nested projections

$$\begin{aligned} y &= \mathbf{x}'_1 \boldsymbol{\gamma}_1 + u_1 \\ y &= \mathbf{x}'_1 \boldsymbol{\delta}_1 + \mathbf{x}'_2 \boldsymbol{\delta}_2 + u_2 \\ y &= \mathbf{x}'_1 \boldsymbol{\beta}_1 + \mathbf{x}'_2 \boldsymbol{\beta}_2 + \mathbf{x}'_3 \boldsymbol{\beta}_3 + e. \end{aligned}$$

We can call them the short, medium, and long regressions. Suppose that the parameter of interest is  $\boldsymbol{\beta}_1$  in the long regression. We are interested in the consequences of omitting  $\mathbf{x}_3$  when estimating the medium regression, and of omitting both  $\mathbf{x}_2$  and  $\mathbf{x}_3$  when estimating the short regression. In particular we are interested in the question: Is it better to estimate the short or medium regression, given that both omit  $\mathbf{x}_3$ ? Intuition suggests that the medium regression should be “less biased” but it is worth investigating in greater detail. By similar calculations to those above, we find that

$$\begin{aligned} \boldsymbol{\gamma}_1 &= \boldsymbol{\beta}_1 + \Gamma_{12}\boldsymbol{\beta}_2 + \Gamma_{13}\boldsymbol{\beta}_3 \\ \boldsymbol{\delta}_1 &= \boldsymbol{\beta}_1 + \Gamma_{13\cdot 2}\boldsymbol{\beta}_3 \end{aligned}$$

where  $\Gamma_{13\cdot 2} = \mathbf{Q}_{11\cdot 2}^{-1} \mathbf{Q}_{13\cdot 2}$  using the notation from Section 2.22.

We see that the bias in the short regression coefficient is  $\Gamma_{12}\boldsymbol{\beta}_2 + \Gamma_{13}\boldsymbol{\beta}_3$  which depends on both  $\boldsymbol{\beta}_2$  and  $\boldsymbol{\beta}_3$ , while that for the medium regression coefficient is  $\Gamma_{13\cdot 2}\boldsymbol{\beta}_3$  which only depends on  $\boldsymbol{\beta}_3$ . So the bias for the medium regression is less complicated, and intuitively seems more likely to be smaller than that of the short regression. However it is impossible to strictly rank the two. It is quite possible that  $\boldsymbol{\gamma}_1$  is less biased than  $\boldsymbol{\delta}_1$ . Thus as a general rule it is strictly impossible to state that estimation of the medium regression will be less biased than estimation of the short regression.

## 2.25 Best Linear Approximation

There are alternative ways we could construct a linear approximation  $\mathbf{x}'\boldsymbol{\beta}$  to the conditional mean  $m(\mathbf{x})$ . In this section we show that one alternative approach turns out to yield the same answer as the best linear predictor.

We start by defining the mean-square approximation error of  $\mathbf{x}'\boldsymbol{\beta}$  to  $m(\mathbf{x})$  as the expected squared difference between  $\mathbf{x}'\boldsymbol{\beta}$  and the conditional mean  $m(\mathbf{x})$

$$d(\boldsymbol{\beta}) = \mathbb{E}((m(\mathbf{x}) - \mathbf{x}'\boldsymbol{\beta})^2).$$

The function  $d(\boldsymbol{\beta})$  is a measure of the deviation of  $\mathbf{x}'\boldsymbol{\beta}$  from  $m(\mathbf{x})$ . If the two functions are identical then  $d(\boldsymbol{\beta}) = 0$ , otherwise  $d(\boldsymbol{\beta}) > 0$ . We can also view the mean-square difference  $d(\boldsymbol{\beta})$  as a density-weighted average of the function  $(m(\mathbf{x}) - \mathbf{x}'\boldsymbol{\beta})^2$ , since

$$d(\boldsymbol{\beta}) = \int_{\mathbb{R}^k} (m(\mathbf{x}) - \mathbf{x}'\boldsymbol{\beta})^2 f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}$$

where  $f_{\mathbf{x}}(\mathbf{x})$  is the marginal density of  $\mathbf{x}$ .

We can then define the best linear approximation to the conditional  $m(\mathbf{x})$  as the function  $\mathbf{x}'\boldsymbol{\beta}$  obtained by selecting  $\boldsymbol{\beta}$  to minimize  $d(\boldsymbol{\beta})$ :

$$\boldsymbol{\beta} = \underset{\boldsymbol{b} \in \mathbb{R}^k}{\operatorname{argmin}} d(\boldsymbol{b}). \quad (2.44)$$

Similar to the best linear predictor we are measuring accuracy by expected squared error. The difference is that the best linear predictor (2.16) selects  $\boldsymbol{\beta}$  to minimize the expected squared prediction error, while the best linear approximation (2.44) selects  $\boldsymbol{\beta}$  to minimize the expected squared approximation error.

Despite the different definitions, it turns out that the best linear predictor and the best linear approximation are identical. By the same steps as in (2.18) plus an application of conditional expectations we can find that

$$\boldsymbol{\beta} = (\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-1} \mathbb{E}(\mathbf{x}m(\mathbf{x})) \quad (2.45)$$

$$= (\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-1} \mathbb{E}(\mathbf{x}y) \quad (2.46)$$

(see Exercise 2.19). Thus (2.44) equals (2.16). We conclude that the definition (2.44) can be viewed as an alternative motivation for the linear projection coefficient.

## 2.26 Regression to the Mean

The term **regression** originated in an influential paper by Francis Galton (1886), where he examined the joint distribution of the stature (height) of parents and children. Effectively, he was estimating the conditional mean of children's height given their parent's height. Galton discovered that this conditional mean was approximately linear with a slope of  $2/3$ . This implies that *on average* a child's height is more mediocre (average) than his or her parent's height. Galton called this phenomenon **regression to the mean**, and the label **regression** has stuck to this day to describe most conditional relationships.

One of Galton's fundamental insights was to recognize that if the marginal distributions of  $y$  and  $x$  are the same (e.g. the heights of children and parents in a stable environment) then the regression slope in a linear projection is always less than one.

To be more precise, take the simple linear projection

$$y = x\beta + \alpha + e \quad (2.47)$$

where  $y$  equals the height of the child and  $x$  equals the height of the parent. Assume that  $y$  and  $x$  have the same mean, so that  $\mu_y = \mu_x = \mu$ . Then from (2.37)

$$\alpha = (1 - \beta)\mu$$

so we can write the linear projection (2.47) as

$$\mathcal{P}(y | x) = (1 - \beta)\mu + x\beta.$$

This shows that the projected height of the child is a weighted average of the population average height  $\mu$  and the parent's height  $x$ , with the weight equal to the regression slope  $\beta$ . When the height distribution is stable across generations, so that  $\text{var}(y) = \text{var}(x)$ , then this slope is the simple correlation of  $y$  and  $x$ . Using (2.38)

$$\beta = \frac{\text{cov}(x, y)}{\text{var}(x)} = \text{corr}(x, y).$$

By the Cauchy-Schwarz inequality (B.31),  $-1 \leq \text{corr}(x, y) \leq 1$ , with  $\text{corr}(x, y) = 1$  only in the degenerate case  $y = x$ . Thus if we exclude degeneracy,  $\beta$  is strictly less than 1.

This means that on average a child's height is more mediocre (closer to the population average) than the parent's.

A common error – known as the **regression fallacy** – is to infer from  $\beta < 1$  that the population is **converging**, meaning that its variance is declining towards zero. This is a fallacy because we derived the implication  $\beta < 1$  under the assumption of constant means and variances. So certainly  $\beta < 1$  does not imply that the variance  $y$  is less than than the variance of  $x$ .

Another way of seeing this is to examine the conditions for convergence in the context of equation (2.47). Since  $x$  and  $e$  are uncorrelated, it follows that

$$\text{var}(y) = \beta^2 \text{var}(x) + \text{var}(e).$$

Then  $\text{var}(y) < \text{var}(x)$  if and only if

$$\beta^2 < 1 - \frac{\text{var}(e)}{\text{var}(x)}$$

which is not implied by the simple condition  $|\beta| < 1$ .

The regression fallacy arises in related empirical situations. Suppose you sort families into groups by the heights of the parents, and then plot the average heights of each subsequent generation over time. If the population is stable, the regression property implies that the plots lines will converge – children's height will be more average than their parents. The regression fallacy is to incorrectly conclude that the population is converging. A message to be learned from this example is that such plots are misleading for inferences about convergence.

The regression fallacy is subtle. It is easy for intelligent economists to succumb to its temptation. A famous example is *The Triumph of Mediocrity in Business* by Horace Secrist, published in 1933. In this book, Secrist carefully and with great detail documented that in a sample of department stores over 1920–1930, when he divided the stores into groups based on 1920–1921 profits, and plotted the average profits of these groups for the subsequent 10 years, he found clear and persuasive evidence for convergence “toward mediocrity”. Of course, there was no discovery – regression to the mean is a necessary feature of stable distributions.

## 2.27 Reverse Regression

Galton noticed another interesting feature of the bivariate distribution. There is nothing special about a regression of  $y$  on  $x$ . We can also regress  $x$  on  $y$ . (In his heredity example this is the best linear predictor of the height of parents given the height of their children.) This regression takes the form

$$x = y\beta^* + \alpha^* + e^*. \quad (2.48)$$

This is sometimes called the **reverse regression**. In this equation, the coefficients  $\alpha^*$ ,  $\beta^*$  and error  $e^*$  are defined by linear projection. In a stable population we find that

$$\beta^* = \text{corr}(x, y) = \beta$$

$$\alpha^* = (1 - \beta)\mu = \alpha$$

which are exactly the same as in the projection of  $y$  on  $x$ ! The intercept and slope have exactly the same values in the forward and reverse projections!

While this algebraic discovery is quite simple, it is counter-intuitive. Instead, a common yet mistaken guess for the form of the reverse regression is to take the equation (2.47), divide through by  $\beta$  and rewrite to find the equation

$$x = y\frac{1}{\beta} - \frac{\alpha}{\beta} - \frac{1}{\beta}e \quad (2.49)$$

suggesting that the projection of  $x$  on  $y$  should have a slope coefficient of  $1/\beta$  instead of  $\beta$ , and intercept of  $-\alpha/\beta$  rather than  $\alpha$ . What went wrong? Equation (2.49) is perfectly valid, because it is a simple manipulation of the valid equation (2.47). The trouble is that (2.49) is neither a CEF nor a linear projection.

Inverting a projection (or CEF) does not yield a projection (or CEF). Instead, (2.48) is a valid projection, not (2.49).

In any event, Galton's finding was that when the variables are standardized, the slope in both projections ( $y$  on  $x$ , and  $x$  and  $y$ ) equals the correlation, and both equations exhibit regression to the mean. It is not a causal relation, but a natural feature of all joint distributions.

## 2.28 Limitations of the Best Linear Projection

Let's compare the linear projection and linear CEF models.

From Theorem 2.4.4 we know that the CEF error has the property  $\mathbb{E}(\mathbf{x}e) = \mathbf{0}$ . Thus a linear CEF is the best linear projection. However, the converse is not true as the projection error does not necessarily satisfy  $\mathbb{E}(e | \mathbf{x}) = 0$ . Furthermore, the linear projection may be a poor approximation to the CEF.

To see these points in a simple example, suppose that the true process is  $y = x + x^2$  with  $x \sim N(0, 1)$ . In this case the true CEF is  $m(x) = x + x^2$  and there is no error. Now consider the linear projection of  $y$  on  $x$  and a constant, namely the model  $y = \beta x + \alpha + u$ . Since  $x \sim N(0, 1)$  then  $x$  and  $x^2$  are uncorrelated and the linear projection takes the form  $\mathcal{P}(y | x) = x + 1$ . This is quite different from the true CEF  $m(x) = x + x^2$ . The projection error equals  $e = x^2 - 1$ , which is a deterministic function of  $x$ , yet is uncorrelated with  $x$ . We see in this example that a projection error need not be a CEF error, and a linear projection can be a poor approximation to the CEF.

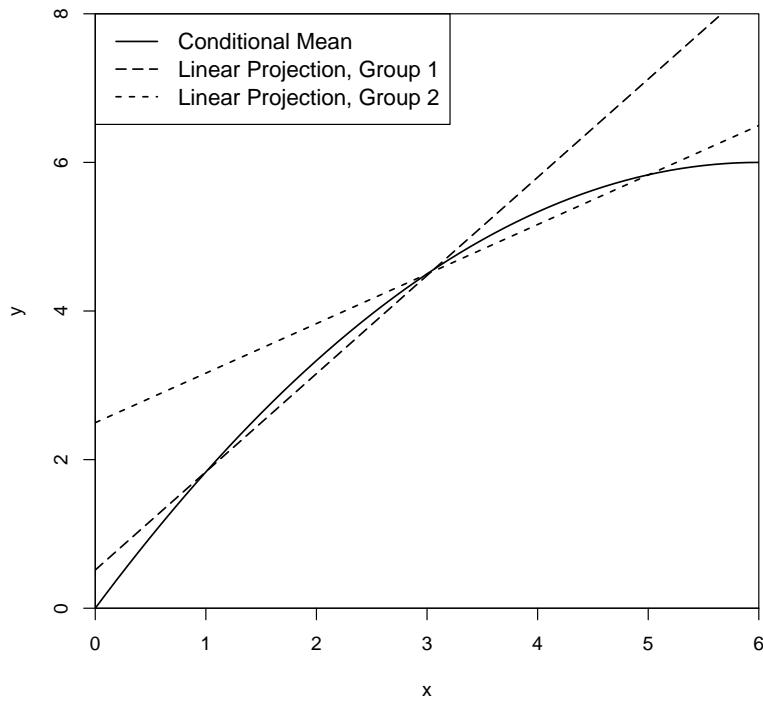


Figure 2.10: Conditional Mean and Two Linear Projections

Another defect of linear projection is that it is sensitive to the marginal distribution of the regressors when the conditional mean is non-linear. We illustrate the issue in Figure 2.10 for a constructed<sup>12</sup> joint distribution of  $y$  and  $x$ . The solid line is the non-linear CEF of  $y$  given  $x$ . The data are divided in two groups – Group 1 and Group 2 – which have different marginal distributions for the regressor  $x$ , and

<sup>12</sup>The  $x$  in Group 1 are  $N(2, 1)$  and those in Group 2 are  $N(4, 1)$ , and the conditional distribution of  $y$  given  $x$  is  $N(m(x), 1)$  where  $m(x) = 2x - x^2/6$ .

Group 1 has a lower mean value of  $x$  than Group 2. The separate linear projections of  $y$  on  $x$  for these two groups are displayed in the Figure by the dashed lines. These two projections are distinct approximations to the CEF. A defect with linear projection is that it leads to the incorrect conclusion that the effect of  $x$  on  $y$  is different for individuals in the two groups. This conclusion is incorrect because in fact there is no difference in the conditional mean function. The apparent difference is a by-product of a linear approximation to a nonlinear mean, combined with different marginal distributions for the conditioning variables.

## 2.29 Random Coefficient Model

A model which is notationally similar to but conceptually distinct from the linear CEF model is the linear random coefficient model. It takes the form

$$y = \mathbf{x}'\boldsymbol{\eta}$$

where the individual-specific coefficient  $\boldsymbol{\eta}$  is random and independent of  $\mathbf{x}$ . For example, if  $\mathbf{x}$  is years of schooling and  $y$  is log wages, then  $\boldsymbol{\eta}$  is the individual-specific returns to schooling. If a person obtains an extra year of schooling,  $\boldsymbol{\eta}$  is the actual change in their wage. The random coefficient model allows the returns to schooling to vary in the population. Some individuals might have a high return to education (a high  $\boldsymbol{\eta}$ ) and others a low return, possibly 0, or even negative.

In the linear CEF model the regressor coefficient equals the regression derivative – the change in the conditional mean due to a change in the regressors,  $\boldsymbol{\beta} = \nabla m(\mathbf{x})$ . This is not the effect on a given individual, it is the effect on the population average. In contrast, in the random coefficient model, the random vector  $\boldsymbol{\eta} = \nabla(\mathbf{x}'\boldsymbol{\eta})$  is the true causal effect – the change in the response variable  $y$  itself due to a change in the regressors.

It is interesting, however, to discover that the linear random coefficient model implies a linear CEF. To see this, let  $\boldsymbol{\beta}$  and  $\Sigma$  denote the mean and covariance matrix of  $\boldsymbol{\eta}$ :

$$\begin{aligned}\boldsymbol{\beta} &= \mathbb{E}(\boldsymbol{\eta}) \\ \Sigma &= \text{var}(\boldsymbol{\eta})\end{aligned}$$

and then decompose the random coefficient as

$$\boldsymbol{\eta} = \boldsymbol{\beta} + \mathbf{u}$$

where  $\mathbf{u}$  is distributed independently of  $\mathbf{x}$  with mean zero and covariance matrix  $\Sigma$ . Then we can write

$$\mathbb{E}(y | \mathbf{x}) = \mathbf{x}'\mathbb{E}(\boldsymbol{\eta} | \mathbf{x}) = \mathbf{x}'\mathbb{E}(\boldsymbol{\eta}) = \mathbf{x}'\boldsymbol{\beta}$$

so the CEF is linear in  $\mathbf{x}$ , and the coefficients  $\boldsymbol{\beta}$  equal the mean of the random coefficient  $\boldsymbol{\eta}$ .

We can thus write the equation as a linear CEF

$$y = \mathbf{x}'\boldsymbol{\beta} + e$$

where  $e = \mathbf{x}'\mathbf{u}$  and  $\mathbf{u} = \boldsymbol{\eta} - \boldsymbol{\beta}$ . The error is conditionally mean zero:

$$\mathbb{E}(e | \mathbf{x}) = 0.$$

Furthermore

$$\begin{aligned}\text{var}(e | \mathbf{x}) &= \mathbf{x}'\text{var}(\boldsymbol{\eta})\mathbf{x} \\ &= \mathbf{x}'\Sigma\mathbf{x}\end{aligned}$$

so the error is conditionally heteroskedastic with its variance a quadratic function of  $\mathbf{x}$ .

**Theorem 2.10** In the linear random coefficient model  $y = \mathbf{x}'\boldsymbol{\eta}$  with  $\boldsymbol{\eta}$  independent of  $\mathbf{x}$ ,  $\mathbb{E}\|\mathbf{x}\|^2 < \infty$ , and  $\mathbb{E}\|\boldsymbol{\eta}\|^2 < \infty$ , then

$$\begin{aligned}\mathbb{E}(y | \mathbf{x}) &= \mathbf{x}'\boldsymbol{\beta} \\ \text{var}(y | \mathbf{x}) &= \mathbf{x}'\boldsymbol{\Sigma}\mathbf{x}\end{aligned}$$

where  $\boldsymbol{\beta} = \mathbb{E}(\boldsymbol{\eta})$  and  $\boldsymbol{\Sigma} = \text{var}(\boldsymbol{\eta})$ .

## 2.30 Causal Effects

So far we have avoided the concept of causality, yet often the underlying goal of an econometric analysis is to uncover a causal relationship between variables. It is often of great interest to understand the causes and effects of decisions, actions, and policies. For example, we may be interested in the effect of class sizes on test scores, police expenditures on crime rates, climate change on economic activity, years of schooling on wages, institutional structure on growth, the effectiveness of rewards on behavior, the consequences of medical procedures for health outcomes, or any variety of possible causal relationships. In each case, the goal is to understand what is the actual effect on the outcome  $y$  due to a change in the input  $x$ . We are not just interested in the conditional mean or linear projection, we would like to know the actual change.

Two inherent barriers are that the causal effect is typically specific to an individual and that it is unobserved.

Consider the effect of schooling on wages. The causal effect is the actual difference a person would receive in wages if we could change their level of education *holding all else constant*. This is specific to each individual as their employment outcomes in these two distinct situations is individual. The causal effect is unobserved because the most we can observe is their actual level of education and their actual wage, but not the counterfactual wage if their education had been different.

To be even more specific, suppose that there are two individuals, Jennifer and George, and both have the possibility of being high-school graduates or college graduates, but both would have received different wages given their choices. For example, suppose that Jennifer would have earned \$10 an hour as a high-school graduate and \$20 an hour as a college graduate while George would have earned \$8 as a high-school graduate and \$12 as a college graduate. In this example the causal effect of schooling is \$10 an hour for Jennifer and \$4 an hour for George. The causal effects are specific to the individual and neither causal effect is observed.

A variable  $x_1$  can be said to have a causal effect on the response variable  $y$  if the latter changes when all other inputs are held constant. To make this precise we need a mathematical formulation. We can write a full model for the response variable  $y$  as

$$y = h(x_1, \mathbf{x}_2, \mathbf{u}) \tag{2.50}$$

where  $x_1$  and  $\mathbf{x}_2$  are the observed variables,  $\mathbf{u}$  is an  $\ell \times 1$  unobserved random factor, and  $h$  is a functional relationship. This framework, called the **potential outcomes** framework, includes as a special case the random coefficient model (2.29) studied earlier. We define the causal effect of  $x_1$  within this model as the change in  $y$  due to a change in  $x_1$  holding the other variables  $\mathbf{x}_2$  and  $\mathbf{u}$  constant.

**Definition 2.6** In the model (2.50) the **causal effect** of  $x_1$  on  $y$  is

$$C(x_1, \mathbf{x}_2, \mathbf{u}) = \nabla_{x_1} h(x_1, \mathbf{x}_2, \mathbf{u}), \tag{2.51}$$

the change in  $y$  due to a change in  $x_1$ , holding  $\mathbf{x}_2$  and  $\mathbf{u}$  constant.

To understand this concept, imagine taking a single individual. As far as our structural model is concerned, this person is described by their observables  $x_1$  and  $\mathbf{x}_2$  and their unobservables  $\mathbf{u}$ . In a wage regression the unobservables would include characteristics such as the person's abilities, skills, work ethic, interpersonal connections, and preferences. The causal effect of  $x_1$  (say, education) is the change in the wage as  $x_1$  changes, holding constant all other observables and unobservables.

It may be helpful to understand that (2.51) is a definition, and does not necessarily describe causality in a fundamental or experimental sense. Perhaps it would be more appropriate to label (2.51) as a **structural effect** (the effect within the structural model).

Sometimes it is useful to write this relationship as a potential outcome function

$$y(x_1) = h(x_1, \mathbf{x}_2, \mathbf{u})$$

where the notation implies that  $y(x_1)$  is holding  $\mathbf{x}_2$  and  $\mathbf{u}$  constant.

A popular example arises in the analysis of treatment effects with a binary regressor  $x_1$ . Let  $x_1 = 1$  indicate treatment (e.g. a medical procedure) and  $x_1 = 0$  indicate non-treatment. In this case  $y(x_1)$  can be written

$$\begin{aligned} y(0) &= h(0, \mathbf{x}_2, \mathbf{u}) \\ y(1) &= h(1, \mathbf{x}_2, \mathbf{u}). \end{aligned}$$

In the literature on treatment effects, it is common to refer to  $y(0)$  and  $y(1)$  as the latent outcomes associated with non-treatment and treatment, respectively. That is, for a given individual,  $y(0)$  is the health outcome if there is no treatment, and  $y(1)$  is the health outcome if there is treatment. The causal effect of treatment for the individual is the change in their health outcome due to treatment – the change in  $y$  as we hold both  $\mathbf{x}_2$  and  $\mathbf{u}$  constant:

$$C(\mathbf{x}_2, \mathbf{u}) = y(1) - y(0).$$

This is random (a function of  $\mathbf{x}_2$  and  $\mathbf{u}$ ) as both potential outcomes  $y(0)$  and  $y(1)$  are different across individuals.

In a sample, we cannot observe both outcomes from the same individual, we only observe the realized value

$$y = \begin{cases} y(0) & \text{if } x_1 = 0 \\ y(1) & \text{if } x_1 = 1. \end{cases}$$

As the causal effect varies across individuals and is not observable, it cannot be measured on the individual level. We therefore focus on aggregate causal effects, in particular what is known as the average causal effect.

**Definition 2.7** In the model (2.50) the **average causal effect** of  $x_1$  on  $y$  conditional on  $\mathbf{x}_2$  is

$$\begin{aligned} \text{ACE}(x_1, \mathbf{x}_2) &= \mathbb{E}(C(x_1, \mathbf{x}_2, \mathbf{u}) | x_1, \mathbf{x}_2) \\ &= \int_{\mathbb{R}^\ell} \nabla_1 h(x_1, \mathbf{x}_2, \mathbf{u}) f(\mathbf{u} | x_1, \mathbf{x}_2) d\mathbf{u} \end{aligned}$$

where  $f(\mathbf{u} | x_1, \mathbf{x}_2)$  is the conditional density of  $\mathbf{u}$  given  $x_1, \mathbf{x}_2$ .

We can think of the average causal effect  $\text{ACE}(x_1, \mathbf{x}_2)$  as the average effect in the general population. In our Jennifer & George schooling example given earlier, supposing that half of the population are Jennifer's and the other half George's, then the average causal effect of college is  $(10 + 4)/2 = \$7$  an hour. This is not the individual causal effect, it is the average of the causal effect across all individuals in the

population. Given data on only educational attainment and wages, the ACE of \$7 is the best we can hope to learn.

When we conduct a regression analysis (that is, consider the regression of observed wages on educational attainment) we might hope that the regression reveals the average causal effect. Technically, that the regression derivative (the coefficient on education) equals the ACE. Is this the case? In other words, what is the relationship between the average causal effect  $\text{ACE}(x_1, \mathbf{x}_2)$  and the regression derivative  $\nabla_1 m(x_1, \mathbf{x}_2)$ ? Equation (2.50) implies that the CEF is

$$\begin{aligned} m(x_1, \mathbf{x}_2) &= \mathbb{E}(h(x_1, \mathbf{x}_2, \mathbf{u}) | x_1, \mathbf{x}_2) \\ &= \int_{\mathbb{R}^\ell} h(x_1, \mathbf{x}_2, \mathbf{u}) f(\mathbf{u} | x_1, \mathbf{x}_2) d\mathbf{u}, \end{aligned}$$

the average causal equation, averaged over the conditional distribution of the unobserved component  $\mathbf{u}$ .

Applying the marginal effect operator, the regression derivative is

$$\begin{aligned} \nabla_1 m(x_1, \mathbf{x}_2) &= \int_{\mathbb{R}^\ell} \nabla_1 h(x_1, \mathbf{x}_2, \mathbf{u}) f(\mathbf{u} | x_1, \mathbf{x}_2) d\mathbf{u} \\ &\quad + \int_{\mathbb{R}^\ell} h(x_1, \mathbf{x}_2, \mathbf{u}) \nabla_1 f(\mathbf{u} | x_1, \mathbf{x}_2) d\mathbf{u} \\ &= \text{ACE}(x_1, \mathbf{x}_2) + \int_{\mathbb{R}^\ell} h(x_1, \mathbf{x}_2, \mathbf{u}) \nabla_1 f(\mathbf{u} | x_1, \mathbf{x}_2) d\mathbf{u}. \end{aligned} \quad (2.52)$$

Equation (2.52) shows that in general, the regression derivative does not equal the average causal effect. The difference is the second term on the right-hand-side of (2.52). The regression derivative and ACE equal in the special case when this term equals zero, which occurs when  $\nabla_1 f(\mathbf{u} | x_1, \mathbf{x}_2) = 0$ , that is, when the conditional density of  $\mathbf{u}$  given  $(x_1, \mathbf{x}_2)$  does not depend on  $x_1$ . When this condition holds then the regression derivative equals the ACE, which means that regression analysis can be interpreted causally, in the sense that it uncovers average causal effects.

The condition is sufficiently important that it has a special name in the treatment effects literature.

**Definition 2.8 Conditional Independence Assumption (CIA).** Conditional on  $\mathbf{x}_2$ , the random variables  $x_1$  and  $\mathbf{u}$  are statistically independent.

The CIA implies  $f(\mathbf{u} | x_1, \mathbf{x}_2) = f(\mathbf{u} | \mathbf{x}_2)$  does not depend on  $x_1$ , and thus  $\nabla_1 f(\mathbf{u} | x_1, \mathbf{x}_2) = 0$ . Thus the CIA implies that  $\nabla_1 m(x_1, \mathbf{x}_2) = \text{ACE}(x_1, \mathbf{x}_2)$ , the regression derivative equals the average causal effect.

**Theorem 2.11** In the structural model (2.50), the Conditional Independence Assumption implies

$$\nabla_1 m(x_1, \mathbf{x}_2) = \text{ACE}(x_1, \mathbf{x}_2)$$

the regression derivative equals the average causal effect for  $x_1$  on  $y$  conditional on  $\mathbf{x}_2$ .

This is a fascinating result. It shows that whenever the unobservable is independent of the treatment variable (after conditioning on appropriate regressors) the regression derivative equals the average causal effect. In this case, the CEF has causal economic meaning, giving strong justification to estimation of the CEF. Our derivation also shows the critical role of the CIA. If CIA fails, then the equality of the regression derivative and ACE fails.

Table 2.2: Example Distribution

	\$8	\$10	\$12	\$20	Mean
High-School Graduate	10	6	0	0	\$8.75
College Graduate	0	0	6	10	\$17.00

This theorem is quite general. It applies equally to the treatment-effects model where  $x_1$  is binary or to more general settings where  $x_1$  is continuous.

It is also helpful to understand that the CIA is weaker than full independence of  $\mathbf{u}$  from the regressors  $(x_1, \mathbf{x}_2)$ . The CIA was introduced precisely as a minimal sufficient condition to obtain the desired result. Full independence implies the CIA and implies that each regression derivative equals that variable's average causal effect, but full independence is not necessary in order to causally interpret a subset of the regressors.

To illustrate, let's return to our education example involving a population with equal numbers of Jennifer's and George's. Recall that Jennifer earns \$10 as a high-school graduate and \$20 as a college graduate (and so has a causal effect of \$10) while George earns \$8 as a high-school graduate and \$12 as a college graduate (so has a causal effect of \$4). Given this information, the average causal effect of college is \$7, which is what we hope to learn from a regression analysis.

Now suppose that while in high school all students take an aptitude test, and if a student gets a high (H) score he or she goes to college with probability 3/4, and if a student gets a low (L) score he or she goes to college with probability 1/4. Suppose further that Jennifer's get an aptitude score of H with probability 3/4, while George's get a score of H with probability 1/4. Given this situation, 62.5% of Jennifer's will go to college<sup>13</sup>, while 37.5% of George's will go to college<sup>14</sup>.

An econometrician who randomly samples 32 individuals and collects data on educational attainment and wages will find the wage distribution in Table 2.2.

Let  $college$  denote a dummy variable taking the value of 1 for a college graduate, otherwise 0. Thus the regression of wages on college attendance takes the form

$$\mathbb{E}(wage | college) = 8.25college + 8.75.$$

The coefficient on the college dummy, \$8.25, is the regression derivative, and the implied wage effect of college attendance. But \$8.25 overstates the average causal effect of \$7. The reason is because the CIA fails. In this model the unobservable  $\mathbf{u}$  is the individual's type (Jennifer or George) which is not independent of the regressor  $x_1$  (education), since Jennifer is more likely to go to college than George. Since Jennifer's causal effect is higher than George's, the regression derivative overstates the ACE. The coefficient \$8.25 is not the average benefit of college attendance, rather it is the observed difference in realized wages in a population whose decision to attend college is correlated with their individual causal effect. At the risk of repeating myself, in this example, \$8.25 is the true regression derivative, it is the difference in average wages between those with a college education and those without. It is not, however, the average causal effect of college education in the population.

This does not mean that it is impossible to estimate the ACE. The key is conditioning on the appropriate variables. The CIA says that we need to find a variable  $x_2$  such that conditional on  $x_2$ ,  $\mathbf{u}$  and  $x_1$  (type and education) are independent. In this example a variable which will achieve this is the aptitude test score. The decision to attend college was based on the test score, not on an individual's type. Thus educational attainment and type are independent once we condition on the test score.

This also alters the ACE. Notice that Definition 2.7 is a function of  $x_2$  (the test score). Among the students who receive a high test score, 3/4 are Jennifer's and 1/4 are George's. Thus the ACE for students with a score of H is  $(3/4) \times 10 + (1/4) \times 4 = \$8.50$ . Among the students who receive a low test score, 1/4 are Jennifer's and 3/4 are George's. Thus the ACE for students with a score of L is  $(1/4) \times 10 + (3/4) \times 4 = \$5.50$ .

<sup>13</sup> $\mathbb{P}(\text{College}| \text{Jennifer}) = \mathbb{P}(\text{College}|H)\mathbb{P}(H| \text{Jennifer}) + \mathbb{P}(\text{College}|L)\mathbb{P}(L| \text{Jennifer}) = (3/4)^2 + (1/4)^2$

<sup>14</sup> $\mathbb{P}(\text{College}| \text{George}) = \mathbb{P}(\text{College}|H)\mathbb{P}(H| \text{George}) + \mathbb{P}(\text{College}|L)\mathbb{P}(L| \text{George}) = (3/4)(1/4) + (1/4)(3/4)$

Table 2.3: Example Distribution 2

	\$8	\$10	\$12	\$20	Mean
High-School Graduate + High Test Score	1	3	0	0	\$9.50
College Graduate + High Test Score	0	0	3	9	\$18.00
High-School Graduate + Low Test Score	9	3	0	0	\$8.50
College Graduate + Low Test Score	0	0	3	1	\$14.00

The ACE varies between these two observable groups (those with high test scores and those with low test scores). Again, we would hope to be able to learn the ACE from a regression analysis, this time from a regression of wages on education and test scores.

To see this in the wage distribution, suppose that the econometrician collects data on the aptitude test score as well as education and wages. Given a random sample of 32 individuals we would expect to find the wage distribution in Table 2.3.

Define the dummy variable *highscore* which takes the value 1 for students who received a high test score, else zero. The regression of wages on college attendance and test scores (with interactions) takes the form

$$\mathbb{E}(\text{wage} | \text{college}, \text{highscore}) = 1.00\text{highscore} + 5.50\text{college} + 3.00\text{highscore} \times \text{college} + 8.50.$$

The coefficient on *college*, \$5.50, is the regression derivative of college attendance for those with low test scores, and the sum of this coefficient with the interaction coefficient, \$8.50, is the regression derivative for college attendance for those with high test scores. These equal the average causal effect as calculated above. Furthermore, since 1/2 of the population achieves a high test score and 1/2 achieve a low test score, the measured average causal effect in the entire population is \$7, which precisely equals the true value.

In this example, by conditioning on the aptitude test score, the average causal effect of education on wages can be learned from a regression analysis. What this shows is that by conditioning on the proper variables, it may be possible to achieve the CIA, in which case regression analysis measures average causal effects.

### 2.31 Expectation: Mathematical Details\*

We define the **mean** or **expectation**  $\mathbb{E}(y)$  of a random variable  $y$  as follows. If  $y$  is discrete on the set  $\{\tau_1, \tau_2, \dots\}$  then

$$\mathbb{E}(y) = \sum_{j=1}^{\infty} \tau_j \mathbb{P}(y = \tau_j),$$

and if  $y$  is continuous with density  $f$  then

$$\mathbb{E}(y) = \int_{-\infty}^{\infty} y f(y) dy.$$

We can unify these definitions by writing the expectation as the Lebesgue integral with respect to the distribution function  $F$

$$\mathbb{E}(y) = \int_{-\infty}^{\infty} y dF(y).$$

In the event that the above integral is not finite, separately evaluate the two integrals

$$I_1 = \int_0^{\infty} y dF(y) \tag{2.53}$$

$$I_2 = - \int_{-\infty}^0 y dF(y). \tag{2.54}$$

If  $I_1 = \infty$  and  $I_2 < \infty$  then it is typical to define  $\mathbb{E}(y) = \infty$ . If  $I_1 < \infty$  and  $I_2 = \infty$  then we define  $\mathbb{E}(y) = -\infty$ . However, if both  $I_1 = \infty$  and  $I_2 = \infty$  then  $\mathbb{E}(y)$  is undefined. If

$$\mathbb{E}|y| = \int_{-\infty}^{\infty} |y| dF(y) = I_1 + I_2 < \infty$$

then  $\mathbb{E}(y)$  exists and is finite. In this case it is common to say that the mean  $\mathbb{E}(y)$  is “well-defined”.

More generally,  $y$  has a finite  $r^{th}$  moment if

$$\mathbb{E}|y|^r < \infty. \quad (2.55)$$

By Liapunov's Inequality (B.34), (2.55) implies  $\mathbb{E}|y|^s < \infty$  for all  $1 \leq s \leq r$ . Thus, for example, if the fourth moment is finite then the first, second and third moments are also finite, and so is the 3.9<sup>th</sup> moment.

It is common in econometric theory to assume that the variables, or certain transformations of the variables, have finite moments of a certain order. How should we interpret this assumption? How restrictive is it?

One way to visualize the importance is to consider the class of Pareto densities given by

$$f(y) = ay^{-a-1}, \quad y > 1.$$

The parameter  $a$  of the Pareto distribution indexes the rate of decay of the tail of the density. Larger  $a$  means that the tail declines to zero more quickly. See Figure 2.11 below where we plot the Pareto density for  $a = 1$  and  $a = 2$ . The parameter  $a$  also determines which moments are finite. We can calculate that

$$\mathbb{E}|y|^r = \begin{cases} a \int_1^\infty y^{r-a-1} dy = \frac{a}{a-r} & \text{if } r < a \\ \infty & \text{if } r \geq a. \end{cases}$$

This shows that if  $y$  is Pareto distributed with parameter  $a$ , then the  $r^{th}$  moment of  $y$  is finite if and only if  $r < a$ . Higher  $a$  means higher finite moments. Equivalently, the faster the tail of the density declines to zero, the more moments are finite.

This connection between tail decay and finite moments is not limited to the Pareto distribution. We can make a similar analysis using a tail bound. Suppose that  $y$  has density  $f(y)$  which satisfies the bound  $f(y) \leq A|y|^{-a-1}$  for some  $A < \infty$  and  $a > 0$ . Since  $f(y)$  is bounded below a scale of a Pareto density, its tail behavior is similarly bounded. This means that for  $r < a$

$$\mathbb{E}|y|^r = \int_{-\infty}^{\infty} |y|^r f(y) dy \leq \int_{-1}^1 f(y) dy + 2A \int_1^\infty y^{r-a-1} dy \leq 1 + \frac{2A}{a-r} < \infty.$$

Thus if the tail of the density declines at the rate  $|y|^{-a-1}$  or faster, then  $y$  has finite moments up to (but not including)  $a$ . Broadly speaking, the restriction that  $y$  has a finite  $r^{th}$  moment means that the tail of  $y$ 's density declines to zero faster than  $y^{-r-1}$ . The faster decline of the tail means that the probability of observing an extreme value of  $y$  is a more rare event.

We complete this section by adding an alternative representation of expectation in terms of the distribution function.

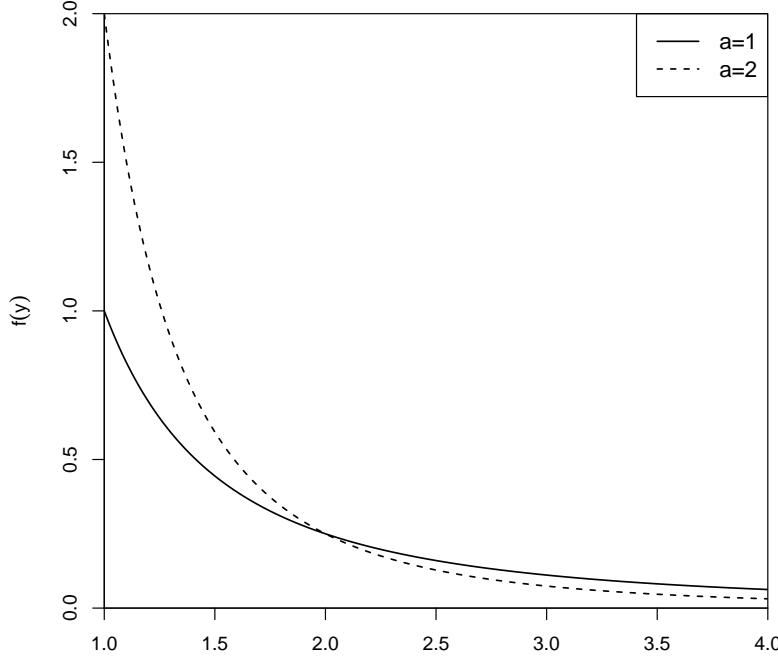
**Theorem 2.12** For any non-negative random variable  $y$

$$\mathbb{E}(y) = \int_0^\infty \mathbb{P}(y > u) du$$

**Proof of Theorem 2.12:** Let  $F^*(x) = \mathbb{P}(y > x) = 1 - F(x)$ , where  $F(x)$  is the distribution function. By integration by parts

$$\mathbb{E}(y) = \int_0^\infty y dF(y) = - \int_0^\infty y dF^*(y) = -[yF^*(y)]_0^\infty + \int_0^\infty F^*(y) dy = \int_0^\infty \mathbb{P}(y > u) du$$

as stated. ■

Figure 2.11: Pareto Densities,  $a = 1$  and  $a = 2$ 

## 2.32 Moment Generating and Characteristic Functions\*

For a random variable  $y$  with distribution  $F$  its **moment generating function** (MGF) is

$$M(t) = \mathbb{E}(\exp(ty)) = \int \exp(ty) dF(y). \quad (2.56)$$

This is also known as the Laplace transformation of the density of  $y$ . The MGF is a function of the argument  $t$ , and is an alternative representation of the distribution  $F$ . It is called the moment generating function since the  $r^{th}$  derivative evaluated at zero is the  $r^{th}$  uncentered moment. Indeed,

$$M^{(r)}(t) = \mathbb{E}\left(\frac{d^r}{dt^r} \exp(ty)\right) = \mathbb{E}(y^r \exp(ty))$$

and thus the  $r^{th}$  derivative at  $t = 0$  is

$$M^{(r)}(0) = \mathbb{E}(y^r).$$

A major limitation with the MGF is that it does not exist for many random variables. Essentially, existence of the integral (2.56) requires the tail of the density of  $y$  to decline exponentially. This excludes thick-tailed distributions such as the Pareto.

This limitation is removed if we consider the **characteristic function** (CF) of  $y$ , which is defined as

$$C(t) = \mathbb{E}(\exp(ity)) = \int \exp(ity) dF(y)$$

where  $i = \sqrt{-1}$ . Like the MGF, the CF is a function of its argument  $t$  and is a representation of the distribution function  $F$ . The CF is also known as the Fourier transformation of the density of  $y$ . Unlike the MGF, the CF exists for all random variables and all values of  $t$  since  $\exp(ity) = \cos(ty) + i\sin(ty)$  is bounded.

Similarly to the MGF, the  $r^{th}$  derivative of the characteristic function evaluated at zero takes the simple form

$$C^{(r)}(0) = i^r \mathbb{E}(y^r)$$

when such expectations exist. A further connection is that the  $r^{th}$  moment is finite if and only if  $C^{(r)}(t)$  is continuous at zero.

For random vectors  $\mathbf{y}$  with distribution  $F$  we define the multivariate MGF as

$$M(\mathbf{t}) = \mathbb{E}(\exp(\mathbf{t}' \mathbf{y})) = \int \exp(\mathbf{t}' \mathbf{y}) dF(\mathbf{y})$$

when it exists. Similarly, we define the multivariate CF as

$$C(\mathbf{t}) = \mathbb{E}(\exp(i\mathbf{t}' \mathbf{y})) = \int \exp(i\mathbf{t}' \mathbf{y}) dF(\mathbf{y}).$$

### 2.33 Moments and Cumulants\*

For a random variable  $y$  it is common to define its  $r^{th}$  **moment** as

$$\mu'_r = \mathbb{E}(y^r).$$

For example, the mean is the  $1^{st}$  moment,  $\mu = \mu'_1$ . As described in Section 2.32, the moments can be expressed in terms of the derivatives of the moment generating function, that is  $\mu'_r = M^{(r)}(0)$ .

We define the  $r^{th}$  **central moment** as

$$\mu_r = \mathbb{E}(y - \mathbb{E}(y))^r.$$

Note  $\sigma^2 = \mu_2$ .

The **cumulant generating function** is the natural log of the moment generating function

$$K(t) = \log M(t).$$

Since  $M(0) = 1$  we see  $K(0) = 0$ . Expanding as a power series we obtain

$$K(t) = \sum_{r=1}^{\infty} \kappa_r \frac{t^r}{r!}$$

where

$$\kappa_r = K^{(r)}(0)$$

is the  $r^{th}$  derivative of  $K(t)$ , evaluated at  $t = 0$ . The constants  $\kappa_r$  are known as the **cumulants** of the distribution of  $y$ .

The cumulants are related to the central moments. We can calculate that

$$\begin{aligned} K^{(1)}(t) &= \frac{M^{(1)}(t)}{M(t)} \\ K^{(2)}(t) &= \frac{M^{(2)}(t)}{M(t)} - \left( \frac{M^{(1)}(t)}{M(t)} \right)^2 \end{aligned}$$

so  $\kappa_1 = \mu$  and  $\kappa_2 = \mu'_2 - \mu^2 = \mu_2$ . The first six cumulants are as follows.

$$\kappa_1 = \mu$$

$$\kappa_2 = \mu_2$$

$$\kappa_3 = \mu_3$$

$$\kappa_4 = \mu_4 - 3\mu_2^2$$

$$\kappa_5 = \mu_5 - 10\mu_3\mu_2$$

$$\kappa_6 = \mu_6 - 15\mu_4\mu_2 - 10\mu_3^2 + 30\mu_2^3.$$

We see that the first three cumulants correspond to the central moments, but higher cumulants are polynomial functions of the central moments.

Inverting, we can also express the central moments in terms of the cumulants, for example, the 4<sup>th</sup> through 6<sup>th</sup> are as follows.

$$\begin{aligned}\mu_4 &= \kappa_4 + 3\kappa_2^2 \\ \mu_5 &= \kappa_5 + 10\kappa_3\kappa_2 \\ \mu_6 &= \kappa_6 + 15\kappa_4\kappa_2 + 10\kappa_3^2 + 15\kappa_2^3.\end{aligned}$$

## 2.34 Existence and Uniqueness of the Conditional Expectation\*

In Sections 2.3 and 2.6 we defined the conditional mean when the conditioning variables  $\mathbf{x}$  are discrete and when the variables  $(y, \mathbf{x})$  have a joint density. We have explored these cases because these are the situations where the conditional mean is easiest to describe and understand. However, the conditional mean exists quite generally without appealing to the properties of either discrete or continuous random variables.

To justify this claim we now present a deep result from probability theory. What it says is that the conditional mean exists for all joint distributions  $(y, \mathbf{x})$  for which  $y$  has a finite mean.

**Theorem 2.13 Existence of the Conditional Mean**

If  $\mathbb{E}|y| < \infty$  then there exists a function  $m(\mathbf{x})$  such that for all sets  $\mathcal{X}$  for which  $\mathbb{P}(\mathbf{x} \in \mathcal{X})$  is defined,

$$\mathbb{E}(1_{(\mathbf{x} \in \mathcal{X})} y) = \mathbb{E}(1_{(\mathbf{x} \in \mathcal{X})} m(\mathbf{x})). \quad (2.57)$$

The function  $m(\mathbf{x})$  is almost everywhere unique, in the sense that if  $h(\mathbf{x})$  satisfies (2.57), then there is a set  $S$  such that  $\mathbb{P}(S) = 1$  and  $m(\mathbf{x}) = h(\mathbf{x})$  for  $\mathbf{x} \in S$ . The function  $m(\mathbf{x})$  is called the **conditional mean** and is written  $m(\mathbf{x}) = \mathbb{E}(y | \mathbf{x})$ .

See, for example, Ash (1972), Theorem 6.3.3.

The conditional mean  $m(\mathbf{x})$  defined by (2.57) specializes to (2.4) when  $(y, \mathbf{x})$  have a joint density. The usefulness of definition (2.57) is that Theorem 2.13 shows that the conditional mean  $m(\mathbf{x})$  exists for all finite-mean distributions. This definition allows  $y$  to be discrete or continuous, for  $\mathbf{x}$  to be scalar or vector-valued, and for the components of  $\mathbf{x}$  to be discrete or continuously distributed.

You may have noticed that Theorem 2.13 applies only to sets  $\mathcal{X}$  for which  $\mathbb{P}(\mathbf{x} \in \mathcal{X})$  is defined. This is a technical issue –measurability – which we largely side-step in this textbook. Formal probability theory only applies to sets which are measurable – for which probabilities are defined – as it turns out that not all sets satisfy measurability. This is not a practical concern for applications, so we defer such distinctions for formal theoretical treatments.

## 2.35 Identification\*

A critical and important issue in structural econometric modeling is identification, meaning that a parameter is uniquely determined by the distribution of the observed variables. It is relatively straightforward in the context of the unconditional and conditional mean, but it is worthwhile to introduce and explore the concept at this point for clarity.

Let  $F$  denote the distribution of the observed data, for example the distribution of the pair  $(y, \mathbf{x})$ . Let  $\mathcal{F}$  be a collection of distributions  $F$ . Let  $\theta$  be a parameter of interest (for example, the mean  $\mathbb{E}(y)$ ).

**Definition 2.9** A parameter  $\theta \in \mathbb{R}$  is identified on  $\mathcal{F}$  if for all  $F \in \mathcal{F}$ , there is a uniquely determined value of  $\theta$ .

Equivalently,  $\theta$  is identified if we can write it as a mapping  $\theta = g(F)$  on the set  $\mathcal{F}$ . The restriction to the set  $\mathcal{F}$  is important. Most parameters are identified only on a strict subset of the space of all distributions.

Take, for example, the mean  $\mu = \mathbb{E}(y)$ . It is uniquely determined if  $\mathbb{E}|y| < \infty$ , so it is clear that  $\mu$  is identified for the set  $\mathcal{F} = \{F : \int_{-\infty}^{\infty} |y| dF(y) < \infty\}$ . However,  $\mu$  is also well defined when it is either positive or negative infinity. Hence, defining  $I_1$  and  $I_2$  as in (2.53) and (2.54), we can deduce that  $\mu$  is identified on the set  $\mathcal{F} = \{F : \{I_1 < \infty\} \cup \{I_2 < \infty\}\}$ .

Next, consider the conditional mean. Theorem 2.13 demonstrates that  $\mathbb{E}|y| < \infty$  is a sufficient condition for identification.

**Theorem 2.14 Identification of the Conditional Mean**

If  $\mathbb{E}|y| < \infty$ , the conditional mean  $m(\mathbf{x}) = \mathbb{E}(y | \mathbf{x})$  is identified almost everywhere.

It might seem as if identification is a general property for parameters, so long as we exclude degenerate cases. This is true for moments of observed data, but not necessarily for more complicated models. As a case in point, consider the context of censoring. Let  $y$  be a random variable with distribution  $F$ . Instead of observing  $y$ , we observe  $y^*$  defined by the censoring rule

$$y^* = \begin{cases} y & \text{if } y \leq \tau \\ \tau & \text{if } y > \tau \end{cases} .$$

That is,  $y^*$  is capped at the value  $\tau$ . A common example is income surveys, where income responses are “top-coded”, meaning that incomes above the top code  $\tau$  are recorded as the top code. The observed variable  $y^*$  has distribution

$$F^*(u) = \begin{cases} F(u) & \text{for } u \leq \tau \\ 1 & \text{for } u \geq \tau. \end{cases}$$

We are interested in features of the distribution  $F$  not the censored distribution  $F^*$ . For example, we are interested in the mean wage  $\mu = \mathbb{E}(y)$ . The difficulty is that we cannot calculate  $\mu$  from  $F^*$  except in the trivial case where there is no censoring  $\mathbb{P}(y \geq \tau) = 0$ . Thus the mean  $\mu$  is not generically identified from the censored distribution.

A typical solution to the identification problem is to assume a parametric distribution. For example, let  $\mathcal{F}$  be the set of normal distributions  $y \sim N(\mu, \sigma^2)$ . It is possible to show that the parameters  $(\mu, \sigma^2)$  are identified for all  $F \in \mathcal{F}$ . That is, if we know that the uncensored distribution is normal, we can uniquely determine the parameters from the censored distribution. This is often called **parametric identification** as identification is restricted to a parametric class of distributions. In modern econometrics this is generally viewed as a second-best solution, as identification has been achieved only through the use of an arbitrary and unverifiable parametric assumption.

A pessimistic conclusion might be that it is impossible to identify parameters of interest from censored data without parametric assumptions. Interestingly, this pessimism is unwarranted. It turns out that we can identify the quantiles  $q_\alpha$  of  $F$  for  $\alpha \leq \mathbb{P}(y \leq \tau)$ . For example, if 20% of the distribution is censored, we can identify all quantiles for  $\alpha \in (0, 0.8)$ . This is often called **nonparametric identification** as the parameters are identified without restriction to a parametric class.

What we have learned from this little exercise is that in the context of censored data, moments can only be parametrically identified, while non-censored quantiles are nonparametrically identified. Part of the message is that a study of identification can help focus attention on what can be learned from the data distributions available.

## 2.36 Technical Proofs\*

**Proof of Theorem 2.1:** For convenience, assume that the variables have a joint density  $f(y, \mathbf{x})$ . Since  $\mathbb{E}(y | \mathbf{x})$  is a function of the random vector  $\mathbf{x}$  only, to calculate its expectation we integrate with respect to the density  $f_{\mathbf{x}}(\mathbf{x})$  of  $\mathbf{x}$ , that is

$$\mathbb{E}(\mathbb{E}(y | \mathbf{x})) = \int_{\mathbb{R}^k} \mathbb{E}(y | \mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}.$$

Substituting in (2.4) and noting that  $f_{y|\mathbf{x}}(y|\mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) = f(y, \mathbf{x})$ , we find that the above expression equals

$$\int_{\mathbb{R}^k} \left( \int_{\mathbb{R}} y f_{y|\mathbf{x}}(y|\mathbf{x}) dy \right) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} = \int_{\mathbb{R}^k} \int_{\mathbb{R}} y f(y, \mathbf{x}) dy d\mathbf{x} = \mathbb{E}(y)$$

the unconditional mean of  $y$ . ■

**Proof of Theorem 2.2:** Again assume that the variables have a joint density. It is useful to observe that

$$f(y|\mathbf{x}_1, \mathbf{x}_2) f(\mathbf{x}_2|\mathbf{x}_1) = \frac{f(y, \mathbf{x}_1, \mathbf{x}_2)}{f(\mathbf{x}_1, \mathbf{x}_2)} \frac{f(\mathbf{x}_1, \mathbf{x}_2)}{f(\mathbf{x}_1)} = f(y, \mathbf{x}_2|\mathbf{x}_1), \quad (2.58)$$

the density of  $(y, \mathbf{x}_2)$  given  $\mathbf{x}_1$ . Here, we have abused notation and used a single symbol  $f$  to denote the various unconditional and conditional densities to reduce notational clutter.

Note that

$$\mathbb{E}(y | \mathbf{x}_1, \mathbf{x}_2) = \int_{\mathbb{R}} y f(y|\mathbf{x}_1, \mathbf{x}_2) dy. \quad (2.59)$$

Integrating (2.59) with respect to the conditional density of  $\mathbf{x}_2$  given  $\mathbf{x}_1$ , and applying (2.58) we find that

$$\begin{aligned} \mathbb{E}(\mathbb{E}(y | \mathbf{x}_1, \mathbf{x}_2) | \mathbf{x}_1) &= \int_{\mathbb{R}^{k_2}} \mathbb{E}(y | \mathbf{x}_1, \mathbf{x}_2) f(\mathbf{x}_2|\mathbf{x}_1) d\mathbf{x}_2 \\ &= \int_{\mathbb{R}^{k_2}} \left( \int_{\mathbb{R}} y f(y|\mathbf{x}_1, \mathbf{x}_2) dy \right) f(\mathbf{x}_2|\mathbf{x}_1) d\mathbf{x}_2 \\ &= \int_{\mathbb{R}^{k_2}} \int_{\mathbb{R}} y f(y|\mathbf{x}_1, \mathbf{x}_2) f(\mathbf{x}_2|\mathbf{x}_1) dy d\mathbf{x}_2 \\ &= \int_{\mathbb{R}^{k_2}} \int_{\mathbb{R}} y f(y, \mathbf{x}_2|\mathbf{x}_1) dy d\mathbf{x}_2 \\ &= \mathbb{E}(y | \mathbf{x}_1) \end{aligned}$$

as stated. ■

**Proof of Theorem 2.3:**

$$\mathbb{E}(g(\mathbf{x}) y | \mathbf{x}) = \int_{\mathbb{R}} g(\mathbf{x}) y f_{y|\mathbf{x}}(y|\mathbf{x}) dy = g(\mathbf{x}) \int_{\mathbb{R}} y f_{y|\mathbf{x}}(y|\mathbf{x}) dy = g(\mathbf{x}) \mathbb{E}(y | \mathbf{x})$$

This is (2.5). Equation (2.6) follows by applying the simple law of iterated expectations (Theorem 2.1) to (2.5). ■

**Proof of Theorem 2.4.** Applying Minkowski's inequality (B.33) to  $e = y - m(\mathbf{x})$ ,

$$(\mathbb{E}|e|^r)^{1/r} = (\mathbb{E}|y - m(\mathbf{x})|^r)^{1/r} \leq (\mathbb{E}|y|^r)^{1/r} + (\mathbb{E}|m(\mathbf{x})|^r)^{1/r} < \infty,$$

where the two parts on the right-hand are finite since  $\mathbb{E}|y|^r < \infty$  by assumption and  $\mathbb{E}|m(\mathbf{x})|^r < \infty$  by the conditional expectation inequality (B.28). The fact that  $(\mathbb{E}|e|^r)^{1/r} < \infty$  implies  $\mathbb{E}|e|^r < \infty$ . ■

**Proof of Theorem 2.6:** The assumption that  $\mathbb{E}(y^2) < \infty$  implies that all the conditional expectations below exist.

Using the law of iterated expectations (Theorem 2.2)  $\mathbb{E}(y | \mathbf{x}_1) = \mathbb{E}(\mathbb{E}(y | \mathbf{x}_1, \mathbf{x}_2) | \mathbf{x}_1)$  and the conditional Jensen's inequality (B.27),

$$(\mathbb{E}(y | \mathbf{x}_1))^2 = (\mathbb{E}(\mathbb{E}(y | \mathbf{x}_1, \mathbf{x}_2) | \mathbf{x}_1))^2 \leq \mathbb{E}((\mathbb{E}(y | \mathbf{x}_1, \mathbf{x}_2))^2 | \mathbf{x}_1).$$

Taking unconditional expectations, this implies

$$\mathbb{E}((\mathbb{E}(y | \mathbf{x}_1))^2) \leq \mathbb{E}((\mathbb{E}(y | \mathbf{x}_1, \mathbf{x}_2))^2).$$

Similarly,

$$(\mathbb{E}(y))^2 \leq \mathbb{E}((\mathbb{E}(y | \mathbf{x}_1))^2) \leq \mathbb{E}((\mathbb{E}(y | \mathbf{x}_1, \mathbf{x}_2))^2). \quad (2.60)$$

The variables  $y$ ,  $\mathbb{E}(y | \mathbf{x}_1)$  and  $\mathbb{E}(y | \mathbf{x}_1, \mathbf{x}_2)$  all have the same mean  $\mathbb{E}(y)$ , so the inequality (2.60) implies that the variances are ranked monotonically:

$$0 \leq \text{var}(\mathbb{E}(y | \mathbf{x}_1)) \leq \text{var}(\mathbb{E}(y | \mathbf{x}_1, \mathbf{x}_2)). \quad (2.61)$$

Define  $e = y - \mathbb{E}(y | \mathbf{x})$  and  $u = \mathbb{E}(y | \mathbf{x}) - \mu$  so that we have the decomposition

$$y - \mu = e + u.$$

Notice  $\mathbb{E}(e | \mathbf{x}) = 0$  and  $u$  is a function of  $\mathbf{x}$ . Thus by the conditioning theorem (Theorem 2.3),  $\mathbb{E}(eu) = 0$  so  $e$  and  $u$  are uncorrelated. It follows that

$$\text{var}(y) = \text{var}(e) + \text{var}(u) = \text{var}(y - \mathbb{E}(y | \mathbf{x})) + \text{var}(\mathbb{E}(y | \mathbf{x})). \quad (2.62)$$

The monotonicity of the variances of the conditional mean (2.61) applied to the variance decomposition (2.62) implies the reverse monotonicity of the variances of the differences, completing the proof. ■

**Proof of Theorem 2.8.** For part 1, by the expectation inequality (B.29), (A.16) and Assumption 2.1,

$$\|\mathbb{E}(\mathbf{x}\mathbf{x}')\| \leq \mathbb{E}\|\mathbf{x}\mathbf{x}'\| = \mathbb{E}(\|\mathbf{x}\|^2) < \infty.$$

Similarly, using the expectation inequality (B.29), the Cauchy-Schwarz inequality (B.31) and Assumption 2.1,

$$\|\mathbb{E}(\mathbf{x}y)\| \leq \mathbb{E}\|\mathbf{x}y\| \leq (\mathbb{E}(\|\mathbf{x}\|^2))^{1/2} (\mathbb{E}(y^2))^{1/2} < \infty.$$

Thus the moments  $\mathbb{E}(\mathbf{x}y)$  and  $\mathbb{E}(\mathbf{x}\mathbf{x}')$  are finite and well defined.

For part 2, the coefficient  $\boldsymbol{\beta} = (\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-1} \mathbb{E}(\mathbf{x}y)$  is well defined since  $(\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-1}$  exists under Assumption 2.1.

Part 3 follows from Definition 2.5 and part 2.

For part 4, first note that

$$\begin{aligned} \mathbb{E}(e^2) &= \mathbb{E}((y - \mathbf{x}'\boldsymbol{\beta})^2) \\ &= \mathbb{E}(y^2) - 2\mathbb{E}(yx')\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbb{E}(\mathbf{x}\mathbf{x}')\boldsymbol{\beta} \\ &= \mathbb{E}(y^2) - 2\mathbb{E}(yx')(\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-1}\mathbb{E}(\mathbf{x}y) \\ &\leq \mathbb{E}(y^2) \\ &< \infty. \end{aligned}$$

The first inequality holds because  $\mathbb{E}(yx')(\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-1}\mathbb{E}(\mathbf{x}y)$  is a quadratic form and therefore necessarily non-negative. Second, by the expectation inequality (B.29), the Cauchy-Schwarz inequality (B.31) and Assumption 2.1,

$$\|\mathbb{E}(\mathbf{x}e)\| \leq \mathbb{E}\|\mathbf{x}e\| = (\mathbb{E}(\|\mathbf{x}\|^2))^{1/2} (\mathbb{E}(e^2))^{1/2} < \infty.$$

It follows that the expectation  $\mathbb{E}(\mathbf{x}e)$  is finite, and is zero by the calculation (2.24).

For part 6, Applying Minkowski's inequality (B.33) to  $e = y - \mathbf{x}'\boldsymbol{\beta}$ ,

$$\begin{aligned} (\mathbb{E}|e|^r)^{1/r} &= (\mathbb{E}|y - \mathbf{x}'\boldsymbol{\beta}|^r)^{1/r} \\ &\leq (\mathbb{E}|y|^r)^{1/r} + (\mathbb{E}|\mathbf{x}'\boldsymbol{\beta}|^r)^{1/r} \\ &\leq (\mathbb{E}|y|^r)^{1/r} + (\mathbb{E}\|\mathbf{x}\|^r)^{1/r} \|\boldsymbol{\beta}\| \\ &< \infty, \end{aligned}$$

the final inequality by assumption. ■

## Exercises

**Exercise 2.1** Find  $\mathbb{E}(\mathbb{E}(y | \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) | \mathbf{x}_1, \mathbf{x}_2) | \mathbf{x}_1$ .

**Exercise 2.2** If  $\mathbb{E}(y | x) = a + bx$ , find  $\mathbb{E}(yx)$  as a function of moments of  $x$ .

**Exercise 2.3** Prove Theorem 2.4.4 using the law of iterated expectations.

**Exercise 2.4** Suppose that the random variables  $y$  and  $x$  only take the values 0 and 1, and have the following joint probability distribution

	$x = 0$	$x = 1$
$y = 0$	.1	.2
$y = 1$	.4	.3

Find  $\mathbb{E}(y | x)$ ,  $\mathbb{E}(y^2 | x)$  and  $\text{var}(y | x)$  for  $x = 0$  and  $x = 1$ .

**Exercise 2.5** Show that  $\sigma^2(\mathbf{x})$  is the best predictor of  $e^2$  given  $\mathbf{x}$ :

- (a) Write down the mean-squared error of a predictor  $h(\mathbf{x})$  for  $e^2$ .
- (b) What does it mean to be predicting  $e^2$ ?
- (c) Show that  $\sigma^2(\mathbf{x})$  minimizes the mean-squared error and is thus the best predictor.

**Exercise 2.6** Use  $y = m(\mathbf{x}) + e$  to show that

$$\text{var}(y) = \text{var}(m(\mathbf{x})) + \sigma^2$$

**Exercise 2.7** Show that the conditional variance can be written as

$$\sigma^2(\mathbf{x}) = \mathbb{E}(y^2 | \mathbf{x}) - (\mathbb{E}(y | \mathbf{x}))^2.$$

**Exercise 2.8** Suppose that  $y$  is discrete-valued, taking values only on the non-negative integers, and the conditional distribution of  $y$  given  $\mathbf{x}$  is Poisson:

$$\mathbb{P}(y = j | \mathbf{x}) = \frac{\exp(-\mathbf{x}'\boldsymbol{\beta})(\mathbf{x}'\boldsymbol{\beta})^j}{j!}, \quad j = 0, 1, 2, \dots$$

Compute  $\mathbb{E}(y | \mathbf{x})$  and  $\text{var}(y | \mathbf{x})$ . Does this justify a linear regression model of the form  $y = \mathbf{x}'\boldsymbol{\beta} + e$ ?

Hint: If  $\mathbb{P}(y = j) = \frac{\exp(-\lambda)\lambda^j}{j!}$ , then  $\mathbb{E}(y) = \lambda$  and  $\text{var}(y) = \lambda$ .

**Exercise 2.9** Suppose you have two regressors:  $x_1$  is binary (takes values 0 and 1) and  $x_2$  is categorical with 3 categories ( $A, B, C$ ). Write  $\mathbb{E}(y | x_1, x_2)$  as a linear regression.

**Exercise 2.10** True or False. If  $y = x\boldsymbol{\beta} + e$ ,  $x \in \mathbb{R}$ , and  $\mathbb{E}(e | x) = 0$ , then  $\mathbb{E}(x^2 e) = 0$ .

**Exercise 2.11** True or False. If  $y = x\boldsymbol{\beta} + e$ ,  $x \in \mathbb{R}$ , and  $\mathbb{E}(xe) = 0$ , then  $\mathbb{E}(x^2 e) = 0$ .

**Exercise 2.12** True or False. If  $y = \mathbf{x}'\boldsymbol{\beta} + e$  and  $\mathbb{E}(e | \mathbf{x}) = 0$ , then  $e$  is independent of  $\mathbf{x}$ .

**Exercise 2.13** True or False. If  $y = \mathbf{x}'\boldsymbol{\beta} + e$  and  $\mathbb{E}(xe) = \mathbf{0}$ , then  $\mathbb{E}(e | \mathbf{x}) = 0$ .

**Exercise 2.14** True or False. If  $y = \mathbf{x}'\boldsymbol{\beta} + e$ ,  $\mathbb{E}(e | \mathbf{x}) = 0$ , and  $\mathbb{E}(e^2 | \mathbf{x}) = \sigma^2$ , a constant, then  $e$  is independent of  $\mathbf{x}$ .

**Exercise 2.15** Consider the intercept-only model  $y = \alpha + e$  defined as the best linear predictor. Show that  $\alpha = \mathbb{E}(y)$ .

**Exercise 2.16** Let  $x$  and  $y$  have the joint density  $f(x, y) = \frac{3}{2}(x^2 + y^2)$  on  $0 \leq x \leq 1, 0 \leq y \leq 1$ . Compute the coefficients of the best linear predictor  $y = \alpha + \beta x + e$ . Compute the conditional mean  $m(x) = \mathbb{E}(y | x)$ . Are the best linear predictor and conditional mean different?

**Exercise 2.17** Let  $x$  be a random variable with  $\mu = \mathbb{E}(x)$  and  $\sigma^2 = \text{var}(x)$ . Define

$$g(x | \mu, \sigma^2) = \begin{pmatrix} x - \mu \\ (x - \mu)^2 - \sigma^2 \end{pmatrix}.$$

Show that  $\mathbb{E}g(x | m, s) = 0$  if and only if  $m = \mu$  and  $s = \sigma^2$ .

**Exercise 2.18** Suppose that

$$\mathbf{x} = \begin{pmatrix} 1 \\ x_2 \\ x_3 \end{pmatrix}$$

and  $x_3 = \alpha_1 + \alpha_2 x_2$  is a linear function of  $x_2$ .

- (a) Show that  $\mathbf{Q}_{\mathbf{x}\mathbf{x}} = \mathbb{E}(\mathbf{x}\mathbf{x}')$  is not invertible.
- (b) Use a linear transformation of  $\mathbf{x}$  to find an expression for the best linear predictor of  $y$  given  $\mathbf{x}$ . (Be explicit, do not just use the generalized inverse formula.)

**Exercise 2.19** Show (2.45)-(2.46), namely that for

$$d(\boldsymbol{\beta}) = \mathbb{E}(m(\mathbf{x}) - \mathbf{x}'\boldsymbol{\beta})^2$$

then

$$\begin{aligned} \boldsymbol{\beta} &= \underset{\boldsymbol{b} \in \mathbb{R}^k}{\operatorname{argmin}} d(\boldsymbol{b}) \\ &= (\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-1} \mathbb{E}(\mathbf{x}m(\mathbf{x})) \\ &= (\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-1} \mathbb{E}(\mathbf{xy}). \end{aligned}$$

Hint: To show  $\mathbb{E}(\mathbf{x}m(\mathbf{x})) = \mathbb{E}(\mathbf{xy})$  use the law of iterated expectations.

**Exercise 2.20** Verify that (2.57) holds with  $m(\mathbf{x})$  defined in (2.4) when  $(y, \mathbf{x})$  have a joint density  $f(y, \mathbf{x})$ .

**Exercise 2.21** Consider the short and long projections

$$y = x\gamma_1 + e$$

$$y = x\beta_1 + x^2\beta_2 + u$$

- (a) Under what condition does  $\gamma_1 = \beta_1$ ?
- (b) Now suppose the long projection is

$$y = x\theta_1 + x^3\theta_2 + v$$

Is there a similar condition under which  $\gamma_1 = \theta_1$ ?

**Exercise 2.22** Take the homoskedastic model

$$\begin{aligned} y &= \mathbf{x}'_1 \boldsymbol{\beta}_1 + \mathbf{x}'_2 \boldsymbol{\beta}_2 + e \\ \mathbb{E}(e | \mathbf{x}_1, \mathbf{x}_2) &= 0 \\ \mathbb{E}(e^2 | \mathbf{x}_1, \mathbf{x}_2) &= \sigma^2 \\ \mathbb{E}(\mathbf{x}_2 | \mathbf{x}_1) &= \boldsymbol{\Gamma} \mathbf{x}_1 \\ \boldsymbol{\Gamma} &\neq 0 \end{aligned}$$

Suppose the parameter  $\boldsymbol{\beta}_1$  is of interest. We know that the exclusion of  $\mathbf{x}_2$  creates omitted variable bias in the projection coefficient on  $\mathbf{x}_2$ . It also changes the equation error. Our question is: what is the effect on the homoskedasticity property of the induced equation error? Does the exclusion of  $\mathbf{x}_2$  induce heteroskedasticity or not? Be specific.

# Chapter 3

## The Algebra of Least Squares

### 3.1 Introduction

In this chapter we introduce the popular least-squares estimator. Most of the discussion will be algebraic, with questions of distribution and inference deferred to later chapters.

### 3.2 Samples

In Section 2.18 we derived and discussed the best linear predictor of  $y$  given  $\mathbf{x}$  for a pair of random variables  $(y, \mathbf{x}) \in \mathbb{R} \times \mathbb{R}^k$ , and called this the linear projection model. We are now interested in **estimating** the parameters of this model, in particular the projection coefficient

$$\boldsymbol{\beta} = (\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-1} \mathbb{E}(\mathbf{x}y). \quad (3.1)$$

We can estimate  $\boldsymbol{\beta}$  from observational data which includes joint measurements on the variables  $(y, \mathbf{x})$ . For example, supposing we are interested in estimating a wage equation, we would use a dataset with observations on wages (or weekly earnings), education, experience (or age), and demographic characteristics (gender, race, location). One possible dataset is the Current Population Survey (CPS), a survey of U.S. households which includes questions on employment, income, education, and demographic characteristics.

Notationally we wish to distinguish observations from the underlying random variables. The convention in econometrics is to denote observations by appending a subscript  $i$  which runs from 1 to  $n$ , thus the  $i^{th}$  observation is  $(y_i, \mathbf{x}_i)$ , and  $n$  denotes the sample size. The dataset is then  $\{(y_i, \mathbf{x}_i); i = 1, \dots, n\}$ . We call this the **sample** or the **observations**.

From the viewpoint of empirical analysis, a dataset is an array of numbers often organized as a table, where the columns of the table correspond to distinct variables and the rows correspond to distinct observations. For empirical analysis, the dataset and observations are fixed in the sense that they are numbers presented to the researcher. For statistical analysis we need to view the dataset as random, or more precisely as a realization of a random process.

In order for the coefficient  $\boldsymbol{\beta}$  defined in (3.1) to make sense as defined, the expectations over the random variables  $(\mathbf{x}, y)$  need to be common across the observations. The most elegant approach to ensure this is to assume that the observations are draws from an identical underlying population  $F$ . This is the standard assumption that the observations are identically distributed:

**Assumption 3.1** The observations  $\{(y_1, \mathbf{x}_1), \dots, (y_i, \mathbf{x}_i), \dots, (y_n, \mathbf{x}_n)\}$  are identically distributed; they are draws from a common distribution  $F$ .

This assumption does not need to be viewed as literally true, rather it is a useful modeling device so that parameters such as  $\beta$  are well defined. This assumption should be interpreted as how we view an observation *a priori*, before we actually observe it. If I tell you that we have a sample with  $n = 59$  observations set in no particular order, then it makes sense to view two observations, say 17 and 58, as draws from the same distribution. We have no reason to expect anything special about either observation.

In econometric theory, we refer to the underlying common distribution  $F$  as the **population**. Some authors prefer the label the **data-generating-process** (DGP). You can think of it as a theoretical concept or an infinitely-large potential population. In contrast we refer to the observations available to us  $\{(y_i, \mathbf{x}_i) : i = 1, \dots, n\}$  as the **sample** or **dataset**. In some contexts the dataset consists of all potential observations, for example administrative tax records may contain every single taxpayer in a political unit. Even in this case we view the observations as if they are random draws from an underlying infinitely-large population, as this will allow us to apply the tools of statistical theory.

The linear projection model applies to the random observations  $(y_i, \mathbf{x}_i)$ . This means that the probability model for the observations is the same as that described in Section 2.18. We can write the model as

$$y_i = \mathbf{x}'_i \beta + e_i \quad (3.2)$$

where the linear projection coefficient  $\beta$  is defined as

$$\beta = \underset{\mathbf{b} \in \mathbb{R}^k}{\operatorname{argmin}} S(\mathbf{b}), \quad (3.3)$$

the minimizer of the expected squared error

$$S(\beta) = \mathbb{E}((y_i - \mathbf{x}'_i \beta)^2), \quad (3.4)$$

and has the explicit solution

$$\beta = (\mathbb{E}(\mathbf{x}_i \mathbf{x}'_i))^{-1} \mathbb{E}(\mathbf{x}_i y_i). \quad (3.5)$$

### 3.3 Moment Estimators

We want to estimate the coefficient  $\beta$  defined in (3.5) from the sample of observations. Notice that  $\beta$  is written as a function of certain population expectations. In this context an appropriate estimator is the same function of the sample moments. Let's explain this in detail.

To start, suppose that we are interested in the population mean  $\mu$  of a random variable  $y_i$  with distribution function  $F$

$$\mu = \mathbb{E}(y_i) = \int_{-\infty}^{\infty} y dF(y). \quad (3.6)$$

The mean  $\mu$  is a function of the distribution  $F$  as written in (3.6). To estimate  $\mu$  given a sample  $\{y_1, \dots, y_n\}$  a natural estimator is the sample mean

$$\hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

Notice that we have written this using two pieces of notation. The notation  $\bar{y}$  with the bar on top is conventional for a sample mean. The notation  $\hat{\mu}$  with the hat “ $\wedge$ ” is conventional in econometrics to denote an estimator of the parameter  $\mu$ . In this case  $\bar{y}$  is the estimator of  $\mu$ , so  $\hat{\mu}$  and  $\bar{y}$  are the same. The sample mean  $\bar{y}$  can be viewed as the natural analog of the population mean (3.6) because  $\bar{y}$  equals the expectation (3.6) with respect to the empirical distribution – the discrete distribution which puts weight  $1/n$  on each observation  $y_i$ . There are many other justifications for  $\bar{y}$  as an estimator for  $\mu$ , we will defer these discussions for now. Suffice it to say that it is the conventional estimator in the lack of other information about  $\mu$  or the distribution of  $y_i$ .

Now suppose that we are interested in a set of population means of possibly non-linear functions of a random vector  $\mathbf{y}$ , say  $\boldsymbol{\mu} = \mathbb{E}(\mathbf{h}(\mathbf{y}_i))$ . For example, we may be interested in the first two moments of  $y_i$ ,

$\mathbb{E}(y_i)$  and  $\mathbb{E}(y_i^2)$ . In this case the natural estimator is the vector of sample means,

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^n \mathbf{h}(y_i).$$

For example,  $\hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^n y_i$  and  $\hat{\mu}_2 = \frac{1}{n} \sum_{i=1}^n y_i^2$ . This is not really a substantive change. We call  $\hat{\boldsymbol{\mu}}$  the **moment estimator** for  $\boldsymbol{\mu}$ .

Now suppose that we are interested in a nonlinear function of a set of moments. For example, consider the variance of  $y$

$$\sigma^2 = \text{var}(y_i) = \mathbb{E}(y_i^2) - (\mathbb{E}(y_i))^2.$$

In general, many parameters of interest, say  $\boldsymbol{\beta}$ , can be written as a function of moments of  $\mathbf{y}$ . Notationally,

$$\begin{aligned}\boldsymbol{\beta} &= \mathbf{g}(\boldsymbol{\mu}) \\ \boldsymbol{\mu} &= \mathbb{E}(\mathbf{h}(y_i)).\end{aligned}$$

Here,  $y_i$  are the random variables,  $\mathbf{h}(y_i)$  are functions (transformations) of the random variables, and  $\boldsymbol{\mu}$  is the mean (expectation) of these functions.  $\boldsymbol{\beta}$  is the parameter of interest, and is the (nonlinear) function  $\mathbf{g}(\cdot)$  of these means.

In this context a natural estimator of  $\boldsymbol{\beta}$  is obtained by replacing  $\boldsymbol{\mu}$  with  $\hat{\boldsymbol{\mu}}$ .

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \mathbf{g}(\hat{\boldsymbol{\mu}}) \\ \hat{\boldsymbol{\mu}} &= \frac{1}{n} \sum_{i=1}^n \mathbf{h}(y_i).\end{aligned}$$

The estimator  $\hat{\boldsymbol{\beta}}$  is sometimes called a “plug-in” estimator, and sometimes a “substitution” estimator. We typically call  $\hat{\boldsymbol{\beta}}$  a moment, or moment-based, estimator of  $\boldsymbol{\beta}$ , since it is a natural extension of the moment estimator  $\hat{\boldsymbol{\mu}}$ .

Take the example of the variance  $\sigma^2 = \text{var}(y_i)$ . Its moment estimator is

$$\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n} \sum_{i=1}^n y_i^2 - \left( \frac{1}{n} \sum_{i=1}^n y_i \right)^2.$$

This is not the only possible estimator for  $\sigma^2$  (there is the well-known bias-corrected version appropriate for independent observations) but it a straightforward and simple choice.

### 3.4 Least Squares Estimator

The linear projection coefficient  $\boldsymbol{\beta}$  is defined in (3.3) as the minimizer of the expected squared error  $S(\boldsymbol{\beta})$  defined in (3.4). For given  $\boldsymbol{\beta}$ , the expected squared error is the expectation of the squared error  $(y_i - \mathbf{x}'_i \boldsymbol{\beta})^2$ . The moment estimator of  $S(\boldsymbol{\beta})$  is the sample average:

$$\begin{aligned}\hat{S}(\boldsymbol{\beta}) &= \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 \\ &= \frac{1}{n} \text{SSE}(\boldsymbol{\beta})\end{aligned}\tag{3.7}$$

where

$$\text{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2$$

is called the **sum-of-squared-errors** function.

Since  $\hat{S}(\boldsymbol{\beta})$  is a sample average, we can interpret it as an estimator of the expected squared error  $S(\boldsymbol{\beta})$ . Examining  $\hat{S}(\boldsymbol{\beta})$  as a function of  $\boldsymbol{\beta}$  is informative about how  $S(\boldsymbol{\beta})$  varies with  $\boldsymbol{\beta}$ . Since the projection coefficient minimizes  $S(\boldsymbol{\beta})$ , an analog estimator minimizes (3.7).

**Definition 3.1** The least-squares estimator  $\hat{\beta}$  is

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}} \hat{S}(\beta)$$

where

$$\hat{S}(\beta) = \frac{1}{n} \sum_{i=1}^n (y_i - x'_i \beta)^2$$

Alternatively, as  $\hat{S}(\beta)$  is a scale multiple of  $SSE(\beta)$ , we may equivalently define  $\hat{\beta}$  as the minimizer of  $SSE(\beta)$ . Hence  $\hat{\beta}$  is commonly called the **least-squares (LS)** estimator of  $\beta$ . The estimator is also commonly referred to as the **ordinary least-squares (OLS)** estimator. For the origin of this label see the historical discussion on Adrien-Marie Legendre below. Here, as is common in econometrics, we put a hat “ $\hat{\cdot}$ ” over the parameter  $\beta$  to indicate that  $\hat{\beta}$  is a sample estimate of  $\beta$ . This is a helpful convention. Just by seeing the symbol  $\hat{\beta}$  we can immediately interpret it as an estimator (because of the hat) of the parameter  $\beta$ . Sometimes when we want to be explicit about the estimation method, we will write  $\hat{\beta}_{ols}$  to signify that it is the OLS estimator. It is also common to see the notation  $\hat{\beta}_n$ , where the subscript “ $n$ ” indicates that the estimator depends on the sample size  $n$ .

It is important to understand the distinction between population parameters such as  $\beta$  and sample estimators such as  $\hat{\beta}$ . The population parameter  $\beta$  is a non-random feature of the population while the sample estimator  $\hat{\beta}$  is a random feature of a random sample.  $\beta$  is fixed, while  $\hat{\beta}$  varies across samples.

### 3.5 Solving for Least Squares with One Regressor

For simplicity, we start by considering the case  $k = 1$  so that there is a scalar regressor  $x_i$  and a scalar coefficient  $\beta$ . To illustrate, Figure 3.1 displays a scatter plot<sup>1</sup> of 20 pairs  $(y_i, x_i)$ .

The sum of squared errors  $SSE(\beta)$  is a function of  $\beta$ . Given  $\beta$  we calculate the “error”  $y_i - x_i \beta$  by taking the vertical distance between  $y_i$  and  $x_i \beta$ . This can be seen in Figure 3.1 by the vertical lines which connect the observations to the straight line. These vertical lines are the errors  $y_i - x_i \beta$ . The sum of squared errors is the sum of the 20 squared lengths shown in Figure 3.1.

The sum of squared errors is the function

$$\begin{aligned} SSE(\beta) &= \sum_{i=1}^n (y_i - x_i \beta)^2 \\ &= \left( \sum_{i=1}^n y_i^2 \right) - 2\beta \left( \sum_{i=1}^n x_i y_i \right) + \beta^2 \left( \sum_{i=1}^n x_i^2 \right). \end{aligned}$$

This is a quadratic function of  $\beta$ . For example, for the sample displayed in Figure 3.1, the sum of squared error function is displayed in Figure 3.2 over the range [2, 4]. The coefficient  $\beta$  ranges along the  $x$ -axis. The sum-of-squared errors  $SSE(\beta)$  as a function of  $\beta$  is displayed on the  $y$ -axis.

The OLS estimator  $\hat{\beta}$  minimizes this function. From elementary algebra we know that the minimizer of the quadratic function  $a - 2bx + cx^2$  is  $x = b/c$ . Thus the minimizer of  $SSE(\beta)$  is

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}. \quad (3.8)$$

For example, the minimizer of the sum of squared error function displayed in Figure 3.2 is  $\hat{\beta} = 3.07$ , and is marked on the  $x$ -axis.

<sup>1</sup>The observations were generated by simulation as  $x_i \sim U[0, 1]$ ,  $e_i \sim N[0, 1]$ , and  $y_i = 3x_i + e_i$ .

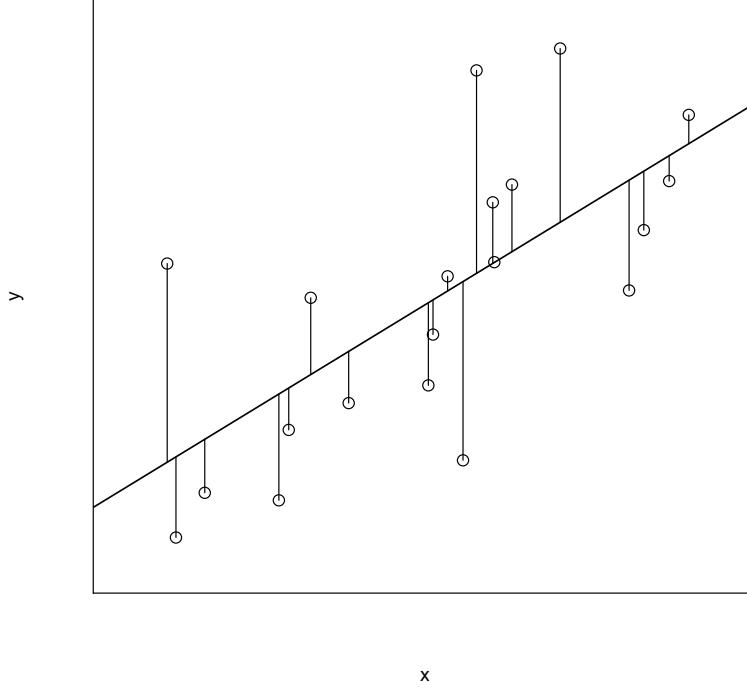


Figure 3.1: Measuring Errors as Deviations from a Fitted Line

The intercept-only model is the special case  $x_i = 1$ . In this case we find

$$\hat{\beta} = \frac{\sum_{i=1}^n 1 y_i}{\sum_{i=1}^n 1^2} = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y}, \quad (3.9)$$

the sample mean of  $y_i$ . Here, as is common, we put a bar “ $\bar{}$ ” over  $y$  to indicate that the quantity is a sample mean. This calculation shows that the OLS estimator in the intercept-only model is the sample mean.

Technically, the estimator  $\hat{\beta}$  in (3.8) only exists if the denominator is non-zero. Since it is a sum of squares it is necessarily non-negative. Thus  $\hat{\beta}$  exists if  $\sum_{i=1}^n x_i^2 > 0$ .

### 3.6 Solving for Least Squares with Multiple Regressors

We now consider the case with  $k > 1$  so that the coefficient  $\beta$  is a vector.

To illustrate, Figure 3.3 displays a scatter plot of 100 triples  $(y_i, x_{1i}, x_{2i})$ . The regression function  $\mathbf{x}'\beta = x_1\beta_1 + x_2\beta_2$  is a 2-dimensional surface, and is shown as the plane in Figure 3.3.

The sum of squared errors  $SSE(\beta)$  is a function of the vector  $\beta$ . For any  $\beta$  the error  $y_i - \mathbf{x}'_i\beta$  is the vertical distance between  $y_i$  and  $\mathbf{x}'_i\beta$ . This can be seen in Figure 3.3 by the vertical lines which connect the observations to the plane. As in the single regressor case, these vertical lines are the errors  $e_i = y_i - \mathbf{x}'_i\beta$ . The sum of squared errors is the sum of the 100 squared lengths shown in Figure 3.3.

The sum of squared errors can be written as

$$SSE(\beta) = \sum_{i=1}^n y_i^2 - 2\beta' \sum_{i=1}^n \mathbf{x}_i y_i + \beta' \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \beta.$$

As in the single regressor case, this is a quadratic function in  $\beta$ . The difference is that in the multiple regressor case this is a vector-valued quadratic function. To visualize the sum of squared errors function,

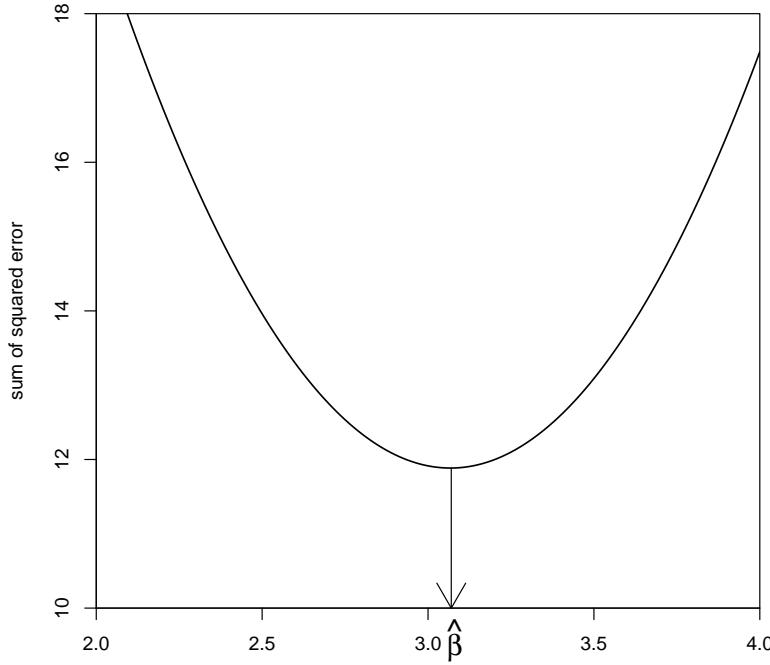


Figure 3.2: Sum of Squared Error Function for One Regressor

Figure 3.4 displays  $\text{SSE}(\boldsymbol{\beta})$  for the data shown in Figure 3.3. Another way to visualize a 3-dimensional surface is by a contour plot. A contour plot of the same  $\text{SSE}(\boldsymbol{\beta})$  function is shown in Figure 3.5. The contour lines are points in the  $(\beta_1, \beta_2)$  space where  $\text{SSE}(\boldsymbol{\beta})$  takes the same value. The contour lines are elliptical.

The least-squares estimator  $\hat{\boldsymbol{\beta}}$  minimizes  $\text{SSE}(\boldsymbol{\beta})$ . A simple way to find the minimum is by solving the first-order conditions. The latter are

$$0 = \frac{\partial}{\partial \boldsymbol{\beta}} \text{SSE}(\hat{\boldsymbol{\beta}}) = -2 \sum_{i=1}^n \mathbf{x}_i y_i + 2 \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \hat{\boldsymbol{\beta}}. \quad (3.10)$$

We have written this using a single expression, but it is actually a system of  $k$  equations with  $k$  unknowns (the elements of  $\hat{\boldsymbol{\beta}}$ ).

The solution for  $\hat{\boldsymbol{\beta}}$  may be found by solving the system of  $k$  equations in (3.10). We can write this solution compactly using matrix algebra. Dividing (3.10) by 2 we obtain

$$\sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \hat{\boldsymbol{\beta}} = \sum_{i=1}^n \mathbf{x}_i y_i. \quad (3.11)$$

This is a system of equations of the form  $\mathbf{Ab} = \mathbf{c}$  where  $\mathbf{A}$  is  $k \times k$  and  $\mathbf{b}$  and  $\mathbf{c}$  are  $k \times 1$ . The solution is  $\mathbf{b} = \mathbf{A}^{-1}\mathbf{c}$ , and can be obtained by pre-multiplying  $\mathbf{Ab} = \mathbf{c}$  by  $\mathbf{A}^{-1}$  and using the matrix inverse property  $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_k$ . Applied to (3.11) we find an explicit formula for the least-squares estimator

$$\hat{\boldsymbol{\beta}} = \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{x}_i y_i \right). \quad (3.12)$$

This is the natural estimator of the best linear projection coefficient  $\boldsymbol{\beta}$  defined in (3.3), and can also be called the linear projection estimator.

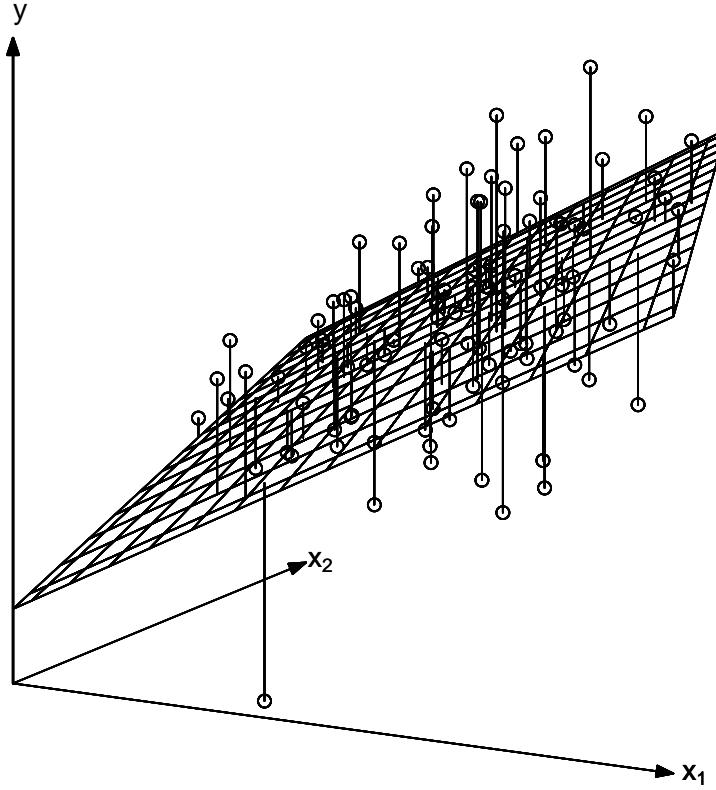


Figure 3.3: Scatter Plot and Regression Plane

Recall that we claim that  $\hat{\beta}$  in (3.12) is the minimizer of  $\text{SSE}(\beta)$ , and we found this by solving the first-order conditions. To be complete we should verify the second-order conditions. We calculate that

$$\frac{\partial^2}{\partial \beta \partial \beta'} \text{SSE}(\beta) = 2 \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i > 0$$

which is a positive definite matrix. This shows that the second-order condition for minimization is satisfied, so  $\hat{\beta}$  is indeed the unique minimizer of  $\text{SSE}(\beta)$ .

Returning to the example sum-of-squared errors function  $\text{SSE}(\beta)$  displayed in Figures 3.4 and 3.5, the least-squares estimator  $\hat{\beta}$  is the pair  $(\hat{\beta}_1, \hat{\beta}_2)$  which minimize this function; visually it is the low spot in the 3-dimensional graph, and is marked in Figure 3.5 as the center point of the contour plots.

Returning to equation (3.12) suppose that  $k = 1$ . In this case  $\mathbf{x}_i$  is scalar so  $\mathbf{x}_i \mathbf{x}'_i = x_i^2$ . Then (3.12) simplifies to the expression (3.8) previously derived. The expression (3.12) is a notationally simple generalization but requires a careful attention to vector and matrix manipulations.

Alternatively, equation (3.5) writes the projection coefficient  $\beta$  as an explicit function of the popula-

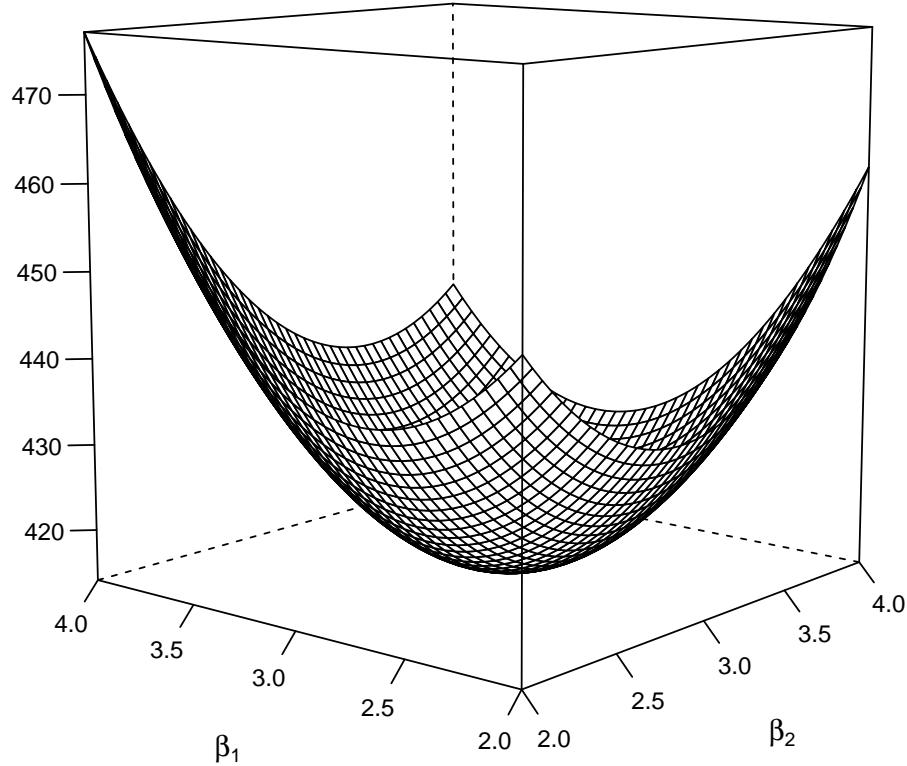


Figure 3.4: Sum-of-Squared Errors Function

tion moments  $\mathbf{Q}_{xy}$  and  $\mathbf{Q}_{xx}$ . Their moment estimators are the sample moments

$$\begin{aligned}\hat{\mathbf{Q}}_{xy} &= \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i y_i \\ \hat{\mathbf{Q}}_{xx} &= \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i.\end{aligned}$$

The moment estimator of  $\boldsymbol{\beta}$  replaces the population moments in (3.5) with the sample moments:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \hat{\mathbf{Q}}_{xx}^{-1} \hat{\mathbf{Q}}_{xy} \\ &= \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i y_i \right) \\ &= \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{x}_i y_i \right)\end{aligned}$$

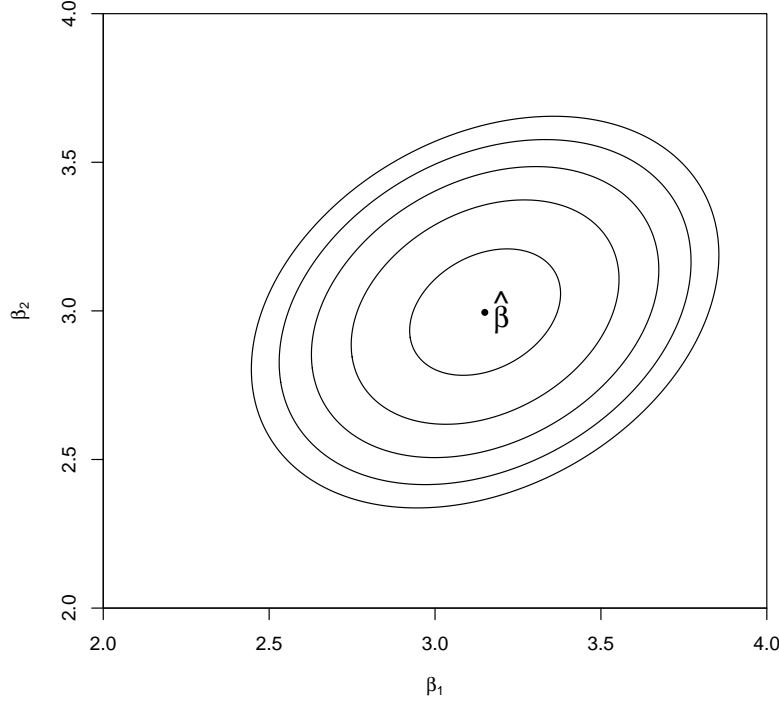


Figure 3.5: Sum-of-Squared Errors Contour

which is identical with (3.12).

Technically, the estimator  $\hat{\beta}$  in (3.12) exists and is unique only if the inverted matrix is actually invertible, which holds if (and only if) this matrix is positive definite. This excludes the case that  $x_i$  contains redundant regressors or regressors with no sample variation. This will be discussed further in Section 3.24.

**Theorem 3.1** If  $\sum_{i=1}^n x_i x_i' > 0$ , the least squares estimator equals

$$\hat{\beta} = \left( \sum_{i=1}^n x_i x_i' \right)^{-1} \left( \sum_{i=1}^n x_i y_i \right).$$

### Adrien-Marie Legendre

The method of least-squares was first published in 1805 by the French mathematician Adrien-Marie Legendre (1752-1833). Legendre proposed least-squares as a solution to the algebraic problem of solving a system of equations when the number of equations exceeded the number of unknowns. This was a vexing and common problem in astronomical measurement. As viewed by Legendre, (3.2) is a set of  $n$  equations with  $k$  unknowns. As the equations cannot be solved exactly, Legendre's goal was to select  $\beta$  to make the set of errors as small as possible. He proposed the sum of squared error criterion, and derived the algebraic solution presented above. As he noted, the first-order conditions (3.10) is a system of  $k$  equations with  $k$  unknowns, which can be solved by "ordinary" methods. Hence the method became known as **Ordinary Least Squares** and to this day we still use the abbreviation OLS to refer to Legendre's estimation method.

## 3.7 Illustration

We illustrate the least-squares estimator in practice with the data set used to calculate the estimates reported in Chapter 2. This is the March 2009 Current Population Survey, which has extensive information on the U.S. population. This data set is described in more detail in Section 3.22. For this illustration, we use the sub-sample of married (spouse present) black female wage earners with 12 years potential work experience. This sub-sample has 20 observations.

In Table 3.1 we display the observations for reference. Each row is an individual observation, which are the data for an individual person. The columns correspond to the variables (measurements) for the individuals. The second column is the reported wage (total annual earnings divided by hours worked). The third column is the natural logarithm of the wage. The fourth column is years of education. The fifth and six columns are further transformations, specifically the square of *education* and the product of *education* and  $\log(wage)$ . The bottom row are the sums of the elements in that column.

Putting the variables into the standard regression notation, let  $y_i$  be log wages and  $x_i$  be years of education and an intercept. Then from the column sums in Table 3.1 we have

$$\sum_{i=1}^n x_i y_i = \begin{pmatrix} 995.86 \\ 62.64 \end{pmatrix}$$

and

$$\sum_{i=1}^n x_i x'_i = \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix}.$$

Taking the inverse we obtain

$$\left( \sum_{i=1}^n x_i x'_i \right)^{-1} = \begin{pmatrix} 0.0125 & -0.196 \\ -0.196 & 3.124 \end{pmatrix}.$$

Thus by matrix multiplication

$$\begin{aligned} \hat{\beta} &= \begin{pmatrix} 0.0125 & -0.196 \\ -0.196 & 3.124 \end{pmatrix} \begin{pmatrix} 995.86 \\ 62.64 \end{pmatrix} \\ &= \begin{pmatrix} 0.155 \\ 0.698 \end{pmatrix}. \end{aligned}$$

In practice, the regression estimates  $\hat{\beta}$  are computed by computer software without the user taking the explicit steps listed above. However, it is useful to understand that the least-squares estimator can

Table 3.1: Observations From CPS Data Set

Observation	Wage	log(Wage)	Education	Education <sup>2</sup>	Education*log(Wage)
1	37.93	3.64	18	324	65.44
2	40.87	3.71	18	324	66.79
3	14.18	2.65	13	169	34.48
4	16.83	2.82	16	256	45.17
5	33.17	3.50	16	256	56.03
6	29.81	3.39	18	324	61.11
7	54.62	4.00	16	256	64.00
8	43.08	3.76	18	324	67.73
9	14.42	2.67	12	144	32.03
10	14.90	2.70	16	256	43.23
11	21.63	3.07	18	324	55.44
12	11.09	2.41	16	256	38.50
13	10.00	2.30	13	169	29.93
14	31.73	3.46	14	196	48.40
15	11.06	2.40	12	144	28.84
16	18.75	2.93	16	256	46.90
17	27.35	3.31	14	196	46.32
18	24.04	3.18	16	256	50.76
19	36.06	3.59	18	324	64.53
20	23.08	3.14	16	256	50.22
Sum		62.64	314	5010	995.86

be calculated by simple algebraic operations. If your data is in a spreadsheet similar to Table 3.1, then the listed transformations (logarithm, squares and cross-products, column sums) can be computed by spreadsheet operations.  $\hat{\beta}$  could then be calculated by matrix inversion and multiplication. One again, this is rarely done by applied economists since computer software is available to ease the process.

We often write the estimated equation using the format

$$\widehat{\log(Wage)} = 0.155 \text{ education} + 0.698. \quad (3.13)$$

An interpretation of the estimated equation is that each year of education is associated with a 16% increase in mean wages.

Equation (3.13) is called a **bivariate regression** as there are two variables. It is also called a **simple regression** as there is a single regressor. A **multiple regression** has two or more regressors, and allows a more detailed investigation. Let's take an example similar to (3.13) but include all levels of experience. This time, we use the sub-sample of single (never married) Asian men, which has 268 observations. Including as regressors years of potential work experience (*experience*) and its square (*experience*<sup>2</sup>/100) (we divide by 100 to simplify reporting), we obtain the estimates

$$\widehat{\log(Wage)} = 0.143 \text{ education} + 0.036 \text{ experience} - 0.071 \text{ experience}^2/100 + 0.575. \quad (3.14)$$

These estimates suggest a 14% increase in mean wages per year of education, holding experience constant.

### 3.8 Least Squares Residuals

As a by-product of estimation, we define the **fitted value**

$$\hat{y}_i = \mathbf{x}'_i \hat{\beta}$$

and the **residual**

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}. \quad (3.15)$$

Sometimes  $\hat{y}_i$  is called the predicted value, but this is a misleading label. The fitted value  $\hat{y}_i$  is a function of the entire sample, including  $y_i$ , and thus cannot be interpreted as a valid prediction of  $y_i$ . It is thus more accurate to describe  $\hat{y}_i$  as a *fitted* rather than a *predicted* value.

Note that  $y_i = \hat{y}_i + \hat{e}_i$  and

$$y_i = \mathbf{x}'_i \hat{\boldsymbol{\beta}} + \hat{e}_i. \quad (3.16)$$

We make a distinction between the **error**  $e_i$  and the **residual**  $\hat{e}_i$ . The error  $e_i$  is unobservable while the residual  $\hat{e}_i$  is a by-product of estimation. These two variables are frequently mislabeled, which can cause confusion.

Equation (3.10) implies that

$$\sum_{i=1}^n \mathbf{x}_i \hat{e}_i = \mathbf{0}. \quad (3.17)$$

To see this by a direct calculation, using (3.15) and (3.12),

$$\begin{aligned} \sum_{i=1}^n \mathbf{x}_i \hat{e}_i &= \sum_{i=1}^n \mathbf{x}_i (y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}) \\ &= \sum_{i=1}^n \mathbf{x}_i y_i - \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \hat{\boldsymbol{\beta}} \\ &= \sum_{i=1}^n \mathbf{x}_i y_i - \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{x}_i y_i \right) \\ &= \sum_{i=1}^n \mathbf{x}_i y_i - \sum_{i=1}^n \mathbf{x}_i y_i \\ &= \mathbf{0}. \end{aligned}$$

When  $\mathbf{x}_i$  contains a constant, an implication of (3.17) is

$$\frac{1}{n} \sum_{i=1}^n \hat{e}_i = 0. \quad (3.18)$$

Thus the residuals have a sample mean of zero and the sample correlation between the regressors and the residual is zero. These are algebraic results, and hold true for all linear regression estimates.

### 3.9 Demeaned Regressors

Sometimes it is useful to separate the constant from the other regressors, and write the linear projection equation in the format

$$y_i = \mathbf{x}'_i \hat{\boldsymbol{\beta}} + \alpha + e_i$$

where  $\alpha$  is the intercept and  $\mathbf{x}_i$  does not contain a constant. The least-squares estimates and residuals can be written as

$$y_i = \mathbf{x}'_i \hat{\boldsymbol{\beta}} + \hat{\alpha} + \hat{e}_i.$$

In this case (3.17) can be written as the equation system

$$\begin{aligned} \sum_{i=1}^n (y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}} - \hat{\alpha}) &= 0 \\ \sum_{i=1}^n \mathbf{x}_i (y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}} - \hat{\alpha}) &= \mathbf{0}. \end{aligned}$$

The first equation implies

$$\hat{\alpha} = \bar{y} - \bar{x}' \hat{\beta}.$$

Subtracting from the second we obtain

$$\sum_{i=1}^n \mathbf{x}_i \left( (y_i - \bar{y}) - (\mathbf{x}_i - \bar{\mathbf{x}})' \hat{\beta} \right) = \mathbf{0}.$$

Solving for  $\hat{\beta}$  we find

$$\begin{aligned} \hat{\beta} &= \left( \sum_{i=1}^n \mathbf{x}_i (\mathbf{x}_i - \bar{\mathbf{x}})' \right)^{-1} \left( \sum_{i=1}^n \mathbf{x}_i (y_i - \bar{y}) \right) \\ &= \left( \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})' \right)^{-1} \left( \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) (y_i - \bar{y}) \right). \end{aligned} \quad (3.19)$$

Thus the OLS estimator for the slope coefficients is a regression with demeaned data.

The representation (3.19) is known as the demeaned formula for the least-squares estimator.

### 3.10 Model in Matrix Notation

For many purposes, including computation, it is convenient to write the model and statistics in matrix notation. The linear equation (2.22) is a system of  $n$  equations, one for each observation. We can stack these  $n$  equations together as

$$\begin{aligned} y_1 &= \mathbf{x}'_1 \boldsymbol{\beta} + e_1 \\ y_2 &= \mathbf{x}'_2 \boldsymbol{\beta} + e_2 \\ &\vdots \\ y_n &= \mathbf{x}'_n \boldsymbol{\beta} + e_n. \end{aligned}$$

Now define

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_n \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

Observe that  $\mathbf{y}$  and  $\mathbf{e}$  are  $n \times 1$  vectors, and  $\mathbf{X}$  is an  $n \times k$  matrix. Then the system of  $n$  equations can be compactly written in the single equation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}. \quad (3.20)$$

Sample sums can be written in matrix notation. For example

$$\begin{aligned} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i &= \mathbf{X}' \mathbf{X} \\ \sum_{i=1}^n \mathbf{x}_i y_i &= \mathbf{X}' \mathbf{y}. \end{aligned}$$

Therefore the least-squares estimator can be written as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{y}).$$

The matrix version of (3.16) and estimated version of (3.20) is

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{e}},$$

or equivalently the residual vector is

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}.$$

Using the residual vector, we can write (3.17) as

$$\mathbf{X}'\hat{\mathbf{e}} = \mathbf{0}.$$

It can also be useful to write the sum-of-squared error criterion as

$$SSE(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Using matrix notation we have simple expressions for most estimators. This is particularly convenient for computer programming, as most languages allow matrix notation and manipulation.

### Theorem 3.2 Important Matrix Expressions

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'\mathbf{y}) \\ \hat{\mathbf{e}} &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ \mathbf{X}'\hat{\mathbf{e}} &= \mathbf{0}.\end{aligned}$$

### Early Use of Matrices

The earliest known treatment of the use of matrix methods to solve simultaneous systems is found in Chapter 8 of the Chinese text *The Nine Chapters on the Mathematical Art*, written by several generations of scholars from the 10th to 2nd century BCE.

## 3.11 Projection Matrix

Define the matrix

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Observe that

$$\mathbf{P}\mathbf{X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{X}.$$

This is a property of a **projection matrix**. More generally, for any matrix  $\mathbf{Z}$  which can be written as  $\mathbf{Z} = \mathbf{X}\boldsymbol{\Gamma}$  for some matrix  $\boldsymbol{\Gamma}$  (we say that  $\mathbf{Z}$  lies in the **range space** of  $\mathbf{X}$ ), then

$$\mathbf{P}\mathbf{Z} = \mathbf{P}\mathbf{X}\boldsymbol{\Gamma} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\Gamma} = \mathbf{X}\boldsymbol{\Gamma} = \mathbf{Z}.$$

As an important example, if we partition the matrix  $\mathbf{X}$  into two matrices  $\mathbf{X}_1$  and  $\mathbf{X}_2$  so that

$$\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2],$$

then  $\mathbf{P}\mathbf{X}_1 = \mathbf{X}_1$ . (See Exercise 3.7.)

The projection matrix  $\mathbf{P}$  has the algebraic property that it is an idempotent matrix  $\mathbf{P}\mathbf{P} = \mathbf{P}$ . See Theorem 3.3.2 below. For the general properties of projection matrices see Section A.11.

The matrix  $\mathbf{P}$  creates the fitted values in a least-squares regression:

$$\mathbf{P}\mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{y}}.$$

Because of this property,  $\mathbf{P}$  is also known as the “hat matrix”.

A special example of a projection matrix occurs when  $\mathbf{X} = \mathbf{1}_n$  is an  $n$ -vector of ones. Then

$$\begin{aligned}\mathbf{P} &= \mathbf{1}_n(\mathbf{1}'_n\mathbf{1}_n)^{-1}\mathbf{1}'_n \\ &= \frac{1}{n}\mathbf{1}_n\mathbf{1}'_n.\end{aligned}$$

Note that in this case

$$\begin{aligned}\mathbf{P}\mathbf{y} &= \mathbf{1}_n(\mathbf{1}'_n\mathbf{1}_n)^{-1}\mathbf{1}'_n\mathbf{y} \\ &= \mathbf{1}_n\bar{y}\end{aligned}$$

creates an  $n$ -vector whose elements are the sample mean  $\bar{y}$  of  $y_i$ .

The projection matrix  $\mathbf{P}$  appears frequently in algebraic manipulations in least squares regression. The matrix has the following important properties.

**Theorem 3.3** The projection matrix  $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  for any  $n \times k$   $\mathbf{X}$  with  $n \geq k$  has the following algebraic properties

1.  $\mathbf{P}$  is **symmetric** ( $\mathbf{P}' = \mathbf{P}$ ).
2.  $\mathbf{P}$  is **idempotent** ( $\mathbf{P}\mathbf{P} = \mathbf{P}$ ).
3.  $\text{tr } \mathbf{P} = k$ .
4. The eigenvalues of  $\mathbf{P}$  are 1 and 0. There are  $k$  eigenvalues equalling 1 and  $n - k$  equalling 0.
5.  $\text{rank}(\mathbf{P}) = k$ .

We close this section by proving the claims in Theorem 3.3. Part 1 holds since

$$\begin{aligned}\mathbf{P}' &= \left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' \\ &= (\mathbf{X}')'\left((\mathbf{X}'\mathbf{X})^{-1}\right)'(\mathbf{X})' \\ &= \mathbf{X}\left((\mathbf{X}'\mathbf{X})'\right)^{-1}\mathbf{X}' \\ &= \mathbf{X}\left((\mathbf{X})'(\mathbf{X}')'\right)^{-1}\mathbf{X}' \\ &= \mathbf{P}.\end{aligned}$$

To establish part 2, the fact that  $\mathbf{P}\mathbf{X} = \mathbf{X}$  implies that

$$\begin{aligned}\mathbf{P}\mathbf{P} &= \mathbf{P}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{P}\end{aligned}$$

as claimed.

For part 3,

$$\begin{aligned}\text{tr } \mathbf{P} &= \text{tr} \left( \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \right) \\ &= \text{tr} \left( (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{X} \right) \\ &= \text{tr} (\mathbf{I}_k) \\ &= k.\end{aligned}$$

See Appendix A.5 for definition and properties of the trace operator.

For part 4, it is shown in Appendix A.11 that the eigenvalues  $\lambda_i$  of an idempotent matrix are all 1 and 0. Since  $\text{tr } \mathbf{P}$  equals the sum of the  $n$  eigenvalues and  $\text{tr } \mathbf{P} = k$  by part 3, it follows that there are  $k$  eigenvalues equalling 1 and the remainder  $(n - k)$  equalling  $n - k$ .

For part 5, observe that  $\mathbf{P}$  is positive semi-definite since its eigenvalues are all non-negative. By Theorem A.4.5, its rank equals the number of positive eigenvalues, which is  $k$  as claimed.

## 3.12 Orthogonal Projection

Define

$$\begin{aligned}\mathbf{M} &= \mathbf{I}_n - \mathbf{P} \\ &= \mathbf{I}_n - \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'\end{aligned}$$

where  $\mathbf{I}_n$  is the  $n \times n$  identity matrix. Note that

$$\mathbf{M}\mathbf{X} = (\mathbf{I}_n - \mathbf{P})\mathbf{X} = \mathbf{X} - \mathbf{P}\mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{0}. \quad (3.22)$$

Thus  $\mathbf{M}$  and  $\mathbf{X}$  are orthogonal. We call  $\mathbf{M}$  an **orthogonal projection matrix**, or more colorfully an **annihilator matrix**, due to the property that for any matrix  $\mathbf{Z}$  in the range space of  $\mathbf{X}$  then

$$\mathbf{M}\mathbf{Z} = \mathbf{Z} - \mathbf{P}\mathbf{Z} = \mathbf{0}.$$

For example,  $\mathbf{M}\mathbf{X}_1 = \mathbf{0}$  for any subcomponent  $\mathbf{X}_1$  of  $\mathbf{X}$ , and  $\mathbf{M}\mathbf{P} = \mathbf{0}$  (see Exercise 3.7).

The orthogonal projection matrix  $\mathbf{M}$  has similar properties with  $\mathbf{P}$ , including that  $\mathbf{M}$  is symmetric ( $\mathbf{M}' = \mathbf{M}$ ) and idempotent ( $\mathbf{M}\mathbf{M} = \mathbf{M}$ ). Similarly to Theorem 3.3.3 we can calculate

$$\text{tr } \mathbf{M} = n - k. \quad (3.23)$$

(See Exercise 3.9.) One implication is that the rank of  $\mathbf{M}$  is  $n - k$ .

While  $\mathbf{P}$  creates fitted values,  $\mathbf{M}$  creates least-squares residuals:

$$\mathbf{My} = \mathbf{y} - \mathbf{Py} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{e}}. \quad (3.24)$$

As discussed in the previous section, a special example of a projection matrix occurs when  $\mathbf{X} = \mathbf{1}_n$  is an  $n$ -vector of ones, so that  $\mathbf{P} = \mathbf{1}_n (\mathbf{1}'_n \mathbf{1}_n)^{-1} \mathbf{1}'_n$ . In this case the orthogonal projection matrix is

$$\begin{aligned}\mathbf{M} &= \mathbf{I}_n - \mathbf{P} \\ &= \mathbf{I}_n - \mathbf{1}_n (\mathbf{1}'_n \mathbf{1}_n)^{-1} \mathbf{1}'_n.\end{aligned}$$

While  $\mathbf{P}$  creates a vector of sample means,  $\mathbf{M}$  creates demeaned values:

$$\mathbf{My} = \mathbf{y} - \mathbf{1}_n \bar{y}.$$

For simplicity we will often write the right-hand-side as  $\mathbf{y} - \bar{y}$ . The  $i^{th}$  element is  $y_i - \bar{y}$ , the **demeaned** value of  $y_i$ .

We can also use (3.24) to write an alternative expression for the residual vector. Substituting  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$  into  $\hat{\mathbf{e}} = \mathbf{My}$  and using  $\mathbf{M}\mathbf{X} = \mathbf{0}$  we find

$$\hat{\mathbf{e}} = \mathbf{My} = \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) = \mathbf{Me} \quad (3.25)$$

which is free of dependence on the regression coefficient  $\boldsymbol{\beta}$ .

### 3.13 Estimation of Error Variance

The error variance  $\sigma^2 = \mathbb{E}(e_i^2)$  is a moment, so a natural estimator is a moment estimator. If  $e_i$  were observed we would estimate  $\sigma^2$  by

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n e_i^2. \quad (3.26)$$

However, this is infeasible as  $e_i$  is not observed. In this case it is common to take a two-step approach to estimation. The residuals  $\hat{e}_i$  are calculated in the first step, and then we substitute  $\hat{e}_i$  for  $e_i$  in expression (3.26) to obtain the feasible estimator

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{e}_i^2. \quad (3.27)$$

In matrix notation, we can write (3.26) and (3.27) as

$$\tilde{\sigma}^2 = n^{-1} \mathbf{e}' \mathbf{e}$$

and

$$\hat{\sigma}^2 = n^{-1} \hat{\mathbf{e}}' \hat{\mathbf{e}}. \quad (3.28)$$

Recall the expressions  $\hat{\mathbf{e}} = \mathbf{M}\mathbf{y} = \mathbf{M}\mathbf{e}$  from (3.24) and (3.25). Applied to (3.28) we find

$$\begin{aligned} \hat{\sigma}^2 &= n^{-1} \hat{\mathbf{e}}' \hat{\mathbf{e}} \\ &= n^{-1} \mathbf{y}' \mathbf{M} \mathbf{M} \mathbf{y} \\ &= n^{-1} \mathbf{y}' \mathbf{M} \mathbf{y} \\ &= n^{-1} \mathbf{e}' \mathbf{M} \mathbf{e} \end{aligned} \quad (3.29)$$

the third equality since  $\mathbf{M} \mathbf{M} = \mathbf{M}$ .

An interesting implication is that

$$\begin{aligned} \tilde{\sigma}^2 - \hat{\sigma}^2 &= n^{-1} \mathbf{e}' \mathbf{e} - n^{-1} \mathbf{e}' \mathbf{M} \mathbf{e} \\ &= n^{-1} \mathbf{e}' \mathbf{P} \mathbf{e} \\ &\geq 0. \end{aligned}$$

The final inequality holds because  $\mathbf{P}$  is positive semi-definite and  $\mathbf{e}' \mathbf{P} \mathbf{e}$  is a quadratic form. This shows that the feasible estimator  $\hat{\sigma}^2$  is numerically smaller than the idealized estimator (3.26).

### 3.14 Analysis of Variance

Another way of writing (3.24) is

$$\mathbf{y} = \mathbf{P}\mathbf{y} + \mathbf{M}\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}}. \quad (3.30)$$

This decomposition is **orthogonal**, that is

$$\hat{\mathbf{y}}' \hat{\mathbf{e}} = (\mathbf{P}\mathbf{y})' (\mathbf{M}\mathbf{y}) = \mathbf{y}' \mathbf{P} \mathbf{M} \mathbf{y} = 0. \quad (3.31)$$

It follows that

$$\mathbf{y}' \mathbf{y} = \hat{\mathbf{y}}' \hat{\mathbf{y}} + 2\hat{\mathbf{y}}' \hat{\mathbf{e}} + \hat{\mathbf{e}}' \hat{\mathbf{e}} = \hat{\mathbf{y}}' \hat{\mathbf{y}} + \hat{\mathbf{e}}' \hat{\mathbf{e}}$$

or

$$\sum_{i=1}^n y_i^2 = \sum_{i=1}^n \hat{y}_i^2 + \sum_{i=1}^n \hat{e}_i^2.$$

Subtracting  $\bar{y}$  from both sides of (3.30) we obtain

$$\mathbf{y} - \mathbf{1}_n \bar{y} = \hat{\mathbf{y}} - \mathbf{1}_n \bar{y} + \hat{\mathbf{e}}.$$

This decomposition is also orthogonal when  $X$  contains a constant, as

$$(\hat{\mathbf{y}} - \mathbf{1}_n \bar{y})' \hat{\mathbf{e}} = \hat{\mathbf{y}}' \hat{\mathbf{e}} - \bar{y} \mathbf{1}_n' \hat{\mathbf{e}} = 0$$

under (3.18). It follows that

$$(\mathbf{y} - \mathbf{1}_n \bar{y})' (\mathbf{y} - \mathbf{1}_n \bar{y}) = (\hat{\mathbf{y}} - \mathbf{1}_n \bar{y})' (\hat{\mathbf{y}} - \mathbf{1}_n \bar{y}) + \hat{\mathbf{e}}' \hat{\mathbf{e}}$$

or

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{e}_i^2.$$

This is commonly called the **analysis-of-variance** formula for least squares regression.

A commonly reported statistic is the **coefficient of determination or R-squared**:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n \hat{e}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

It is often described as the fraction of the sample variance of  $y_i$  which is explained by the least-squares fit.  $R^2$  is a crude measure of regression fit. We have better measures of fit, but these require a statistical (not just algebraic) analysis and we will return to these issues later. One deficiency with  $R^2$  is that it increases when regressors are added to a regression (see Exercise 3.16) so the “fit” can be always increased by increasing the number of regressors.

The coefficient of determination was introduced by Wright (1921).

## 3.15 Projections

One way to visualize least squares fitting is as a projection operation.

Write the regressor matrix as  $\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2 \dots \ \mathbf{X}_k]$  where  $\mathbf{X}_j$  is the  $j^{th}$  column of  $\mathbf{X}$ . The range space  $\mathcal{R}(\mathbf{X})$  of  $\mathbf{X}$  is the space consisting of all linear combinations of the columns  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_k$ .  $\mathcal{R}(\mathbf{X})$  is a  $k$  dimensional surface contained in  $\mathbb{R}^n$ . If  $k = 2$  then  $\mathcal{R}(\mathbf{X})$  is a plane. The operator  $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  projects vectors onto the  $\mathcal{R}(\mathbf{X})$ . In particular, the fitted values  $\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$  are the projection of  $\mathbf{y}$  onto  $\mathcal{R}(\mathbf{X})$ .

To visualize, examine Figure 3.6. This displays the case  $n = 3$  and  $k = 2$ . Displayed are three vectors  $\mathbf{y}$ ,  $\mathbf{X}_1$ , and  $\mathbf{X}_2$ , which are each elements of  $\mathbb{R}^3$ . The plane which is created by  $\mathbf{X}_1$  and  $\mathbf{X}_2$  is the range space  $\mathcal{R}(\mathbf{X})$ . Regression fitted values must be linear combinations of  $\mathbf{X}_1$  and  $\mathbf{X}_2$ , and so lie on this plane. The fitted value  $\hat{\mathbf{y}}$  is the vector on this plane which is closest to  $\mathbf{y}$ . The residual  $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}$  is the difference between the two. The angle between the vectors  $\hat{\mathbf{y}}$  and  $\hat{\mathbf{e}}$  must be  $90^\circ$ , and therefore are orthogonal as shown.

## 3.16 Regression Components

Partition

$$\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2]$$

and

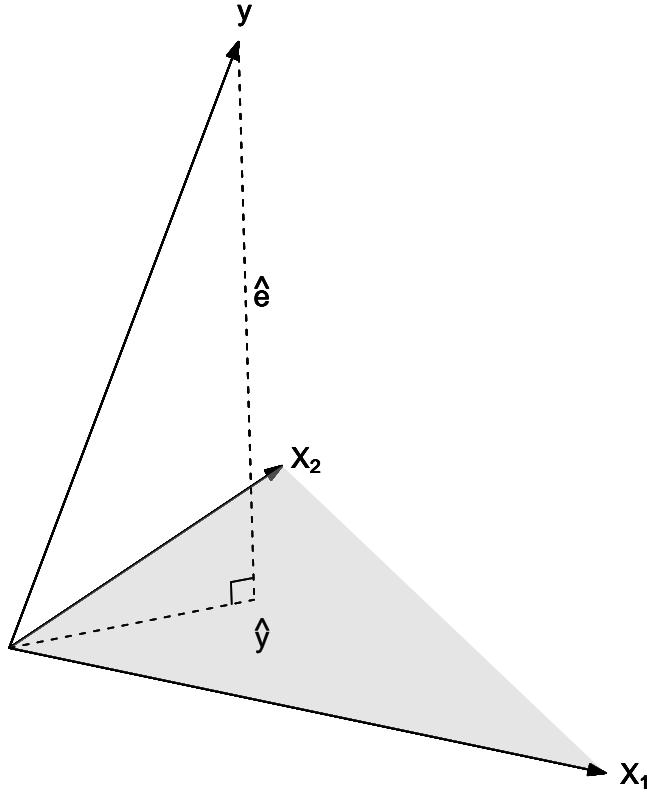
$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}.$$

Then the regression model can be rewritten as

$$\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \mathbf{e}. \quad (3.32)$$

The OLS estimator of  $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$  is obtained by regression of  $\mathbf{y}$  on  $\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2]$  and can be written as

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{e}} = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2\hat{\boldsymbol{\beta}}_2 + \hat{\mathbf{e}}. \quad (3.33)$$

Figure 3.6: Projection of  $y$  onto  $X_1$  and  $X_2$ 

We are interested in algebraic expressions for  $\hat{\beta}_1$  and  $\hat{\beta}_2$ .

Let's focus on finding an algebraic expression for  $\hat{\beta}_1$ . The least-squares estimator by definition is found by the joint minimization

$$(\hat{\beta}_1, \hat{\beta}_2) = \underset{\beta_1, \beta_2}{\operatorname{argmin}} \text{SSE}(\beta_1, \beta_2) \quad (3.34)$$

where

$$\text{SSE}(\beta_1, \beta_2) = (y - X_1\beta_1 - X_2\beta_2)'(y - X_1\beta_1 - X_2\beta_2).$$

An equivalent expression for  $\hat{\beta}_1$  can be obtained by concentration. The solution (3.34) can be written as

$$\hat{\beta}_1 = \underset{\beta_1}{\operatorname{argmin}} \left( \underset{\beta_2}{\operatorname{min}} \text{SSE}(\beta_1, \beta_2) \right). \quad (3.35)$$

The inner expression  $\underset{\beta_2}{\operatorname{min}} \text{SSE}(\beta_1, \beta_2)$  minimizes over  $\beta_2$  while holding  $\beta_1$  fixed. It is the lowest possible sum of squared errors given  $\beta_1$ . The outer minimization  $\underset{\beta_1}{\operatorname{argmin}}$  finds the coefficient  $\beta_1$  which

minimizes the “lowest possible sum of squared errors given  $\beta_1$ ”. This means that  $\hat{\beta}_1$  as defined in (3.34) and (3.35) are algebraically identical.

Examine the inner minimization problem in (3.35). This is simply the least squares regression of  $\mathbf{y} - \mathbf{X}_1\beta_1$  on  $\mathbf{X}_2$ . This has solution

$$\underset{\beta_2}{\operatorname{argmin}} \text{SSE}(\beta_1, \beta_2) = (\mathbf{X}'_2 \mathbf{X}_2)^{-1} (\mathbf{X}'_2 (\mathbf{y} - \mathbf{X}_1 \beta_1))$$

with residuals

$$\begin{aligned} \mathbf{y} - \mathbf{X}_1 \beta_1 - \mathbf{X}_2 (\mathbf{X}'_2 \mathbf{X}_2)^{-1} (\mathbf{X}'_2 (\mathbf{y} - \mathbf{X}_1 \beta_1)) &= (\mathbf{M}_2 \mathbf{y} - \mathbf{M}_2 \mathbf{X}_1 \beta_1) \\ &= \mathbf{M}_2 (\mathbf{y} - \mathbf{X}_1 \beta_1) \end{aligned}$$

where

$$\mathbf{M}_2 = \mathbf{I}_n - \mathbf{X}_2 (\mathbf{X}'_2 \mathbf{X}_2)^{-1} \mathbf{X}'_2 \quad (3.36)$$

is the orthogonal projection matrix for  $\mathbf{X}_2$ . This means that the inner minimization problem (3.35) has minimized value

$$\begin{aligned} \underset{\beta_2}{\operatorname{min}} \text{SSE}(\beta_1, \beta_2) &= (\mathbf{y} - \mathbf{X}_1 \beta_1)' \mathbf{M}_2 \mathbf{M}_2 (\mathbf{y} - \mathbf{X}_1 \beta_1) \\ &= (\mathbf{y} - \mathbf{X}_1 \beta_1)' \mathbf{M}_2 (\mathbf{y} - \mathbf{X}_1 \beta_1) \end{aligned}$$

where the second equality holds since  $\mathbf{M}_2$  is idempotent. Substituting this into (3.35) we find

$$\begin{aligned} \hat{\beta}_1 &= \underset{\beta_1}{\operatorname{argmin}} (\mathbf{y} - \mathbf{X}_1 \beta_1)' \mathbf{M}_2 (\mathbf{y} - \mathbf{X}_1 \beta_1) \\ &= (\mathbf{X}'_1 \mathbf{M}_2 \mathbf{X}_1)^{-1} (\mathbf{X}'_1 \mathbf{M}_2 \mathbf{y}). \end{aligned}$$

By a similar argument we can find

$$\hat{\beta}_2 = (\mathbf{X}'_2 \mathbf{M}_1 \mathbf{X}_2)^{-1} (\mathbf{X}'_2 \mathbf{M}_1 \mathbf{y})$$

where

$$\mathbf{M}_1 = \mathbf{I}_n - \mathbf{X}_1 (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \quad (3.37)$$

is the orthogonal projection matrix for  $\mathbf{X}_1$ .

**Theorem 3.4** The least-squares estimator  $(\hat{\beta}_1, \hat{\beta}_2)$  for (3.33) has the algebraic solution

$$\hat{\beta}_1 = (\mathbf{X}'_1 \mathbf{M}_2 \mathbf{X}_1)^{-1} (\mathbf{X}'_1 \mathbf{M}_2 \mathbf{y}) \quad (3.38)$$

$$\hat{\beta}_2 = (\mathbf{X}'_2 \mathbf{M}_1 \mathbf{X}_2)^{-1} (\mathbf{X}'_2 \mathbf{M}_1 \mathbf{y}) \quad (3.39)$$

where  $\mathbf{M}_1$  and  $\mathbf{M}_2$  are defined in (3.37) and (3.36), respectively.

### 3.17 Regression Components (Alternative Derivation)\*

An alternative proof of Theorem 3.4 uses an algebraic argument which is identical to that for the population coefficients as presented in Section 2.22. Since this is a classic derivation we present it here for completeness.

Partition  $\hat{\mathbf{Q}}_{xx}$  as

$$\hat{\mathbf{Q}}_{xx} = \begin{bmatrix} \hat{\mathbf{Q}}_{11} & \hat{\mathbf{Q}}_{12} \\ \hat{\mathbf{Q}}_{21} & \hat{\mathbf{Q}}_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\mathbf{X}'_1\mathbf{X}_1 & \frac{1}{n}\mathbf{X}'_1\mathbf{X}_2 \\ \frac{1}{n}\mathbf{X}'_2\mathbf{X}_1 & \frac{1}{n}\mathbf{X}'_2\mathbf{X}_2 \end{bmatrix}$$

and similarly  $\hat{\mathbf{Q}}_{xy}$  as

$$\hat{\mathbf{Q}}_{xy} = \begin{bmatrix} \hat{\mathbf{Q}}_{1y} \\ \hat{\mathbf{Q}}_{2y} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\mathbf{X}'_1\mathbf{y} \\ \frac{1}{n}\mathbf{X}'_2\mathbf{y} \end{bmatrix}.$$

By the partitioned matrix inversion formula (A.3)

$$\hat{\mathbf{Q}}_{xx}^{-1} = \begin{bmatrix} \hat{\mathbf{Q}}_{11} & \hat{\mathbf{Q}}_{12} \\ \hat{\mathbf{Q}}_{21} & \hat{\mathbf{Q}}_{22} \end{bmatrix}^{-1} \stackrel{\text{def}}{=} \begin{bmatrix} \hat{\mathbf{Q}}^{11} & \hat{\mathbf{Q}}^{12} \\ \hat{\mathbf{Q}}^{21} & \hat{\mathbf{Q}}^{22} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{Q}}_{11\cdot 2}^{-1} & -\hat{\mathbf{Q}}_{11\cdot 2}^{-1}\hat{\mathbf{Q}}_{12}\hat{\mathbf{Q}}_{22}^{-1} \\ -\hat{\mathbf{Q}}_{22\cdot 1}^{-1}\hat{\mathbf{Q}}_{21}\hat{\mathbf{Q}}_{11}^{-1} & \hat{\mathbf{Q}}_{22\cdot 1}^{-1} \end{bmatrix} \quad (3.40)$$

where  $\hat{\mathbf{Q}}_{11\cdot 2} = \hat{\mathbf{Q}}_{11} - \hat{\mathbf{Q}}_{12}\hat{\mathbf{Q}}_{22}^{-1}\hat{\mathbf{Q}}_{21}$  and  $\hat{\mathbf{Q}}_{22\cdot 1} = \hat{\mathbf{Q}}_{22} - \hat{\mathbf{Q}}_{21}\hat{\mathbf{Q}}_{11}^{-1}\hat{\mathbf{Q}}_{12}$ . Thus

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} \\ &= \begin{bmatrix} \hat{\mathbf{Q}}_{11\cdot 2}^{-1} & -\hat{\mathbf{Q}}_{11\cdot 2}^{-1}\hat{\mathbf{Q}}_{12}\hat{\mathbf{Q}}_{22}^{-1} \\ -\hat{\mathbf{Q}}_{22\cdot 1}^{-1}\hat{\mathbf{Q}}_{21}\hat{\mathbf{Q}}_{11}^{-1} & \hat{\mathbf{Q}}_{22\cdot 1}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{Q}}_{1y} \\ \hat{\mathbf{Q}}_{2y} \end{bmatrix} \\ &= \begin{pmatrix} \hat{\mathbf{Q}}_{11\cdot 2}^{-1}\hat{\mathbf{Q}}_{1y\cdot 2} \\ \hat{\mathbf{Q}}_{22\cdot 1}^{-1}\hat{\mathbf{Q}}_{2y\cdot 1} \end{pmatrix}. \end{aligned}$$

Now

$$\begin{aligned} \hat{\mathbf{Q}}_{11\cdot 2} &= \hat{\mathbf{Q}}_{11} - \hat{\mathbf{Q}}_{12}\hat{\mathbf{Q}}_{22}^{-1}\hat{\mathbf{Q}}_{21} \\ &= \frac{1}{n}\mathbf{X}'_1\mathbf{X}_1 - \frac{1}{n}\mathbf{X}'_1\mathbf{X}_2 \left( \frac{1}{n}\mathbf{X}'_2\mathbf{X}_2 \right)^{-1} \frac{1}{n}\mathbf{X}'_2\mathbf{X}_1 \\ &= \frac{1}{n}\mathbf{X}'_1\mathbf{M}_2\mathbf{X}_1 \end{aligned}$$

and

$$\begin{aligned} \hat{\mathbf{Q}}_{1y\cdot 2} &= \hat{\mathbf{Q}}_{1y} - \hat{\mathbf{Q}}_{12}\hat{\mathbf{Q}}_{22}^{-1}\hat{\mathbf{Q}}_{2y} \\ &= \frac{1}{n}\mathbf{X}'_1\mathbf{y} - \frac{1}{n}\mathbf{X}'_1\mathbf{X}_2 \left( \frac{1}{n}\mathbf{X}'_2\mathbf{X}_2 \right)^{-1} \frac{1}{n}\mathbf{X}'_2\mathbf{y} \\ &= \frac{1}{n}\mathbf{X}'_1\mathbf{M}_2\mathbf{y}. \end{aligned}$$

Equation (3.39) follows.

Similarly to the calculation for  $\hat{\mathbf{Q}}_{11\cdot 2}$  and  $\hat{\mathbf{Q}}_{1y\cdot 2}$  you can show that  $\hat{\mathbf{Q}}_{2y\cdot 1} = \frac{1}{n}\mathbf{X}'_2\mathbf{M}_1\mathbf{y}$  and  $\hat{\mathbf{Q}}_{22\cdot 1} = \frac{1}{n}\mathbf{X}'_2\mathbf{M}_1\mathbf{X}_2$ . This establishes (3.38). Together, this is Theorem 3.4.

### 3.18 Residual Regression

As first recognized by Frisch and Waugh (1933) and extended by Lovell (1963), expressions (3.38) and (3.39) can be used to show that the least-squares estimators  $\hat{\beta}_1$  and  $\hat{\beta}_2$  can be found by a two-step regression procedure.

Take (3.39). Since  $\mathbf{M}_1$  is idempotent,  $\mathbf{M}_1 = \mathbf{M}_1 \mathbf{M}_1$  and thus

$$\begin{aligned}\hat{\beta}_2 &= (\mathbf{X}'_2 \mathbf{M}_1 \mathbf{X}_2)^{-1} (\mathbf{X}'_2 \mathbf{M}_1 \mathbf{y}) \\ &= (\mathbf{X}'_2 \mathbf{M}_1 \mathbf{M}_1 \mathbf{X}_2)^{-1} (\mathbf{X}'_2 \mathbf{M}_1 \mathbf{M}_1 \mathbf{y}) \\ &= (\tilde{\mathbf{X}}'_2 \tilde{\mathbf{X}}_2)^{-1} (\tilde{\mathbf{X}}'_2 \tilde{\mathbf{e}}_1)\end{aligned}$$

where

$$\tilde{\mathbf{X}}_2 = \mathbf{M}_1 \mathbf{X}_2$$

and

$$\tilde{\mathbf{e}}_1 = \mathbf{M}_1 \mathbf{y}.$$

Thus the coefficient estimate  $\hat{\beta}_2$  is algebraically equal to the least-squares regression of  $\tilde{\mathbf{e}}_1$  on  $\tilde{\mathbf{X}}_2$ . Notice that these two are  $\mathbf{y}$  and  $\mathbf{X}_2$ , respectively, premultiplied by  $\mathbf{M}_1$ . But we know that multiplication by  $\mathbf{M}_1$  is equivalent to creating least-squares residuals. Therefore  $\tilde{\mathbf{e}}_1$  is simply the least-squares residual from a regression of  $\mathbf{y}$  on  $\mathbf{X}_1$ , and the columns of  $\tilde{\mathbf{X}}_2$  are the least-squares residuals from the regressions of the columns of  $\mathbf{X}_2$  on  $\mathbf{X}_1$ .

We have proven the following theorem.

**Theorem 3.5 Frisch-Waugh-Lovell (FWL)**

In the model (3.32), the OLS estimator of  $\beta_2$  and the OLS residuals  $\hat{\mathbf{e}}$  may be equivalently computed by either the OLS regression (3.33) or via the following algorithm:

1. Regress  $\mathbf{y}$  on  $\mathbf{X}_1$ , obtain residuals  $\tilde{\mathbf{e}}_1$ ;
2. Regress  $\mathbf{X}_2$  on  $\mathbf{X}_1$ , obtain residuals  $\tilde{\mathbf{X}}_2$ ;
3. Regress  $\tilde{\mathbf{e}}_1$  on  $\tilde{\mathbf{X}}_2$ , obtain OLS estimates  $\hat{\beta}_2$  and residuals  $\hat{\mathbf{e}}$ .

In some contexts (such as panel data models, to be introduced in Chapter 17), the FWL theorem can be used to greatly speed computation.

The FWL theorem is a direct analogy of the coefficient representation obtained in Section 2.23. The result obtained in that section concerned the population projection coefficients, the result obtained here concern the least-squares estimates. The key message is the same. In the least-squares regression (3.33), the estimated coefficient  $\hat{\beta}_2$  numerically equals the regression of  $\mathbf{y}$  on the regressors  $\mathbf{X}_2$ , only after the regressors  $\mathbf{X}_1$  have been linearly projected out. Similarly, the coefficient estimate  $\hat{\beta}_1$  numerically equals the regression of  $\mathbf{y}$  on the regressors  $\mathbf{X}_1$ , after the regressors  $\mathbf{X}_2$  have been linearly projected out. This result can be very insightful when interpreting regression coefficients.

A common application of the FWL theorem is the demeaning formula for regression obtained in (3.19). Partition  $\mathbf{X} = [\mathbf{X}_1 \mathbf{X}_2]$  where  $\mathbf{X}_1 = \mathbf{1}_n$  is a vector of ones and  $\mathbf{X}_2$  is a matrix of observed regressors. In this case,

$$\mathbf{M}_1 = \mathbf{I}_n - \mathbf{1}_n (\mathbf{1}'_n \mathbf{1}_n)^{-1} \mathbf{1}'_n.$$

Observe that

$$\tilde{\mathbf{X}}_2 = \mathbf{M}_1 \mathbf{X}_2 = \mathbf{X}_2 - \bar{\mathbf{X}}_2$$

and

$$\mathbf{M}_1 \mathbf{y} = \mathbf{y} - \bar{\mathbf{y}}$$

are the “demeaned” variables. The FWL theorem says that  $\hat{\beta}_2$  is the OLS estimate from a regression of  $y_i - \bar{y}$  on  $\mathbf{x}_{2i} - \bar{\mathbf{x}}_2$ :

$$\hat{\beta}_2 = \left( \sum_{i=1}^n (\mathbf{x}_{2i} - \bar{\mathbf{x}}_2) (\mathbf{x}_{2i} - \bar{\mathbf{x}}_2)' \right)^{-1} \left( \sum_{i=1}^n (\mathbf{x}_{2i} - \bar{\mathbf{x}}_2) (y_i - \bar{y}) \right).$$

This is (3.19).

### Ragnar Frisch

Ragnar Frisch (1895-1973) was co-winner with Jan Tinbergen of the first Nobel Memorial Prize in Economic Sciences in 1969 for their work in developing and applying dynamic models for the analysis of economic problems. Frisch made a number of foundational contributions to modern economics beyond the Frisch-Waugh-Lovell Theorem, including formalizing consumer theory, production theory, and business cycle theory.

## 3.19 Leverage Values

The **leverage** values for the regressor matrix  $\mathbf{X}$  are the diagonal elements of the projection matrix  $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ . There are  $n$  leverage values, and are typically written as  $h_{ii}$  for  $i = 1, \dots, n$ . Since

$$\mathbf{P} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} (\mathbf{X}'\mathbf{X})^{-1} \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{pmatrix}$$

they are

$$h_{ii} = \mathbf{x}_i' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i. \quad (3.41)$$

The leverage value  $h_{ii}$  is a normalized length of the observed regressor vector  $\mathbf{x}_i$ . They appear frequently in the algebraic and statistical analysis of least-squares regression, including leave-one-out regression, influential observations, robust covariance matrix estimation, and cross-validation.

A few properties of the leverage values are now listed.

### Theorem 3.6

1.  $0 \leq h_{ii} \leq 1$ .
2.  $h_{ii} \geq 1/n$  if  $\mathbf{X}$  includes an intercept.
3.  $\sum_{i=1}^n h_{ii} = k$ .

We prove Theorem 3.6 below.

The leverage values  $h_{ii}$  measure how unusual the  $i^{th}$  observation  $\mathbf{x}_i$  is relative to the other values in the sample. A large  $h_{ii}$  occurs when  $\mathbf{x}_i$  is quite different from the other sample values. A measure of overall unusualness is the maximum leverage value

$$\bar{h} = \max_{1 \leq i \leq n} h_{ii}. \quad (3.42)$$

It is common to say that a regression design is **balanced** when the leverage values are all roughly equal to one another. From Theorem 3.6.3 we can deduce that complete balance implies  $h_{ii} = \bar{h} = k/n$ . An example where complete balance occurs is when the regressors are all orthogonal dummy variables, each of which have equal occurrence of 0's and 1's.

A regression design is **unbalanced** if some leverage values are highly unequal from the others. The most extreme case is  $\bar{h} = 1$ . An example where this occurs is when there is a dummy regressor which takes the value 1 for only one observation in the sample.

The maximal leverage value (3.42) will change depending on the choice of regressors. For example, consider equation (3.14), the wage regression for single asian men which has  $n = 268$  observations. This regression has  $\bar{h} = 0.33$ . If the squared experience regressor is omitted, the leverage drops to  $\bar{h} = 0.10$ . If a cubic in experience is added, it increases to  $\bar{h} = 0.76$ . And if a fourth and fifth power are added, it increases to  $\bar{h} = 0.99$ .

In general, there is no reason to check the leverage values, as in general there is no problem if the leverage values are balanced, unbalanced, or even highly unbalanced. However, the fact that leverage values can easily be large and close to one suggests that we should take this into consideration when examining procedures (such as robust covariance matrix estimation and cross-validation) which make use of leverage values. We will return to these issues later when leverage values arise.

We now prove Theorem 3.6. For part 1, let  $\mathbf{s}_i$  be an  $n \times 1$  unit vector with a 1 in the  $i^{th}$  place and zeros elsewhere, so that  $h_{ii} = \mathbf{s}_i' \mathbf{P} \mathbf{s}_i$ . Then applying the Quadratic Inequality (B.18) and Theorem 3.3.4,

$$h_{ii} = \mathbf{s}_i' \mathbf{P} \mathbf{s}_i \leq \mathbf{s}_i' \mathbf{s}_i \lambda_{\max}(\mathbf{P}) = 1$$

as claimed.

For part 2, partition  $\mathbf{x}_i = (1, \mathbf{z}'_i)'$ . Without loss of generality we can replace  $\mathbf{z}_i$  with the demeaned values  $\mathbf{z}_i^* = \mathbf{z}_i - \bar{\mathbf{z}}$ . Then since  $\mathbf{z}_i^*$  and the intercept are orthogonal,

$$\begin{aligned} h_{ii} &= (1, \mathbf{z}_i^{*\prime}) \begin{bmatrix} n & 0 \\ 0 & \mathbf{Z}^{*\prime} \mathbf{Z}^* \end{bmatrix}^{-1} \begin{pmatrix} 1 \\ \mathbf{z}_i^* \end{pmatrix} \\ &= \frac{1}{n} + \mathbf{z}_i^{*\prime} (\mathbf{Z}^{*\prime} \mathbf{Z}^*)^{-1} \mathbf{z}_i^* \\ &\geq \frac{1}{n}. \end{aligned}$$

For part 3,  $\sum_{i=1}^n h_{ii} = \text{tr } \mathbf{P} = k$  where the second equality is Theorem 3.3.3.

## 3.20 Leave-One-Out Regression

There are a number of statistical procedures – residual analysis, jackknife variance estimation, cross-validation, two-step estimation, hold-out sample evaluation – which make use of estimators constructed on sub-samples. Of particular importance is the case where we exclude a single observation and then repeat this for all observations. This is called **leave-one-out** (LOO) regression.

Specifically, the leave-one-out least-squares estimator of the regression coefficient  $\boldsymbol{\beta}$  is the least-squares estimator constructed using the full sample excluding a single observation  $i$ . This can be written

as

$$\begin{aligned}\hat{\beta}_{(-i)} &= \left( \sum_{j \neq i} \mathbf{x}_j \mathbf{x}'_j \right)^{-1} \left( \sum_{j \neq i} \mathbf{x}_j y_j \right) \\ &= (\mathbf{X}' \mathbf{X} - \mathbf{x}_i \mathbf{x}'_i)^{-1} (\mathbf{X}' \mathbf{y} - \mathbf{x}_i y_i) \\ &= (\mathbf{X}'_{(-i)} \mathbf{X}_{(-i)})^{-1} \mathbf{X}'_{(-i)} \mathbf{y}_{(-i)}.\end{aligned}\quad (3.43)$$

Here,  $\mathbf{X}_{(-i)}$  and  $\mathbf{y}_{(-i)}$  are the data matrices omitting the  $i^{th}$  row. The notation  $\hat{\beta}_{(-i)}$  or  $\hat{\beta}_{-i}$  is commonly used to denote an estimator with the  $i^{th}$  observation omitted.

There is a leave-one-out estimator for each observation,  $i = 1, \dots, n$ , so we have  $n$  such estimators.

The leave-one-out predicted value for  $y_i$  is

$$\tilde{y}_i = \mathbf{x}'_i \hat{\beta}_{(-i)}.$$

This is the predicted value obtained by estimating  $\beta$  on the sample without observation  $i$ , and then using the covariate vector  $\mathbf{x}_i$  to predict  $y_i$ . Notice that  $\tilde{y}_i$  is an authentic prediction as  $y_i$  is not used to construct  $\tilde{y}_i$ . This is in contrast to the fitted values  $\hat{y}_i$  which are functions of  $y_i$ .

The **leave-one-out residual, prediction error, or prediction residual** is

$$\tilde{e}_i = y_i - \tilde{y}_i.$$

The prediction errors may be used as estimates of the errors instead of the residuals. The prediction errors are better estimates than the residuals, since the former are based on authentic predictions.

The leave-one-out formula (3.43) gives the unfortunate impression that the leave-one-out coefficients and errors are computationally cumbersome, requiring  $n$  separate regressions. In the context of linear regression this is fortunately not the case. There are simple linear expressions for  $\hat{\beta}_{(-i)}$  and  $\tilde{e}_i$ .

**Theorem 3.7** The leave-one-out least-squares estimator and prediction error can be calculated as

$$\hat{\beta}_{(-i)} = \hat{\beta} - (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i \tilde{e}_i \quad (3.44)$$

and

$$\tilde{e}_i = (1 - h_{ii})^{-1} \hat{e}_i \quad (3.45)$$

where  $h_{ii}$  are the leverage values as defined in (3.41).

We prove Theorem 3.7 at the end of the section.

Equation (3.44) shows that the leave-one-out coefficients can be calculated by a simple linear operation and do not need to be calculated using  $n$  separate regressions. Equation (3.45) for the prediction error is particularly convenient. It shows that the leave-one-out residuals are a simple scaling of the standard least-squares residuals.

Equations (3.44) and (3.45) both show the usefulness of the leverage values  $h_{ii}$ .

Another interesting feature of equation (3.45) is that the prediction errors  $\tilde{e}_i$  are a simple scaling of the residuals  $\hat{e}_i$ , with the scaling depending on the leverage values  $h_{ii}$ . If  $h_{ii}$  is small then  $\tilde{e}_i \approx \hat{e}_i$ . However if  $h_{ii}$  is large then  $\tilde{e}_i$  can be quite different from  $\hat{e}_i$ . Thus the difference between the residuals and predicted values depends on the leverage values, that is, how unusual  $\mathbf{x}_i$  is relative to the other observations.

To write (3.45) in vector notation, define

$$\begin{aligned}\mathbf{M}^* &= (\mathbf{I}_n - \text{diag}\{h_{11}, \dots, h_{nn}\})^{-1} \\ &= \text{diag}\{(1 - h_{11})^{-1}, \dots, (1 - h_{nn})^{-1}\}.\end{aligned}$$

Then (3.45) is equivalent to

$$\tilde{\mathbf{e}} = \mathbf{M}^* \hat{\mathbf{e}}. \quad (3.46)$$

One use of the prediction errors is to estimate the out-of-sample mean squared error. The natural estimator is

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \tilde{e}_i^2 = \frac{1}{n} \sum_{i=1}^n (1 - h_{ii})^{-2} \hat{e}_i^2. \quad (3.47)$$

This is also known as the **sample mean squared prediction error**. Its square root  $\tilde{\sigma} = \sqrt{\tilde{\sigma}^2}$  is the **prediction standard error**.

We complete the section by presenting a proof of Theorem 3.7. The leave-one-out estimator (3.43) can be written as

$$\hat{\boldsymbol{\beta}}_{(-i)} = (\mathbf{X}' \mathbf{X} - \mathbf{x}_i \mathbf{x}_i')^{-1} (\mathbf{X}' \mathbf{y} - \mathbf{x}_i y_i). \quad (3.48)$$

Multiply (3.48) by  $(\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{X} - \mathbf{x}_i \mathbf{x}_i')$ . We obtain

$$\hat{\boldsymbol{\beta}}_{(-i)} - (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{(-i)} = (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{y} - \mathbf{x}_i y_i) = \hat{\boldsymbol{\beta}} - (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i y_i.$$

Rewriting

$$\hat{\boldsymbol{\beta}}_{(-i)} = \hat{\boldsymbol{\beta}} - (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i (y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{(-i)}) = \hat{\boldsymbol{\beta}} - (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i \tilde{e}_i$$

which is (3.44). Premultiplying this expression by  $\mathbf{x}_i'$  and using definition (3.41) we obtain

$$\mathbf{x}_i' \hat{\boldsymbol{\beta}}_{(-i)} = \mathbf{x}_i' \hat{\boldsymbol{\beta}} - \mathbf{x}_i' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i \tilde{e}_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}} - h_{ii} \tilde{e}_i.$$

Using the definitions for  $\tilde{e}_i$  and  $\tilde{e}_i$  we obtain  $\tilde{e}_i = \hat{e}_i - h_{ii} \tilde{e}_i$ . Re-writing we obtain (3.45).

## 3.21 Influential Observations

Another use of the leave-one-out estimator is to investigate the impact of **influential observations**, sometimes called **outliers**. We say that observation  $i$  is influential if its omission from the sample induces a substantial change in a parameter estimate of interest.

For illustration, consider Figure 3.7 which shows a scatter plot of random variables  $(y_i, x_i)$ . The 25 observations shown with the open circles are generated by  $x_i \sim U[1, 10]$  and  $y_i \sim N(x_i, 4)$ . The 26<sup>th</sup> observation shown with the filled circle is  $x_{26} = 9$ ,  $y_{26} = 0$ . (Imagine that  $y_{26} = 0$  was incorrectly recorded due to a mistaken key entry.) The figure shows both the least-squares fitted line from the full sample and that obtained after deletion of the 26<sup>th</sup> observation from the sample. In this example we can see how the 26<sup>th</sup> observation (the “outlier”) greatly tilts the least-squares fitted line towards the 26<sup>th</sup> observation. In fact, the slope coefficient decreases from 0.97 (which is close to the true value of 1.00) to 0.56, which is substantially reduced. Neither  $y_{26}$  nor  $x_{26}$  are unusual values relative to their marginal distributions, so this outlier would not have been detected from examination of the marginal distributions of the data. The change in the slope coefficient of  $-0.41$  is meaningful and should raise concern to an applied economist.

From (3.44) we know that

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i \tilde{e}_i. \quad (3.49)$$

By direct calculation of this quantity for each observation  $i$ , we can directly discover if a specific observation  $i$  is influential for a coefficient estimate of interest.

For a general assessment, we can focus on the predicted values. The difference between the full-sample and leave-one-out predicted values is

$$\begin{aligned} \hat{y}_i - \tilde{y}_i &= \mathbf{x}_i' \hat{\boldsymbol{\beta}} - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{(-i)} \\ &= \mathbf{x}_i' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i \tilde{e}_i \\ &= h_{ii} \tilde{e}_i \end{aligned}$$

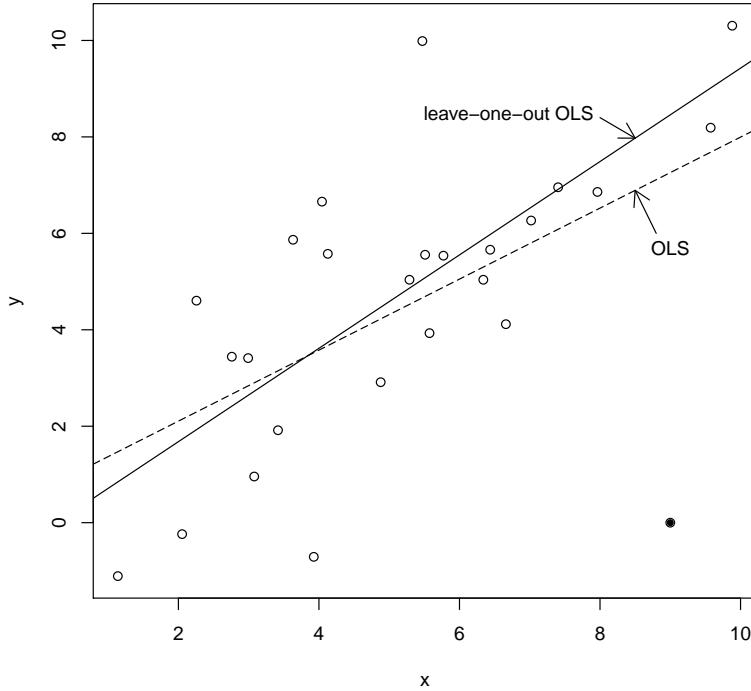


Figure 3.7: Impact of an Influential Observation on the Least-Squares Estimator

which is a simple function of the leverage values  $h_{ii}$  and prediction errors  $\tilde{e}_i$ . Observation  $i$  is influential for the predicted value if  $|h_{ii}\tilde{e}_i|$  is large, which requires that both  $h_{ii}$  and  $|\tilde{e}_i|$  are large.

One way to think about this is that a large leverage value  $h_{ii}$  gives the potential for observation  $i$  to be influential. A large  $h_{ii}$  means that observation  $i$  is unusual in the sense that the regressor  $x_i$  is far from its sample mean. We call an observation with large  $h_{ii}$  a **leverage point**. A leverage point is not necessarily influential as the latter also requires that the prediction error  $\tilde{e}_i$  is large.

To determine if any individual observations are influential in this sense, several diagnostics have been proposed (some names include DFITS, Cook's Distance, and Welsch Distance). Unfortunately, from a statistical perspective it is difficult to recommend these diagnostics for applications as they are not based on statistical theory. Probably the most relevant measure is the change in the coefficient estimates given in (3.49). The ratio of these changes to the coefficient's standard error is called its DFBETA, and is a postestimation diagnostic available in Stata. While there is no magic threshold, the concern is whether or not an individual observation meaningfully changes an estimated coefficient of interest. A simple diagnostic for influential observations is to calculate

$$\text{Influence} = \max_{1 \leq i \leq n} |\hat{y}_i - \tilde{y}_i| = \max_{1 \leq i \leq n} |h_{ii}\tilde{e}_i|.$$

This is the largest (absolute) change in the predicted value due to a single observation. If this diagnostic is large relative to the distribution of  $y_i$ , it may indicate that that observation is influential.

If an observation is determined to be influential, what should be done? As a common cause of influential observations is data entry error, the influential observations should be examined for evidence that the observation was mis-recorded. Perhaps the observation falls outside of permitted ranges, or some observables are inconsistent (for example, a person is listed as having a job but receives earnings of \$0). If it is determined that an observation is incorrectly recorded, then the observation is typically deleted from the sample. This process is often called "cleaning the data". The decisions made in this process involve a fair amount of individual judgment. [When this is done the proper practice is to retain the source

data in its original form and create a program file which executes all cleaning operations (for example deletion of individual observations). The cleaned data file can be saved at this point, and then used for the subsequent statistical analysis. The point of retaining the source data and a specific program file which cleans the data is twofold: so that all decisions are documented, and so that modifications can be made in revisions and future research.] It is also possible that an observation is correctly measured, but unusual and influential. In this case it is unclear how to proceed. Some researchers will try to alter the specification to properly model the influential observation. Other researchers will delete the observation from the sample. The motivation for this choice is to prevent the results from being skewed or determined by individual observations, but this practice is viewed skeptically by many researchers who believe it reduces the integrity of reported empirical results.

For an empirical illustration, consider the log wage regression (3.14) for single Asian males. This regression, which has 268 observations, has  $\text{Influence} = 0.29$ . This means that the most influential observation, when deleted, changes the predicted (fitted) value of the dependent variable  $\log(Wage)$  by 0.29, or equivalently the wage by 29%. This is a meaningful change and suggests further investigation. We examine the influential observation, and find that its leverage  $h_{ii}$  is 0.33, which is the maximum in the sample as described in Section 3.19. It is a rather large leverage value, meaning that the regressor  $x_i$  is unusual. Examining further, we find that this individual is 65 years old with 8 years education, so that his potential experience is 51 years. This is the highest experience in the subsample – the next highest is 41 years. The large leverage is due to his unusual characteristics (very low education and very high experience) within this sample. Essentially, regression (3.14) is attempting to estimate the conditional mean at  $experience = 51$  with only one observation, so it is not surprising that this observation determines the fit and is thus influential. A reasonable conclusion is the regression function can only be estimated over a smaller range of  $experience$ . We restrict the sample to individuals with less than 45 years experience, re-estimate, and obtain the following estimates.

$$\widehat{\log(Wage)} = 0.144 \text{ education} + 0.043 \text{ experience} - 0.095 \text{ experience}^2 / 100 + 0.531. \quad (3.50)$$

For this regression, we calculate that  $\text{Influence} = 0.11$ , which is greatly reduced relative to the regression (3.14). Comparing (3.50) with (3.14), the slope coefficient for education is essentially unchanged, but the coefficients in experience and its square have slightly increased.

By eliminating the influential observation, equation (3.50) can be viewed as a more robust estimate of the conditional mean for most levels of  $experience$ . Whether to report (3.14) or (3.50) in an application is largely a matter of judgment.

## 3.22 CPS Data Set

In this section we describe the data set used in the empirical illustrations.

The Current Population Survey (CPS) is a monthly survey of about 57,000 U.S. households conducted by the Bureau of the Census of the Bureau of Labor Statistics. The CPS is the primary source of information on the labor force characteristics of the U.S. population. The survey covers employment, earnings, educational attainment, income, poverty, health insurance coverage, job experience, voting and registration, computer usage, veteran status, and other variables. Details can be found at [www.census.gov/cps](http://www.census.gov/cps) and [dataferrett.census.gov](http://dataferrett.census.gov).

From the March 2009 survey we extracted the individuals with non-allocated variables who were full-time employed (defined as those who had worked at least 36 hours per week for at least 48 weeks the past year), and excluded those in the military. This sample has 50,742 individuals. We extracted 14 variables from the CPS on these individuals and created the data files `cps09mar.dta` (Stata format), `cps09mar.xlsx` (Excel format) and `cps09mar.txt` (text format). The variables are described in the file `cps09mar_description.pdf`. All data files are available at <http://www.ssc.wisc.edu/~bhansen/econometrics/>

## 3.23 Numerical Computation

Modern econometric estimation involves large samples and many covariates. Consequently calculation of even simple statistics such as the least squares estimator requires a large number (millions) of arithmetic operations. In practice most economists don't need to think much about this as it is done swiftly and effortlessly on our personal computers. Nevertheless it is useful to understand the underlying calculation methods as occasionally choices can make substantive differences.

While today nearly all statistical computations are made using statistical software running on personal computers, this was not always the case. In the nineteenth and early twentieth centuries, "computer" was a job label for workers who made computations by hand. Computers were employed by astronomers and statistical laboratories to execute numerical calculations. This fascinating job (and the fact that most computers employed in laboratories were women) has entered popular culture. For example the lives of several computers who worked for the early U.S. space program is described in the book and popular movie *Hidden Figures*, and the life of computer/astronomer Henrietta Swan Leavitt is dramatized in the moving play *Silent Sky*.

Until programmable electronic computers became available in the 1960s, economics graduate students were routinely employed as computers. Sample sizes were considerably smaller than those seen today, but still the effort required to calculate by hand (for example) a regression with  $n = 100$  observations and  $k = 5$  variables is considerable! If you are a current graduate student, you should feel fortunate that the profession has moved on from the era of human computers! (Now research assistants do more elevated tasks such as writing Stata and Matlab code.)

To obtain the least squares estimator  $\hat{\beta} = (X'X)^{-1}(X'y)$  we need to either invert  $X'X$  or solve a system of equations. To be specific, let  $A = X'X$  and  $c = X'y$  so that the least squares estimator can be written as either the solution to

$$A\hat{\beta} = c \quad (3.51)$$

or as

$$\hat{\beta} = A^{-1}c. \quad (3.52)$$

The equations (3.51) and (3.52) are algebraically identical, but they suggest two distinct numerical approaches to obtain  $\hat{\beta}$ . (3.51) suggests solving a system of  $k$  equations. (3.52) suggests finding  $A^{-1}$  and then multiplying by  $c$ . While the two expressions are algebraically identical, the implied numerical approaches are different.

In a nutshell, solving the system of equations (3.51) is numerically preferred to the matrix inversion problem (3.52). Directly solving (3.51) is faster and produces a solution with a higher degree of numerical accuracy. Thus (3.51) is generally recommended over (3.52). However, in most practical applications the choice will not make any practical difference. Contexts where the choice may make a difference is when the matrix  $A$  is ill-conditioned (to be discussed in Section 3.24) or of extremely high dimension.

Numerical methods to solve the system of equations (3.51) and calculate  $A^{-1}$  are discussed in Sections A.18 and A.19, respectively.

Statistical packages use a variety of matrix to solve (3.51). Stata uses the sweep algorithm, which is a variant of the Gauss-Jordan algorithm discussed in Section A.18. (For the sweep algorithm see Goodnight (1979).) In R, `solve(A, b)` uses the QR decomposition. In Matlab, `A\b` uses the Cholesky decomposition when  $A$  is positive definite and the QR decomposition otherwise.

## 3.24 Collinearity Errors

For the least squares estimator to be uniquely defined the regressors cannot be linearly dependent. However, it is quite easy to *attempt* to calculate a regression with linearly dependent regressors. This can occur for many reasons, including the following.

1. Including the same regressor twice.

2. Including regressors which are a linear combination of one another, such as *education*, *experience* and *age* in the CPS data set example (recall, *experience* is defined as *age-education-6*).
3. Including a dummy variable and its square.
4. Estimating a regression on a sub-sample for which a dummy variable is either all zeros or all ones.
5. Including a dummy variable interaction which yields all zeros.
6. Including more regressors than observations.

In any of the above cases the regressors are linearly dependent so  $\mathbf{X}'\mathbf{X}$  is singular and the least squares estimator is not defined. If you attempt to estimate the regression, you are likely to encounter an error message. (A possible exception is Matlab using “ $\mathbf{A}\backslash\mathbf{b}$ ”, as discussed below.) The message may be that “system is exactly singular”, “system is computationally singular”, a variable is “omitted because of collinearity”, or a coefficient is listed as “NA”. In some cases (such as estimation in R using explicit matrix computation or Matlab using the `regress` command) the program will stop execution. In other cases the program will continue to run. In Stata (and in the `lm` package in R), a regression will be reported but one or more variables will be omitted to achieve non-singularity.

If any of these warnings or error messages appear, the correct response is to stop and examine the regression coding and data. Did you make an unintended mistake? Have you included a linearly dependent regressor? Are you estimating on a subsample for which the variables (in particular dummy variables) have no variation? If you can determine that one of these scenarios caused the error, the solution is immediately apparent. You need to respecify your model (either sample or regressors) so that the redundancy is eliminated. All empirical researchers encounter this error in the course of empirical work. You should not, however, simply accept output if the package has selected variables for omission. It is the researcher’s job to understand the underlying cause and enact a suitable remedy.

There is also a possibility that the statistical package will not detect and report the matrix singularity. If you compute in Matlab using explicit matrix operations and use the recommended `A\b` command to compute the least squares estimator, Matlab may return a numerical solution without an error message even when the regressors are algebraically dependent. It is therefore recommended that you perform a numerical check for matrix singularity when using explicit matrix operations in Matlab.

How can we numerically check if a matrix  $\mathbf{A}$  is singular? A standard diagnostic is the **reciprocal condition number**

$$C = \frac{\lambda_{\min}(\mathbf{A})}{\lambda_{\max}(\mathbf{A})}.$$

If  $C = 0$  then  $\mathbf{A}$  is singular. If  $C = 1$  then  $\mathbf{A}$  is perfectly balanced. If  $C$  is extremely small we say that  $\mathbf{A}$  is **ill-conditioned**. The reciprocal condition number can be calculated in Matlab or R by the `rcond` command. Unfortunately, there is no accepted tolerance for how small  $C$  should be before regarding  $\mathbf{A}$  as numerically singular, in part since `rcond(A)` can return a positive (but small) result even if  $\mathbf{A}$  is algebraically singular. However, in double precision (which is typically used for computation) numerical accuracy is bounded by  $2^{-52} \approx 2e-16$ , suggesting the minimum bound  $C \geq 2e-16$ .

Checking for numerical singularity is complicated by the fact that low values of  $C$  can also be caused by unbalanced or highly correlated regressors.

To illustrate, consider a wage regression using the sample from (3.14) on powers of experience  $x$  from 1 through  $k$  (e.g.  $x, x^2, x^3, \dots, x^k$ ). We calculated the reciprocal condition number  $C$  for each  $k$ , and found that  $C$  is decreasing as  $k$  increases, indicating increasing ill-conditioning. Indeed, for  $k = 5$ , we find  $C = 6e-17$ , which is lower than double precision accuracy. This means that a regression on  $(x, x^2, x^3, x^4, x^5)$  is ill-conditioned. The regressor matrix, however, is not singular. The low value of  $C$  is not due to algebraic singularity, but rather is due to a lack of balance and high collinearity.

Ill-conditioned regressors have the potential problem that the numerical results (the reported coefficient estimates) will be inaccurate. It is not a major concern, as this only occurs in extreme cases,

and because high numerical accuracy is typically not a goal in econometric estimation. Nevertheless, we should try and avoid ill-conditioned regressions when possible.

There are strategies which can reduce or even eliminate ill-conditioning. Often it is sufficient to rescale the regressors. A simple rescaling which often works for non-negative regressors is to divide each by its sample mean, thus replace  $x_{ji}$  with  $x_{ji}/\bar{x}_j$ . In the above example with the powers of experience, this means replacing  $x_i^2$  with  $x_i^2/(n^{-1}\sum_{i=1}^n x_i^2)$ , etc. Doing so dramatically reduces the ill-conditioning. With this scaling, regressions for  $k \leq 11$  satisfy  $C \geq 1e-15$ . A rescaling specific to a regression with powers is to first rescale the regressor to lie in  $[-1, 1]$  before taking powers. With this scaling, regressions for  $k \leq 16$  satisfy  $C \geq 1e-15$ . A simpler version is to first rescale the regressor to lie in  $[0, 1]$  before taking powers. With this scaling, regressions for  $k \leq 9$  satisfy  $C \geq 1e-15$ . This is often sufficient for applications.

Ill-conditioning can often be completely eliminated by orthogonalization of the regressors. This is achieved by sequentially regressing each variable (each column in  $\mathbf{X}$ ) on the preceding variables (each preceding column), taking the residual, and then rescaling to have a unit variance. This will produce regressors which algebraically satisfy  $\mathbf{X}'\mathbf{X} = n\mathbf{I}_n$  and have a condition number of  $C = 1$ . If we apply this method to the above example, we obtain a condition number close to 1 for  $k \leq 20$ .

What this shows is that when a regression has a small condition number it is important to examine the specification carefully. It is possible that the regressors are linearly dependent in which case one or more regressors will need to be omitted. It is also possible that the regressors are badly scaled, in which case it may be useful to rescale some of the regressors. It is also possible that the variables are highly collinear, in which case a possible solution is orthogonalization. These choices should be made by the researcher, not by an automated software program.

## 3.25 Programming

Most packages allow both interactive programming (where you enter commands one-by-one) and batch programming (where you run a pre-written sequence of commands from a file). Interactive programming can be useful for exploratory analysis, but eventually all work should be executed in batch mode. This is the best way to control and document your work.

Batch programs are text files where each line executes a single command. For Stata, this file needs to have the filename extension “.do”, and for MATLAB “.m”. For R there is no specific naming requirements, though it is typical to use the extension “.r”. When writing batch files, it is useful to include comments for documentation and readability. To execute a program file, you type a command within the program.

Stata: do chapter3 executes the file *chapter3.do*

MATLAB: run chapter3 executes the file *chapter3.m*

R: source(“chapter3.r”) or source(‘chapter3.r’) executes the file *chapter3.r*

There are other similarities and differences between the commands used in these packages. For example:

1. Different symbols are used to create comments. \* in Stata, # in R, and % in Matlab.
2. Matlab uses the symbol ; to separate lines. Stata and R use a hard return.
3. Stata uses ln() to compute natural logarithms. R and Matlab use log().
4. The symbol = is used to define a variable. R prefers <-. Double equality == is used to test equality.

We now illustrate programming files for Stata, R, and MATLAB, which execute a portion of the empirical illustrations from Sections 3.7 and 3.21. For the R and Matlab code we illustrate using explicit matrix operations. Alternatively, R and Matlab have packages which implement least squares regression without the need for explicit matrix operations. In R, the standard package is lm. In Matlab the standard command is regress. The advantage of using explicit matrix operations as shown below is that you know exactly what computations are done, and it is easier to go “out of the box” to execute new procedures. The advantage of using built-in packages and commands is that coding is simplified.

**Stata do File**

```
*      Clear memory and load the data
clear
use cps09mar.dta
*      Generate transformations
gen wage = ln(earnings/(hours*week))
gen experience = age - education - 6
gen exp2 = (experience^2)/100
*      Create indicator for subsamples
gen mbf = (race == 2) & (marital <= 2) & (female == 1)
gen mbf12 = (mbf == 1) & (experience == 12)
gen sam = (race == 4) & (marital == 7) & (female == 0)
*      Regressions
reg wage education if mbf12 == 1
reg wage education experience exp2 if sam == 1
*      Leverage and influence
predict leverage, hat
predict e, residual
gen d=e*leverage/(1-leverage)
summarize d if sam ==1
```

**R Program File**

```
#      Load the data and create subsamples
dat<- read.table("cps09mar.txt")
experience <- dat[,1]-dat[,4]-6
mbf <- (dat[,11]==2)&(dat[,12]<=2)&(dat[,2]==1)&(experience==12)
sam <- (dat[,11]==4)&(dat[,12]==7)&(dat[,2]==0)
dat1 <- dat[mbf,]
dat2 <- dat[sam,]
#      First regression
y <- as.matrix(log(dat1[,5]/(dat1[,6]*dat1[,7])))
x <- cbind(dat1[,4],matrix(1,nrow(dat1),1))
xx <- t(x)%*%x
xy <- t(x)%*%y
beta <- solve(xx,xy)
print(beta)
#      Second regression
y <- as.matrix(log(dat2[,5]/(dat2[,6]*dat2[,7])))
experience <- dat2[,1]-dat2[,4]-6
exp2 <- (experience^2)/100
x <- cbind(dat2[,4],experience,exp2,matrix(1,nrow(dat2),1))
xx <- t(x)%*%x
xy <- t(x)%*%y
beta <- solve(xx,xy)
print(beta)
#      Create leverage and influence
e <- y-x%*%beta
xxi <- solve(xx)
leverage <- rowSums(x*(x%*%xxi))
r <- e/(1-leverage)
d <- leverage*e/(1-leverage)
print(max(abs(d)))
```

**MATLAB Program File**

```
% Load the data and create subsamples
dat = load cps09mar.txt;
# An alternative to load the data from an excel file is
# dat = xlsread('cps09mar.xlsx');
experience = dat(:,1)-dat(:,4)-6;
mbf = (dat(:,11)==2)&(dat(:,12)<=2)&(dat(:,2)==1)&(experience==12);
sam = (dat(:,11)==4)&(dat(:,12)==7)&(dat(:,2)==0);
dat1 = dat(mbfb,:);
dat2 = dat(sam,:);
% First regression
y = log(dat1(:,5)./(dat1(:,6).*dat1(:,7)));
x = [dat1(:,4),ones(length(dat1),1)];
xx = x'*x
xy = x'*y
beta = xx\xy;
display(beta);
% Second regression
y = log(dat2(:,5)./(dat2(:,6).*dat2(:,7)));
experience = dat2(:,1)-dat2(:,4)-6;
exp2 = (experience.^2)/100;
x = [dat2(:,4),experience,exp2,ones(length(dat2),1)];
xx = x'*x
xy = x'*y
beta = xx\xy;
display(beta);
% Create leverage and influence
e = y-x*beta;
xxi = inv(xx)
leverage = sum((x.*(x*xxi))')';
d = leverage.*e./(1-leverage);
influence = max(abs(d));
display(influence);
```

## Exercises

**Exercise 3.1** Let  $y$  be a random variable with  $\mu = \mathbb{E}(y)$  and  $\sigma^2 = \text{var}(y)$ . Define

$$g(y, \mu, \sigma^2) = \begin{pmatrix} y - \mu \\ (y - \mu)^2 - \sigma^2 \end{pmatrix}.$$

Let  $(\hat{\mu}, \hat{\sigma}^2)$  be the values such that  $\bar{g}_n(\hat{\mu}, \hat{\sigma}^2) = \mathbf{0}$  where  $\bar{g}_n(m, s) = n^{-1} \sum_{i=1}^n g(y_i, m, s)$ . Show that  $\hat{\mu}$  and  $\hat{\sigma}^2$  are the sample mean and variance.

**Exercise 3.2** Consider the OLS regression of the  $n \times 1$  vector  $\mathbf{y}$  on the  $n \times k$  matrix  $\mathbf{X}$ . Consider an alternative set of regressors  $\mathbf{Z} = \mathbf{X}\mathbf{C}$ , where  $\mathbf{C}$  is a  $k \times k$  non-singular matrix. Thus, each column of  $\mathbf{Z}$  is a mixture of some of the columns of  $\mathbf{X}$ . Compare the OLS estimates and residuals from the regression of  $\mathbf{y}$  on  $\mathbf{X}$  to the OLS estimates from the regression of  $\mathbf{y}$  on  $\mathbf{Z}$ .

**Exercise 3.3** Using matrix algebra, show  $\mathbf{X}'\hat{\mathbf{e}} = \mathbf{0}$ .

**Exercise 3.4** Let  $\hat{\mathbf{e}}$  be the OLS residual from a regression of  $\mathbf{y}$  on  $\mathbf{X} = [\mathbf{X}_1 \mathbf{X}_2]$ . Find  $\mathbf{X}_2'\hat{\mathbf{e}}$ .

**Exercise 3.5** Let  $\hat{\mathbf{e}}$  be the OLS residual from a regression of  $\mathbf{y}$  on  $\mathbf{X}$ . Find the OLS coefficient from a regression of  $\hat{\mathbf{e}}$  on  $\mathbf{X}$ .

**Exercise 3.6** Let  $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ . Find the OLS coefficient from a regression of  $\hat{\mathbf{y}}$  on  $\mathbf{X}$ .

**Exercise 3.7** Show that if  $\mathbf{X} = [\mathbf{X}_1 \mathbf{X}_2]$  then  $\mathbf{P}\mathbf{X}_1 = \mathbf{X}_1$  and  $\mathbf{M}\mathbf{X}_1 = \mathbf{0}$ .

**Exercise 3.8** Show that  $\mathbf{M}$  is idempotent:  $\mathbf{M}\mathbf{M} = \mathbf{M}$ .

**Exercise 3.9** Show that  $\text{tr } \mathbf{M} = n - k$ .

**Exercise 3.10** Show that if  $\mathbf{X} = [\mathbf{X}_1 \mathbf{X}_2]$  and  $\mathbf{X}_1'\mathbf{X}_2 = 0$  then  $\mathbf{P} = \mathbf{P}_1 + \mathbf{P}_2$ .

**Exercise 3.11** Show that when  $\mathbf{X}$  contains a constant,  $\frac{1}{n} \sum_{i=1}^n \hat{y}_i = \bar{y}$ .

**Exercise 3.12** A dummy variable takes on only the values 0 and 1. It is used for categorical data, such as an individual's gender. Let  $\mathbf{d}_1$  and  $\mathbf{d}_2$  be vectors of 1's and 0's, with the  $i^{th}$  element of  $\mathbf{d}_1$  equaling 1 and that of  $\mathbf{d}_2$  equaling 0 if the person is a man, and the reverse if the person is a woman. Suppose that there are  $n_1$  men and  $n_2$  women in the sample. Consider fitting the following three equations by OLS

$$\mathbf{y} = \mu + \mathbf{d}_1\alpha_1 + \mathbf{d}_2\alpha_2 + \mathbf{e} \quad (3.53)$$

$$\mathbf{y} = \mathbf{d}_1\alpha_1 + \mathbf{d}_2\alpha_2 + \mathbf{e} \quad (3.54)$$

$$\mathbf{y} = \mu + \mathbf{d}_1\phi + \mathbf{e} \quad (3.55)$$

Can all three equations (3.53), (3.54), and (3.55) be estimated by OLS? Explain if not.

- (a) Compare regressions (3.54) and (3.55). Is one more general than the other? Explain the relationship between the parameters in (3.54) and (3.55).
- (b) Compute  $\mathbf{1}_n'\mathbf{d}_1$  and  $\mathbf{1}_n'\mathbf{d}_2$ , where  $\mathbf{1}_n$  is an  $n \times 1$  vector of ones.
- (c) Letting  $\boldsymbol{\alpha} = (\alpha_1 \alpha_2)'$ , write equation (3.54) as  $\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{e}$ . Consider the assumption  $\mathbb{E}(\mathbf{x}_i e_i) = 0$ . Is there any content to this assumption in this setting?

**Exercise 3.13** Let  $\mathbf{d}_1$  and  $\mathbf{d}_2$  be defined as in the previous exercise.

(a) In the OLS regression

$$\mathbf{y} = \mathbf{d}_1 \hat{\gamma}_1 + \mathbf{d}_2 \hat{\gamma}_2 + \hat{\mathbf{u}},$$

show that  $\hat{\gamma}_1$  is the sample mean of the dependent variable among the men of the sample ( $\bar{y}_1$ ), and that  $\hat{\gamma}_2$  is the sample mean among the women ( $\bar{y}_2$ ).

(b) Let  $\mathbf{X}$  ( $n \times k$ ) be an additional matrix of regressors. Describe in words the transformations

$$\begin{aligned}\mathbf{y}^* &= \mathbf{y} - \mathbf{d}_1 \bar{y}_1 - \mathbf{d}_2 \bar{y}_2 \\ \mathbf{X}^* &= \mathbf{X} - \mathbf{d}_1 \bar{\mathbf{x}}_1' - \mathbf{d}_2 \bar{\mathbf{x}}_2'\end{aligned}$$

where  $\bar{\mathbf{x}}_1$  and  $\bar{\mathbf{x}}_2$  are the  $k \times 1$  means of the regressors for men and women, respectively.

(c) Compare  $\tilde{\boldsymbol{\beta}}$  from the OLS regression

$$\mathbf{y}^* = \mathbf{X}^* \tilde{\boldsymbol{\beta}} + \tilde{\mathbf{e}}$$

with  $\hat{\boldsymbol{\beta}}$  from the OLS regression

$$\mathbf{y} = \mathbf{d}_1 \hat{\alpha}_1 + \mathbf{d}_2 \hat{\alpha}_2 + \mathbf{X} \hat{\boldsymbol{\beta}} + \hat{\mathbf{e}}.$$

**Exercise 3.14** Let  $\hat{\boldsymbol{\beta}}_n = (\mathbf{X}'_n \mathbf{X}_n)^{-1} \mathbf{X}'_n \mathbf{y}_n$  denote the OLS estimate when  $\mathbf{y}_n$  is  $n \times 1$  and  $\mathbf{X}_n$  is  $n \times k$ . A new observation  $(y_{n+1}, \mathbf{x}_{n+1})$  becomes available. Prove that the OLS estimate computed using this additional observation is

$$\hat{\boldsymbol{\beta}}_{n+1} = \hat{\boldsymbol{\beta}}_n + \frac{1}{1 + \mathbf{x}'_{n+1} (\mathbf{X}'_n \mathbf{X}_n)^{-1} \mathbf{x}_{n+1}} (\mathbf{X}'_n \mathbf{X}_n)^{-1} \mathbf{x}_{n+1} (y_{n+1} - \mathbf{x}'_{n+1} \hat{\boldsymbol{\beta}}_n).$$

**Exercise 3.15** Prove that  $R^2$  is the square of the sample correlation between  $\mathbf{y}$  and  $\hat{\mathbf{y}}$ .

**Exercise 3.16** Consider two least-squares regressions

$$\mathbf{y} = \mathbf{X}_1 \tilde{\boldsymbol{\beta}}_1 + \tilde{\mathbf{e}}$$

and

$$\mathbf{y} = \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2 + \hat{\mathbf{e}}.$$

Let  $R_1^2$  and  $R_2^2$  be the  $R$ -squared from the two regressions. Show that  $R_2^2 \geq R_1^2$ . Is there a case (explain) when there is equality  $R_2^2 = R_1^2$ ?

**Exercise 3.17** For  $\tilde{\sigma}^2$  defined in (3.47), show that  $\tilde{\sigma}^2 \geq \hat{\sigma}^2$ . Is equality possible?

**Exercise 3.18** For which observations will  $\hat{\boldsymbol{\beta}}_{(-i)} = \hat{\boldsymbol{\beta}}$ ?

**Exercise 3.19** For the intercept-only model  $y_i = \beta + e_i$ , show that the leave-one-out prediction error is

$$\tilde{e}_i = \left( \frac{n}{n-1} \right) (y_i - \bar{y}).$$

**Exercise 3.20** Define the leave-one-out estimator of  $\sigma^2$ ,

$$\hat{\sigma}_{(-i)}^2 = \frac{1}{n-1} \sum_{j \neq i} \left( y_j - \mathbf{x}'_j \hat{\boldsymbol{\beta}}_{(-i)} \right)^2.$$

This is the estimator obtained from the sample with observation  $i$  omitted. Show that

$$\hat{\sigma}_{(-i)}^2 = \frac{n}{n-1} \hat{\sigma}^2 - \frac{\hat{e}_i^2}{(n-1)(1-h_{ii})}.$$

**Exercise 3.21** Consider the least-squares regression estimators

$$y_i = x_{1i}\hat{\beta}_1 + x_{2i}\hat{\beta}_2 + \hat{e}_i$$

and the “one regressor at a time” regression estimators

$$y_i = \tilde{\beta}_1 x_{1i} + \tilde{e}_{1i} \quad y_i = \tilde{\beta}_2 x_{2i} + \tilde{e}_{2i}$$

Under what condition does  $\tilde{\beta}_1 = \hat{\beta}_1$  and  $\tilde{\beta}_2 = \hat{\beta}_2$ ?

**Exercise 3.22** You estimate a least-squares regression

$$y_i = \mathbf{x}'_{1i}\tilde{\boldsymbol{\beta}}_1 + \tilde{u}_i$$

and then regress the residuals on another set of regressors

$$\tilde{u}_i = \mathbf{x}'_{2i}\tilde{\boldsymbol{\beta}}_2 + \tilde{e}_i$$

Does this second regression give you the same estimated coefficients as from estimation of a least-squares regression on both set of regressors?

$$y_i = \mathbf{x}'_{1i}\hat{\boldsymbol{\beta}}_1 + \mathbf{x}'_{2i}\hat{\boldsymbol{\beta}}_2 + \hat{e}_i$$

In other words, is it true that  $\tilde{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_2$ ? Explain your reasoning.

**Exercise 3.23** The data matrix is  $(\mathbf{y}, \mathbf{X})$  with  $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ , and consider the transformed regressor matrix  $\mathbf{Z} = [\mathbf{X}_1, \mathbf{X}_2 - \mathbf{X}_1]$ . Suppose you do a least-squares regression of  $\mathbf{y}$  on  $\mathbf{X}$ , and a least-squares regression of  $\mathbf{y}$  on  $\mathbf{Z}$ . Let  $\hat{\sigma}^2$  and  $\tilde{\sigma}^2$  denote the residual variance estimates from the two regressions. Give a formula relating  $\hat{\sigma}^2$  and  $\tilde{\sigma}^2$ ? (Explain your reasoning.)

**Exercise 3.24** Use the data set from Section 3.22 and the sub-sample used for equation (3.50) (see Section 3.25) for data construction)

- (a) Estimate equation (3.50) and compute the equation  $R^2$  and sum of squared errors.
- (b) Re-estimate the slope on education using the residual regression approach. Regress  $\log(\text{Wage})$  on experience and its square, regress education on experience and its square, and the residuals on the residuals. Report the estimates from this final regression, along with the equation  $R^2$  and sum of squared errors. Does the slope coefficient equal the value in (3.50)? Explain.
- (c) Are the  $R^2$  and sum-of-squared errors from parts (a) and (b) equal? Explain.

**Exercise 3.25** Estimate equation (3.50) as in part (a) of the previous question. Let  $\hat{e}_i$  be the OLS residual,  $\hat{y}_i$  the predicted value from the regression,  $x_{1i}$  be education and  $x_{2i}$  be experience. Numerically calculate the following:

- (a)  $\sum_{i=1}^n \hat{e}_i$
- (b)  $\sum_{i=1}^n x_{1i}\hat{e}_i$
- (c)  $\sum_{i=1}^n x_{2i}\hat{e}_i$
- (d)  $\sum_{i=1}^n x_{1i}^2\hat{e}_i$
- (e)  $\sum_{i=1}^n x_{2i}^2\hat{e}_i$
- (f)  $\sum_{i=1}^n \hat{y}_i\hat{e}_i$

$$(g) \sum_{i=1}^n \hat{e}_i^2$$

Are these calculations consistent with the theoretical properties of OLS? Explain.

**Exercise 3.26** Use the data set from Section 3.22.

- (a) Estimate a log wage regression for the subsample of white male Hispanics. In addition to education, experience, and its square, include a set of binary variables for regions and marital status. For regions, you create dummy variables for Northeast, South and West so that Midwest is the excluded group. For marital status, create variables for married, widowed or divorced, and separated, so that single (never married) is the excluded group.
- (b) Repeat this estimation using a different econometric package. Compare your results. Do they agree?

# Chapter 4

## Least Squares Regression

### 4.1 Introduction

In this chapter we investigate some finite-sample properties of the least-squares estimator in the linear regression model. In particular, we calculate the finite-sample mean and covariance matrix and propose standard errors for the coefficient estimates.

### 4.2 Random Sampling

Assumption 3.1 specified that the observations have identical distributions. To derive the finite-sample properties of the estimators we will need to additionally specify the dependence structure across the observations.

The simplest context is when the observations are mutually independent, in which case we say that they are **independent and identically distributed**, or **i.i.d.** It is also common to describe i.i.d. observations as a **random sample**. Traditionally, random sampling has been the default assumption in cross-section (e.g. survey) contexts. It is quite convenient as i.i.d. sampling leads to straightforward expressions for estimation variance. The assumption seems appropriate (meaning that it should be approximately valid) when samples are small and relatively dispersed. That is, if you randomly sample 1000 people from a large country such as the United States it seems reasonable to model their responses as mutually independent.

**Assumption 4.1** The observations  $\{(y_1, \mathbf{x}_1), \dots, (y_i, \mathbf{x}_i), \dots, (y_n, \mathbf{x}_n)\}$  are independent and identically distributed.

For most of this chapter, we will use Assumption 4.1 to derive properties of the OLS estimator.

Assumption 4.1 means that if you take any two individuals  $i \neq j$  in a sample, the values  $(y_i, \mathbf{x}_i)$  are independent of the values  $(y_j, \mathbf{x}_j)$  yet have the same distribution. Independence means that the decisions and choices of individual  $i$  do not affect the decisions of individual  $j$ , and conversely.

This assumption may be violated if individuals in the sample are connected in some way, for example if they are neighbors, members of the same village, classmates at a school, or even firms within a specific industry. In this case, it seems plausible that decisions may be inter-connected and thus mutually dependent rather than independent. Allowing for such interactions complicates inference and requires specialized treatment. A currently popular approach which allows for mutual dependence is known as **clustered dependence**, which assumes that that observations are grouped into “clusters” (for example, schools). We will discuss clustering in more detail in Section 4.21.

### 4.3 Sample Mean

To start with the simplest setting, we first consider the intercept-only model

$$\begin{aligned}y_i &= \mu + e_i \\ \mathbb{E}(e_i) &= 0.\end{aligned}$$

which is equivalent to the regression model with  $k = 1$  and  $x_i = 1$ . In the intercept model,  $\mu = \mathbb{E}(y_i)$  is the mean of  $y_i$ . (See Exercise 2.15.) The least-squares estimator  $\hat{\mu} = \bar{y}$  equals the sample mean as shown in equation (3.9).

We now calculate the mean and variance of the estimator  $\bar{y}$ . Since the sample mean is a linear function of the observations, its expectation is simple to calculate

$$\mathbb{E}(\bar{y}) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n y_i\right) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(y_i) = \mu.$$

This shows that the expected value of the least-squares estimator (the sample mean) equals the projection coefficient (the population mean). An estimator with the property that its expectation equals the parameter it is estimating is called **unbiased**.

**Definition 4.1** An estimator  $\hat{\theta}$  for  $\theta$  is **unbiased** if  $\mathbb{E}(\hat{\theta}) = \theta$ .

We next calculate the variance of the estimator  $\bar{y}$  under Assumption 4.1. Making the substitution  $y_i = \mu + e_i$  we find

$$\bar{y} - \mu = \frac{1}{n} \sum_{i=1}^n e_i.$$

Then

$$\begin{aligned}\text{var}(\bar{y}) &= \mathbb{E}(\bar{y} - \mu)^2 \\ &= \mathbb{E}\left(\left(\frac{1}{n} \sum_{i=1}^n e_i\right)\left(\frac{1}{n} \sum_{j=1}^n e_j\right)\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}(e_i e_j) \\ &= \frac{1}{n^2} \sum_{i=1}^n \sigma^2 \\ &= \frac{1}{n} \sigma^2.\end{aligned}$$

The second-to-last inequality is because  $\mathbb{E}(e_i e_j) = \sigma^2$  for  $i = j$  yet  $\mathbb{E}(e_i e_j) = 0$  for  $i \neq j$  due to independence.

We have shown that  $\text{var}(\bar{y}) = \frac{1}{n} \sigma^2$ . This is the familiar formula for the variance of the sample mean.

### 4.4 Linear Regression Model

We now consider the linear regression model. Throughout this chapter we maintain the following.

**Assumption 4.2 Linear Regression Model**

The observations  $(y_i, \mathbf{x}_i)$  satisfy the linear regression equation

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + e_i \quad (4.1)$$

$$\mathbb{E}(e_i | \mathbf{x}_i) = 0. \quad (4.2)$$

The variables have finite second moments

$$\mathbb{E}(y_i^2) < \infty,$$

$$\mathbb{E}\|\mathbf{x}_i\|^2 < \infty,$$

and an invertible design matrix

$$\mathbf{Q}_{\mathbf{xx}} = \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i) > 0.$$

We will consider both the general case of heteroskedastic regression, where the conditional variance

$$\mathbb{E}(e_i^2 | \mathbf{x}_i) = \sigma^2(\mathbf{x}_i) = \sigma_i^2$$

is unrestricted, and the specialized case of homoskedastic regression, where the conditional variance is constant. In the latter case we add the following assumption.

**Assumption 4.3 Homoskedastic Linear Regression Model**

In addition to Assumption 4.2,

$$\mathbb{E}(e_i^2 | \mathbf{x}_i) = \sigma^2(\mathbf{x}_i) = \sigma^2 \quad (4.3)$$

is independent of  $\mathbf{x}_i$ .

## 4.5 Mean of Least-Squares Estimator

In this section we show that the OLS estimator is unbiased in the linear regression model. This calculation can be done using either summation notation or matrix notation. We will use both.

First take summation notation. Observe that under (4.1)-(4.2)

$$\mathbb{E}(y_i | \mathbf{X}) = \mathbb{E}(y_i | \mathbf{x}_i) = \mathbf{x}'_i \boldsymbol{\beta}. \quad (4.4)$$

The first equality states that the conditional expectation of  $y_i$  given  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  only depends on  $\mathbf{x}_i$ , since the observations are independent across  $i$ . The second equality is the assumption of a linear conditional mean.

Using definition (3.12), the conditioning theorem (Theorem 2.3), the linearity of expectations, (4.4),

and properties of the matrix inverse,

$$\begin{aligned}
 \mathbb{E}(\hat{\boldsymbol{\beta}} | \mathbf{X}) &= \mathbb{E}\left(\left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i\right)^{-1} \left(\sum_{i=1}^n \mathbf{x}_i y_i\right) | \mathbf{X}\right) \\
 &= \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i\right)^{-1} \mathbb{E}\left(\left(\sum_{i=1}^n \mathbf{x}_i y_i\right) | \mathbf{X}\right) \\
 &= \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i\right)^{-1} \sum_{i=1}^n \mathbb{E}(\mathbf{x}_i y_i | \mathbf{X}) \\
 &= \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i\right)^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbb{E}(y_i | \mathbf{X}) \\
 &= \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i\right)^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \boldsymbol{\beta} \\
 &= \boldsymbol{\beta}.
 \end{aligned}$$

Now let's show the same result using matrix notation. (4.4) implies

$$\mathbb{E}(\mathbf{y} | \mathbf{X}) = \begin{pmatrix} \vdots \\ \mathbb{E}(y_i | \mathbf{X}) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \mathbf{x}'_i \boldsymbol{\beta} \\ \vdots \end{pmatrix} = \mathbf{X}\boldsymbol{\beta}. \quad (4.5)$$

Similarly

$$\mathbb{E}(\mathbf{e} | \mathbf{X}) = \begin{pmatrix} \vdots \\ \mathbb{E}(e_i | \mathbf{X}) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \mathbb{E}(e_i | \mathbf{x}_i) \\ \vdots \end{pmatrix} = \mathbf{0}.$$

Using  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y})$ , the conditioning theorem, the linearity of expectations, (4.5), and the properties of the matrix inverse,

$$\begin{aligned}
 \mathbb{E}(\hat{\boldsymbol{\beta}} | \mathbf{X}) &= \mathbb{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} | \mathbf{X}) \\
 &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(\mathbf{y} | \mathbf{X}) \\
 &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\
 &= \boldsymbol{\beta}.
 \end{aligned}$$

At the risk of belaboring the derivation, another way to calculate the same result is as follows. Insert  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$  into the formula for  $\hat{\boldsymbol{\beta}}$  to obtain

$$\begin{aligned}
 \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{e})) \\
 &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} \\
 &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}.
 \end{aligned} \quad (4.6)$$

This is a useful linear decomposition of the estimator  $\hat{\boldsymbol{\beta}}$  into the true parameter  $\boldsymbol{\beta}$  and the stochastic component  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}$ . Once again, we can calculate that

$$\begin{aligned}
 \mathbb{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} | \mathbf{X}) &= \mathbb{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} | \mathbf{X}) \\
 &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(\mathbf{e} | \mathbf{X}) \\
 &= \mathbf{0}.
 \end{aligned}$$

Regardless of the method, we have shown that  $\mathbb{E}(\hat{\boldsymbol{\beta}} | \mathbf{X}) = \boldsymbol{\beta}$ .

We have shown the following theorem.

**Theorem 4.1 Mean of Least-Squares Estimator**

In the linear regression model (Assumption 4.2) and i.i.d. sampling (Assumption 4.1)

$$\mathbb{E}(\hat{\boldsymbol{\beta}} | \mathbf{X}) = \boldsymbol{\beta}. \quad (4.7)$$

Equation (4.7) says that the estimator  $\hat{\boldsymbol{\beta}}$  is unbiased for  $\boldsymbol{\beta}$ , conditional on  $\mathbf{X}$ . This means that the conditional distribution of  $\hat{\boldsymbol{\beta}}$  is centered at  $\boldsymbol{\beta}$ . By “conditional on  $\mathbf{X}$ ” this means that the distribution is unbiased (centered at  $\boldsymbol{\beta}$ ) for any realization of the regressor matrix  $\mathbf{X}$ . In conditional models, we simply refer to this as saying “ $\hat{\boldsymbol{\beta}}$  is unbiased for  $\boldsymbol{\beta}$ ”.

## 4.6 Variance of Least Squares Estimator

In this section we calculate the conditional variance of the OLS estimator.

For any  $r \times 1$  random vector  $\mathbf{Z}$  define the  $r \times r$  covariance matrix

$$\begin{aligned} \text{var}(\mathbf{Z}) &= \mathbb{E}((\mathbf{Z} - \mathbb{E}(\mathbf{Z}))(\mathbf{Z} - \mathbb{E}(\mathbf{Z}))') \\ &= \mathbb{E}(\mathbf{Z}\mathbf{Z}') - (\mathbb{E}(\mathbf{Z}))(\mathbb{E}(\mathbf{Z}))' \end{aligned}$$

and for any pair  $(\mathbf{Z}, \mathbf{X})$  define the conditional covariance matrix

$$\text{var}(\mathbf{Z} | \mathbf{X}) = \mathbb{E}((\mathbf{Z} - \mathbb{E}(\mathbf{Z} | \mathbf{X}))(\mathbf{Z} - \mathbb{E}(\mathbf{Z} | \mathbf{X}))' | \mathbf{X}).$$

We define

$$\mathbf{V}_{\hat{\boldsymbol{\beta}}} \stackrel{\text{def}}{=} \text{var}(\hat{\boldsymbol{\beta}} | \mathbf{X})$$

as the conditional covariance matrix of the regression coefficient estimates. We now derive its form.

The conditional covariance matrix of the  $n \times 1$  regression error  $\mathbf{e}$  is the  $n \times n$  matrix

$$\text{var}(\mathbf{e} | \mathbf{X}) = \mathbb{E}(\mathbf{e}\mathbf{e}' | \mathbf{X}) \stackrel{\text{def}}{=} \mathbf{D}.$$

The  $i^{th}$  diagonal element of  $\mathbf{D}$  is

$$\mathbb{E}(e_i^2 | \mathbf{X}) = \mathbb{E}(e_i^2 | \mathbf{x}_i) = \sigma_i^2$$

while the  $i j^{th}$  off-diagonal element of  $\mathbf{D}$  is

$$\mathbb{E}(e_i e_j | \mathbf{X}) = \mathbb{E}(e_i | \mathbf{x}_i) \mathbb{E}(e_j | \mathbf{x}_j) = 0$$

where the first equality uses independence of the observations (Assumption 4.1) and the second is (4.2). Thus  $\mathbf{D}$  is a diagonal matrix with  $i^{th}$  diagonal element  $\sigma_i^2$ :

$$\mathbf{D} = \text{diag}(\sigma_1^2, \dots, \sigma_n^2) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}. \quad (4.8)$$

In the special case of the linear homoskedastic regression model (4.3), then

$$\mathbb{E}(e_i^2 | \mathbf{x}_i) = \sigma_i^2 = \sigma^2$$

and we have the simplification

$$\mathbf{D} = \mathbf{I}_n \sigma^2.$$

In general, however,  $\mathbf{D}$  need not necessarily take this simplified form.

For any  $n \times r$  matrix  $\mathbf{A} = \mathbf{A}(\mathbf{X})$ ,

$$\text{var}(\mathbf{A}'\mathbf{y} | \mathbf{X}) = \text{var}(\mathbf{A}'\mathbf{e} | \mathbf{X}) = \mathbf{A}'\mathbf{D}\mathbf{A}. \quad (4.9)$$

In particular, we can write  $\hat{\boldsymbol{\beta}} = \mathbf{A}'\mathbf{y}$  where  $\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$  and thus

$$\mathbf{V}_{\hat{\boldsymbol{\beta}}} = \text{var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) = \mathbf{A}'\mathbf{D}\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

It is useful to note that

$$\mathbf{X}'\mathbf{D}\mathbf{X} = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \sigma_i^2,$$

a weighted version of  $\mathbf{X}'\mathbf{X}$ .

In the special case of the linear homoskedastic regression model,  $\mathbf{D} = \mathbf{I}_n \sigma^2$ , so  $\mathbf{X}'\mathbf{D}\mathbf{X} = \mathbf{X}'\mathbf{X}\sigma^2$ , and the variance matrix simplifies to

$$\mathbf{V}_{\hat{\boldsymbol{\beta}}} = (\mathbf{X}'\mathbf{X})^{-1} \sigma^2.$$

### Theorem 4.2 Variance of Least-Squares Estimator

In the linear regression model (Assumption 4.2) and i.i.d. sampling (Assumption 4.1)

$$\begin{aligned} \mathbf{V}_{\hat{\boldsymbol{\beta}}} &= \text{var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) \\ &= (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'\mathbf{D}\mathbf{X}) (\mathbf{X}'\mathbf{X})^{-1} \end{aligned} \quad (4.10)$$

where  $\mathbf{D}$  is defined in (4.8).

In the homoskedastic linear regression model (Assumption 4.3) and i.i.d. sampling (Assumption 4.1)

$$\mathbf{V}_{\hat{\boldsymbol{\beta}}} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}.$$

## 4.7 Unconditional Moments

The previous sections derived the form of the conditional mean and variance of least-squares estimator, where we conditioned on the regressor matrix  $\mathbf{X}$ . What about the unconditional mean and variance?

Many authors and textbooks present unconditional results by either assuming or treating the regressor matrix  $\mathbf{X}$  as “fixed”. Statistically, this is appropriate when the values of the regressors are determined by the experiment and the only randomness is through the realizations of  $\mathbf{y}$ . Fixed regressors is not appropriate for observational data. Thus econometric results for fixed regressors are better interpreted as conditional results.

The core question is to state conditions under which the unconditional moments of the estimator are finite. For example, if it determined that  $\mathbb{E} \|\hat{\boldsymbol{\beta}}\| < \infty$ , then applying the law of iterated expectations (Theorem 2.1), we find that the unconditional mean of  $\hat{\boldsymbol{\beta}}$  is also  $\boldsymbol{\beta}$

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \mathbb{E}(\mathbb{E}(\hat{\boldsymbol{\beta}} | \mathbf{X})) = \boldsymbol{\beta}.$$

A challenge is that  $\hat{\boldsymbol{\beta}}$  may not have finite moments. Take the case of a single dummy variable regressor  $d_i$  with no intercept. Assume  $\mathbb{P}(d_i = 1) = p < 1$ . Then

$$\hat{\beta} = \frac{\sum_{i=1}^n d_i y_i}{\sum_{i=1}^n d_i}$$

is well defined if  $\sum_{i=1}^n d_i > 0$ . However,  $\mathbb{P}(\sum_{i=1}^n d_i = 0) = (1-p)^n > 0$ . This means that with positive (but small) probability,  $\hat{\beta}$  does not exist. Consequently  $\hat{\beta}$  has no finite moments! We ignore this complication in practice but it does pose a conundrum for theory. This existence problem arises whenever there are discrete regressors.

A solution can be obtained when the regressors have continuous distributions. A particularly clean statement was obtained by Kinal (1980) under the assumption of normal regressors and errors. While we introduce the normal regression model in Chapter 5 we present this result here for convenience.

**Theorem 4.3** (Kinal, 1980) In the linear regression model, if in addition  $(\mathbf{x}_i, \mathbf{e}_i)$  have a joint normal distribution then for any  $r$ ,  $\mathbb{E} \|\hat{\beta}\|^r < \infty$  if and only if  $r < n - k + 1$ .

This shows that when the errors and regressors are normally distributed that the least-squares estimator possesses all moments up to  $n - k$ , which includes all moments of practical interest. The normality assumption is not particularly critical for this result. What is key is the assumption that the regressors are continuously distributed.

## 4.8 Gauss-Markov Theorem

Now consider the class of estimators of  $\beta$  which are linear functions of the vector  $\mathbf{y}$ , and thus can be written as

$$\tilde{\beta} = \mathbf{A}' \mathbf{y}$$

where  $\mathbf{A}$  is an  $n \times k$  function of  $\mathbf{X}$ . As noted before, the least-squares estimator is the special case obtained by setting  $\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ . What is the best choice of  $\mathbf{A}$ ? The Gauss-Markov theorem<sup>1</sup>, which we now present, says that the least-squares estimator is the best choice among linear unbiased estimators when the errors are homoskedastic, in the sense that the least-squares estimator has the smallest variance among all unbiased linear estimators.

To see this, since  $\mathbb{E}(\mathbf{y} | \mathbf{X}) = \mathbf{X}\beta$ , then for any linear estimator  $\tilde{\beta} = \mathbf{A}'\mathbf{y}$  we have

$$\mathbb{E}(\tilde{\beta} | \mathbf{X}) = \mathbf{A}'\mathbb{E}(\mathbf{y} | \mathbf{X}) = \mathbf{A}'\mathbf{X}\beta,$$

so  $\tilde{\beta}$  is unbiased if (and only if)  $\mathbf{A}'\mathbf{X} = \mathbf{I}_k$ . Furthermore, we saw in (4.9) that

$$\text{var}(\tilde{\beta} | \mathbf{X}) = \text{var}(\mathbf{A}'\mathbf{y} | \mathbf{X}) = \mathbf{A}'\mathbf{D}\mathbf{A} = \mathbf{A}'\mathbf{A}\sigma^2$$

the last equality using the homoskedasticity assumption  $\mathbf{D} = \mathbf{I}_n\sigma^2$ . The “best” unbiased linear estimator is obtained by finding the matrix  $\mathbf{A}_0$  satisfying  $\mathbf{A}_0'\mathbf{X} = \mathbf{I}_k$  such that  $\mathbf{A}_0'\mathbf{A}_0$  is minimized in the positive definite sense, in that for any other matrix  $\mathbf{A}$  satisfying  $\mathbf{A}'\mathbf{X} = \mathbf{I}_k$ , then  $\mathbf{A}'\mathbf{A} - \mathbf{A}_0'\mathbf{A}_0$  is positive semi-definite.

**Theorem 4.4 Gauss-Markov.** In the homoskedastic linear regression model (Assumption 4.3) and i.i.d. sampling (Assumption 4.1), if  $\tilde{\beta}$  is a linear unbiased estimator of  $\beta$  then

$$\text{var}(\tilde{\beta} | \mathbf{X}) \geq \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}.$$

---

<sup>1</sup>Named after the mathematicians Carl Friedrich Gauss and Andrey Markov.

The Gauss-Markov theorem provides a lower bound on the variance matrix of unbiased linear estimators under the assumption of homoskedasticity. It says that no unbiased linear estimator can have a variance matrix smaller (in the positive definite sense) than  $\sigma^2 (\mathbf{X}' \mathbf{X})^{-1}$ . Since the variance of the OLS estimator is exactly equal to this bound, this means that the OLS estimator is efficient in the class of linear unbiased estimator. This gives rise to the description of OLS as BLUE, standing for “best linear unbiased estimator”. This is an efficiency justification for the least-squares estimator. The justification is limited because the class of models is restricted to homoskedastic linear regression and the class of potential estimators is restricted to linear unbiased estimators. This latter restriction is particularly unsatisfactory as the theorem leaves open the possibility that a non-linear or biased estimator could have lower mean squared error than the least-squares estimator.

We complete this section with a proof of the Gauss-Markov theorem.

Let  $\mathbf{A}$  be any  $n \times k$  function of  $\mathbf{X}$  such that  $\mathbf{A}' \mathbf{X} = \mathbf{I}_k$ . The estimator  $\mathbf{A}' \mathbf{y}$  is unbiased for  $\boldsymbol{\beta}$  and has variance  $\mathbf{A}' \mathbf{A} \sigma^2$ . Since the least-squares estimator is unbiased and has variance  $(\mathbf{X}' \mathbf{X})^{-1} \sigma^2$ , it is sufficient to show that the difference in the two variance matrices is positive semi-definite, or

$$\mathbf{A}' \mathbf{A} - (\mathbf{X}' \mathbf{X})^{-1} > 0. \quad (4.11)$$

Set  $\mathbf{C} = \mathbf{A} - \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1}$ . Note that  $\mathbf{X}' \mathbf{C} = \mathbf{0}$ . Then we calculate that

$$\begin{aligned} \mathbf{A}' \mathbf{A} - (\mathbf{X}' \mathbf{X})^{-1} &= \left( \mathbf{C} + \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \right)' \left( \mathbf{C} + \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \right) - (\mathbf{X}' \mathbf{X})^{-1} \\ &= \mathbf{C}' \mathbf{C} + \mathbf{C}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} + (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{C} \\ &\quad + (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} - (\mathbf{X}' \mathbf{X})^{-1} \\ &= \mathbf{C}' \mathbf{C} \\ &> 0. \end{aligned}$$

The final inequality states that the matrix  $\mathbf{C}' \mathbf{C}$  is positive semi-definite, which is a property of quadratic forms (see Appendix A.10). We have shown (4.11) as required.

## 4.9 Generalized Least Squares

Take the linear regression model in matrix format

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{e}. \quad (4.12)$$

Consider a generalized situation where the observation errors are possibly correlated and/or heteroskedastic. Specifically, suppose that

$$\mathbb{E}(\mathbf{e} | \mathbf{X}) = \mathbf{0} \quad (4.13)$$

$$\text{var}(\mathbf{e} | \mathbf{X}) = \boldsymbol{\Omega} \quad (4.14)$$

for some  $n \times n$  covariance matrix  $\boldsymbol{\Omega}$ , possibly a function of  $\mathbf{X}$ . This includes the i.i.d. sampling framework where  $\boldsymbol{\Omega} = \mathbf{D}$  as defined in (4.8) but allows for non-diagonal covariance matrices as well. As a covariance matrix,  $\boldsymbol{\Omega}$  is necessarily symmetric and positive semi-definite.

Under these assumptions, by similar arguments we can calculate the mean and variance of the OLS estimator:

$$\mathbb{E}(\hat{\boldsymbol{\beta}} | \mathbf{X}) = \boldsymbol{\beta} \quad (4.15)$$

$$\text{var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) = (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \boldsymbol{\Omega} \mathbf{X}) (\mathbf{X}' \mathbf{X})^{-1} \quad (4.16)$$

(see Exercise 4.5).

We have an analog of the Gauss-Markov Theorem.

**Theorem 4.5** If (4.13)-(4.14) hold and if  $\tilde{\beta}$  is a linear unbiased estimator of  $\beta$  then

$$\text{var}(\tilde{\beta} | \mathbf{X}) \geq (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1}.$$

We leave the proof for Exercise 4.6.

The theorem provides a lower bound on the variance matrix of unbiased linear estimators. The bound is different from the variance matrix of the OLS estimator as stated in (4.16) except when  $\boldsymbol{\Omega} = \mathbf{I}_n \sigma^2$ . This suggests that we may be able to improve on the OLS estimator.

This is indeed the case when  $\boldsymbol{\Omega}$  is known up to scale. That is, suppose that  $\boldsymbol{\Omega} = c^2 \boldsymbol{\Sigma}$  where  $c^2 > 0$  is real and  $\boldsymbol{\Sigma}$  is  $n \times n$  and known. Take the linear model (4.12) and pre-multiply by  $\boldsymbol{\Sigma}^{-1/2}$ . This produces the equation

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\beta + \tilde{\mathbf{e}}$$

where  $\tilde{\mathbf{y}} = \boldsymbol{\Sigma}^{-1/2}\mathbf{y}$ ,  $\tilde{\mathbf{X}} = \boldsymbol{\Sigma}^{-1/2}\mathbf{X}$ , and  $\tilde{\mathbf{e}} = \boldsymbol{\Sigma}^{-1/2}\mathbf{e}$ . Consider OLS estimation of  $\beta$  in this equation.

$$\begin{aligned}\tilde{\beta}_{\text{gls}} &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} \\ &= ((\boldsymbol{\Sigma}^{-1/2}\mathbf{X})'(\boldsymbol{\Sigma}^{-1/2}\mathbf{X}))^{-1}(\boldsymbol{\Sigma}^{-1/2}\mathbf{X})'(\boldsymbol{\Sigma}^{-1/2}\mathbf{y}) \\ &= (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}.\end{aligned}\tag{4.17}$$

This is called the **Generalized Least Squares** (GLS) estimator of  $\beta$  and was introduced by Aitken (1935).

You can calculate that

$$\mathbb{E}(\tilde{\beta}_{\text{gls}} | \mathbf{X}) = \beta\tag{4.18}$$

$$\text{var}(\tilde{\beta}_{\text{gls}} | \mathbf{X}) = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}.\tag{4.19}$$

This shows that the GLS estimator is unbiased, and has a covariance matrix which equals the lower bound from Theorem 4.5. This shows that the lower bound is sharp when  $\boldsymbol{\Sigma}$  is known. GLS is thus efficient in the class of linear unbiased estimators.

In the linear regression model with independent observations and known conditional variances, so that  $\boldsymbol{\Omega} = \boldsymbol{\Sigma} = \mathbf{D} = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ , the GLS estimator takes the form

$$\begin{aligned}\tilde{\beta}_{\text{gls}} &= (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}^{-1}\mathbf{y} \\ &= \left( \sum_{i=1}^n \sigma_i^{-2} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left( \sum_{i=1}^n \sigma_i^{-2} \mathbf{x}_i y_i \right).\end{aligned}$$

In practice, the covariance matrix  $\boldsymbol{\Omega}$  is unknown, so the GLS estimator as presented here is not feasible. However, the form of the GLS estimator motivates feasible versions, effectively by replacing  $\boldsymbol{\Omega}$  with an estimator. We do not pursue this here, as it is not common in current applied econometric practice.

## 4.10 Residuals

What are some properties of the residuals  $\hat{e}_i = y_i - \mathbf{x}_i' \hat{\beta}$  and prediction errors  $\tilde{e}_i = y_i - \mathbf{x}_i' \hat{\beta}_{(-i)}$ , at least in the context of the linear regression model?

Recall from (3.25) that we can write the residuals in vector notation as  $\hat{\mathbf{e}} = \mathbf{M}\mathbf{e}$  where  $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  is the orthogonal projection matrix. Using the properties of conditional expectation

$$\mathbb{E}(\hat{\mathbf{e}} | \mathbf{X}) = \mathbb{E}(\mathbf{M}\mathbf{e} | \mathbf{X}) = \mathbf{M}\mathbb{E}(\mathbf{e} | \mathbf{X}) = \mathbf{0}$$

and

$$\text{var}(\hat{\mathbf{e}} | \mathbf{X}) = \text{var}(\mathbf{M}\mathbf{e} | \mathbf{X}) = \mathbf{M}\text{var}(\mathbf{e} | \mathbf{X})\mathbf{M} = \mathbf{MDM}'\tag{4.20}$$

where  $\mathbf{D}$  is defined in (4.8).

We can simplify this expression under the assumption of conditional homoskedasticity

$$\mathbb{E}(e_i^2 | \mathbf{x}_i) = \sigma^2.$$

In this case (4.20) simplifies to

$$\text{var}(\hat{\mathbf{e}} | \mathbf{X}) = \mathbf{M}\sigma^2. \quad (4.21)$$

In particular, for a single observation  $i$ , we can find the (conditional) variance of  $\hat{e}_i$  by taking the  $i^{th}$  diagonal element of (4.21). Since the  $i^{th}$  diagonal element of  $\mathbf{M}$  is  $1 - h_{ii}$  as defined in (3.41) we obtain

$$\text{var}(\hat{e}_i | \mathbf{X}) = \mathbb{E}(\hat{e}_i^2 | \mathbf{X}) = (1 - h_{ii})\sigma^2. \quad (4.22)$$

As this variance is a function of  $h_{ii}$  and hence  $\mathbf{x}_i$ , the residuals  $\hat{e}_i$  are heteroskedastic even if the errors  $e_i$  are homoskedastic. Notice as well that (4.22) implies  $\hat{e}_i^2$  is a biased estimator of  $\sigma^2$ .

Similarly, recall from (3.46) that the prediction errors  $\tilde{e}_i = (1 - h_{ii})^{-1} \hat{e}_i$  can be written in vector notation as  $\tilde{\mathbf{e}} = \mathbf{M}^* \hat{\mathbf{e}}$  where  $\mathbf{M}^*$  is a diagonal matrix with  $i^{th}$  diagonal element  $(1 - h_{ii})^{-1}$ . Thus  $\tilde{\mathbf{e}} = \mathbf{M}^* \mathbf{M} \mathbf{e}$ . We can calculate that

$$\mathbb{E}(\tilde{\mathbf{e}} | \mathbf{X}) = \mathbf{M}^* \mathbf{M} \mathbb{E}(\mathbf{e} | \mathbf{X}) = \mathbf{0}$$

and

$$\text{var}(\tilde{\mathbf{e}} | \mathbf{X}) = \mathbf{M}^* \mathbf{M} \text{var}(\mathbf{e} | \mathbf{X}) \mathbf{M} \mathbf{M}^* = \mathbf{M}^* \mathbf{M} \mathbf{D} \mathbf{M} \mathbf{M}^*$$

which simplifies under homoskedasticity to

$$\begin{aligned} \text{var}(\tilde{\mathbf{e}} | \mathbf{X}) &= \mathbf{M}^* \mathbf{M} \mathbf{M} \mathbf{M}^* \sigma^2 \\ &= \mathbf{M}^* \mathbf{M} \mathbf{M}^* \sigma^2. \end{aligned}$$

The variance of the  $i^{th}$  prediction error is then

$$\begin{aligned} \text{var}(\tilde{e}_i | \mathbf{X}) &= \mathbb{E}(\tilde{e}_i^2 | \mathbf{X}) \\ &= (1 - h_{ii})^{-1} (1 - h_{ii})(1 - h_{ii})^{-1} \sigma^2 \\ &= (1 - h_{ii})^{-1} \sigma^2. \end{aligned}$$

A residual with constant conditional variance can be obtained by rescaling. The **standardized residuals** are

$$\bar{e}_i = (1 - h_{ii})^{-1/2} \hat{e}_i, \quad (4.23)$$

and in vector notation

$$\bar{\mathbf{e}} = (\bar{e}_1, \dots, \bar{e}_n)' = \mathbf{M}^{*1/2} \mathbf{M} \mathbf{e}. \quad (4.24)$$

From our above calculations, under homoskedasticity,

$$\text{var}(\bar{\mathbf{e}} | \mathbf{X}) = \mathbf{M}^{*1/2} \mathbf{M} \mathbf{M}^{*1/2} \sigma^2$$

and

$$\text{var}(\bar{e}_i | \mathbf{X}) = \mathbb{E}(\bar{e}_i^2 | \mathbf{X}) = \sigma^2$$

and thus these standardized residuals have the same bias and variance as the original errors when the latter are homoskedastic.

## 4.11 Estimation of Error Variance

The error variance  $\sigma^2 = \mathbb{E}(e_i^2)$  can be a parameter of interest even in a heteroskedastic regression or a projection model.  $\sigma^2$  measures the variation in the “unexplained” part of the regression. Its method of moments estimator (MME) is the sample average of the squared residuals:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{e}_i^2.$$

In the linear regression model we can calculate the mean of  $\hat{\sigma}^2$ . From (3.29) and the properties of the trace operator, observe that

$$\hat{\sigma}^2 = \frac{1}{n} \mathbf{e}' \mathbf{M} \mathbf{e} = \frac{1}{n} \text{tr}(\mathbf{e}' \mathbf{M} \mathbf{e}) = \frac{1}{n} \text{tr}(\mathbf{M} \mathbf{e} \mathbf{e}').$$

Then

$$\begin{aligned} \mathbb{E}(\hat{\sigma}^2 | \mathbf{X}) &= \frac{1}{n} \text{tr}(\mathbb{E}(\mathbf{M} \mathbf{e} \mathbf{e}' | \mathbf{X})) \\ &= \frac{1}{n} \text{tr}(\mathbf{M} \mathbb{E}(\mathbf{e} \mathbf{e}' | \mathbf{X})) \\ &= \frac{1}{n} \text{tr}(\mathbf{M} \mathbf{D}) \\ &= \frac{1}{n} \sum_{i=1}^n (1 - h_{ii}) \sigma_i^2. \end{aligned} \tag{4.25}$$

The final equality holds since the trace is the sum of the diagonal elements of  $\mathbf{M} \mathbf{D}$ , and since  $\mathbf{D}$  is diagonal the diagonal elements of  $\mathbf{M} \mathbf{D}$  are the product of the diagonal elements of  $\mathbf{M}$  and  $\mathbf{D}$ , are which are  $1 - h_{ii}$  and  $\sigma_i^2$ , respectively.

Adding the assumption of conditional homoskedasticity  $\mathbb{E}(e_i^2 | \mathbf{x}_i) = \sigma^2$ , so that  $\mathbf{D} = \mathbf{I}_n \sigma^2$ , then (4.25) simplifies to

$$\begin{aligned} \mathbb{E}(\hat{\sigma}^2 | \mathbf{X}) &= \frac{1}{n} \text{tr}(\mathbf{M} \sigma^2) \\ &= \sigma^2 \left( \frac{n-k}{n} \right) \end{aligned}$$

the final equality by (3.23). This calculation shows that  $\hat{\sigma}^2$  is biased towards zero. The order of the bias depends on  $k/n$ , the ratio of the number of estimated coefficients to the sample size.

Another way to see this is to use (4.22). Note that

$$\begin{aligned} \mathbb{E}(\hat{\sigma}^2 | \mathbf{X}) &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}(\hat{e}_i^2 | \mathbf{X}) \\ &= \frac{1}{n} \sum_{i=1}^n (1 - h_{ii}) \sigma^2 \\ &= \left( \frac{n-k}{n} \right) \sigma^2 \end{aligned}$$

the last equality using Theorem 3.6.

Since the bias takes a scale form, a classic method to obtain an unbiased estimator is by rescaling the estimator. Define

$$s^2 = \frac{1}{n-k} \sum_{i=1}^n \hat{e}_i^2. \tag{4.26}$$

By the above calculation,

$$\mathbb{E}(s^2 | \mathbf{X}) = \sigma^2$$

and

$$\mathbb{E}(s^2) = \sigma^2.$$

Hence the estimator  $s^2$  is unbiased for  $\sigma^2$ . Consequently,  $s^2$  is known as the “bias-corrected estimator” for  $\sigma^2$  and in empirical practice  $s^2$  is the most widely used estimator for  $\sigma^2$ .

Interestingly, this is not the only method to construct an unbiased estimator for  $\sigma^2$ . An estimator constructed with the standardized residuals  $\bar{e}_i$  from (4.23) is

$$\bar{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \bar{e}_i^2 = \frac{1}{n} \sum_{i=1}^n (1 - h_{ii})^{-1} \hat{e}_i^2.$$

You can show (see Exercise 4.9) that

$$\mathbb{E}(\bar{\sigma}^2 | \mathbf{X}) = \sigma^2 \quad (4.27)$$

and thus  $\bar{\sigma}^2$  is unbiased for  $\sigma^2$  (in the homoskedastic linear regression model).

When  $k/n$  is small (typically, this occurs when  $n$  is large), the estimators  $\hat{\sigma}^2$ ,  $s^2$  and  $\bar{\sigma}^2$  are likely to be similar to one another. However, if  $k/n$  is large then  $s^2$  and  $\bar{\sigma}^2$  are generally preferred to  $\hat{\sigma}^2$ . Consequently it is best to use one of the bias-corrected variance estimators in applications.

## 4.12 Mean-Square Forecast Error

One use of an estimated regression is to predict out-of-sample values. Consider an out-of-sample observation  $(y_{n+1}, \mathbf{x}_{n+1})$  where  $\mathbf{x}_{n+1}$  is observed but not  $y_{n+1}$ . Given the coefficient estimate  $\hat{\boldsymbol{\beta}}$  the standard point estimate of  $\mathbb{E}(y_{n+1} | \mathbf{x}_{n+1}) = \mathbf{x}'_{n+1} \boldsymbol{\beta}$  is  $\tilde{y}_{n+1} = \mathbf{x}'_{n+1} \hat{\boldsymbol{\beta}}$ . The forecast error is the difference between the actual value  $y_{n+1}$  and the point forecast  $\tilde{y}_{n+1}$ . This is the forecast error  $\tilde{e}_{n+1} = y_{n+1} - \tilde{y}_{n+1}$ . The mean-squared forecast error (MSFE) is its expected squared value

$$\text{MSFE}_n = \mathbb{E}(\tilde{e}_{n+1}^2).$$

In the linear regression model,  $\tilde{e}_{n+1} = e_{n+1} - \mathbf{x}'_{n+1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ , so

$$\begin{aligned} \text{MSFE}_n &= \mathbb{E}(e_{n+1}^2) - 2\mathbb{E}(e_{n+1} \mathbf{x}'_{n+1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})) \\ &\quad + \mathbb{E}(\mathbf{x}'_{n+1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{x}_{n+1}). \end{aligned} \quad (4.28)$$

The first term in (4.28) is  $\sigma^2$ . The second term in (4.28) is zero since  $e_{n+1} \mathbf{x}'_{n+1}$  is independent of  $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$  and both are mean zero. Using the properties of the trace operator, the third term in (4.28) is

$$\begin{aligned} &\text{tr}(\mathbb{E}(\mathbf{x}_{n+1} \mathbf{x}'_{n+1}) \mathbb{E}((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})')) \\ &= \text{tr}(\mathbb{E}(\mathbf{x}_{n+1} \mathbf{x}'_{n+1}) \mathbb{E}(\mathbb{E}((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' | \mathbf{X}))) \\ &= \text{tr}(\mathbb{E}(\mathbf{x}_{n+1} \mathbf{x}'_{n+1}) \mathbb{E}(V_{\hat{\boldsymbol{\beta}}})) \\ &= \mathbb{E} \text{tr}((\mathbf{x}_{n+1} \mathbf{x}'_{n+1}) V_{\hat{\boldsymbol{\beta}}}) \\ &= \mathbb{E}(\mathbf{x}'_{n+1} V_{\hat{\boldsymbol{\beta}}} \mathbf{x}_{n+1}) \end{aligned} \quad (4.29)$$

where we use the fact that  $\mathbf{x}_{n+1}$  is independent of  $\hat{\boldsymbol{\beta}}$ , the definition  $V_{\hat{\boldsymbol{\beta}}} = \mathbb{E}((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' | \mathbf{X})$  and the fact that  $\mathbf{x}_{n+1}$  is independent of  $V_{\hat{\boldsymbol{\beta}}}$ . Thus

$$\text{MSFE}_n = \sigma^2 + \mathbb{E}(\mathbf{x}'_{n+1} V_{\hat{\boldsymbol{\beta}}} \mathbf{x}_{n+1}).$$

Under conditional homoskedasticity, this simplifies to

$$\text{MSFE}_n = \sigma^2 \left( 1 + \mathbb{E}(\mathbf{x}'_{n+1} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_{n+1}) \right).$$

A simple estimator for the MSFE is obtained by averaging the squared prediction errors (3.47)

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \tilde{e}_i^2$$

where  $\tilde{e}_i = y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{(-i)} = \hat{e}_i(1 - h_{ii})^{-1}$ . Indeed, we can calculate that

$$\begin{aligned}\mathbb{E}(\tilde{\sigma}^2) &= \mathbb{E}(\tilde{e}_i^2) \\ &= \mathbb{E}(e_i - \mathbf{x}'_i (\hat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}))^2 \\ &= \sigma^2 + \mathbb{E}(\mathbf{x}'_i (\hat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta})' \mathbf{x}_i).\end{aligned}$$

By a similar calculation as in (4.29) we find

$$\mathbb{E}(\tilde{\sigma}^2) = \sigma^2 + \mathbb{E}(\mathbf{x}'_i V_{\hat{\boldsymbol{\beta}}_{(-i)}} \mathbf{x}_i) = \text{MSFE}_{n-1}.$$

This is the MSFE based on a sample of size  $n - 1$ , rather than size  $n$ . The difference arises because the in-sample prediction errors  $\tilde{e}_i$  for  $i \leq n$  are calculated using an effective sample size of  $n - 1$ , while the out-of sample prediction error  $\tilde{e}_{n+1}$  is calculated from a sample with the full  $n$  observations. Unless  $n$  is very small we should expect  $\text{MSFE}_{n-1}$  (the MSFE based on  $n - 1$  observations) to be close to  $\text{MSFE}_n$  (the MSFE based on  $n$  observations). Thus  $\tilde{\sigma}^2$  is a reasonable estimator for  $\text{MSFE}_n$ .

**Theorem 4.6 MSFE**

In the linear regression model (Assumption 4.2) and i.i.d. sampling (Assumption 4.1)

$$\text{MSFE}_n = \mathbb{E}(\tilde{e}_{n+1}^2) = \sigma^2 + \mathbb{E}(\mathbf{x}'_{n+1} V_{\hat{\boldsymbol{\beta}}} \mathbf{x}_{n+1})$$

where  $V_{\hat{\boldsymbol{\beta}}} = \text{var}(\hat{\boldsymbol{\beta}} | \mathbf{X})$ . Furthermore,  $\tilde{\sigma}^2$  defined in (3.47) is an unbiased estimator of  $\text{MSFE}_{n-1}$ :

$$\mathbb{E}(\tilde{\sigma}^2) = \text{MSFE}_{n-1}.$$

## 4.13 Covariance Matrix Estimation Under Homoskedasticity

For inference, we need an estimator of the covariance matrix  $V_{\hat{\boldsymbol{\beta}}}$  of the least-squares estimator. In this section we consider the homoskedastic regression model (Assumption 4.3).

Under homoskedasticity, the covariance matrix takes the relatively simple form

$$V_{\hat{\boldsymbol{\beta}}}^0 = (\mathbf{X}' \mathbf{X})^{-1} \sigma^2$$

which is known up to the unknown scale  $\sigma^2$ . In Section 4.11 we discussed three estimators of  $\sigma^2$ . The most commonly used choice is  $s^2$ , leading to the classic covariance matrix estimator

$$\hat{V}_{\hat{\boldsymbol{\beta}}}^0 = (\mathbf{X}' \mathbf{X})^{-1} s^2. \quad (4.30)$$

Since  $s^2$  is conditionally unbiased for  $\sigma^2$ , it is simple to calculate that  $\hat{V}_{\hat{\boldsymbol{\beta}}}^0$  is conditionally unbiased for  $V_{\hat{\boldsymbol{\beta}}}$  under the assumption of homoskedasticity:

$$\begin{aligned}\mathbb{E}(\hat{V}_{\hat{\boldsymbol{\beta}}}^0 | \mathbf{X}) &= (\mathbf{X}' \mathbf{X})^{-1} \mathbb{E}(s^2 | \mathbf{X}) \\ &= (\mathbf{X}' \mathbf{X})^{-1} \sigma^2 \\ &= V_{\hat{\boldsymbol{\beta}}}.\end{aligned}$$

This was the dominant covariance matrix estimator in applied econometrics for many years, and is still the default method in most regression packages. For example, Stata uses the covariance matrix estimator (4.30) by default in linear regression unless an alternative is specified.

If the estimator (4.30) is used, but the regression error is heteroskedastic, it is possible for  $\hat{V}_{\hat{\beta}}^0$  to be quite biased for the correct covariance matrix  $V_{\hat{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{D} \mathbf{X}) (\mathbf{X}' \mathbf{X})^{-1}$ . For example, suppose  $k = 1$  and  $\sigma_i^2 = x_i^2$  with  $\mathbb{E}(x_i) = 0$ . The ratio of the true variance of the least-squares estimator to the expectation of the variance estimator is

$$\frac{V_{\hat{\beta}}}{\mathbb{E}(\hat{V}_{\hat{\beta}}^0 | \mathbf{X})} = \frac{\sum_{i=1}^n x_i^4}{\sigma^2 \sum_{i=1}^n x_i^2} \approx \frac{\mathbb{E}(x_i^4)}{(\mathbb{E}(x_i^2))^2} \stackrel{\text{def}}{=} \kappa.$$

(Notice that we use the fact that  $\sigma_i^2 = x_i^2$  implies  $\sigma^2 = \mathbb{E}(\sigma_i^2) = \mathbb{E}(x_i^2)$ .) The constant  $\kappa$  is the standardized fourth moment (or kurtosis) of the regressor  $x_i$ , and can be any number greater than one. For example, if  $x_i \sim N(0, \sigma^2)$  then  $\kappa = 3$ , so the true variance  $V_{\hat{\beta}}$  is three times larger than the expected homoskedastic estimator  $\hat{V}_{\hat{\beta}}^0$ . But  $\kappa$  can be much larger. Suppose, for example, that  $x_i \sim \chi_1^2 - 1$ . In this case  $\kappa = 15$ , so that the true variance  $V_{\hat{\beta}}$  is fifteen times larger than the expected homoskedastic estimator  $\hat{V}_{\hat{\beta}}^0$ . While this is an extreme and constructed example, the point is that the classic covariance matrix estimator (4.30) may be quite biased when the homoskedasticity assumption fails.

## 4.14 Covariance Matrix Estimation Under Heteroskedasticity

In the previous section we showed that that the classic covariance matrix estimator can be highly biased if homoskedasticity fails. In this section we show how to construct covariance matrix estimators which do not require homoskedasticity.

Recall that the general form for the covariance matrix is

$$V_{\hat{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{D} \mathbf{X}) (\mathbf{X}' \mathbf{X})^{-1}.$$

with  $\mathbf{D}$  defined in (4.8). This depends on the unknown matrix  $\mathbf{D}$  which we can write as

$$\begin{aligned} \mathbf{D} &= \text{diag}(\sigma_1^2, \dots, \sigma_n^2) \\ &= \mathbb{E}(\mathbf{e}\mathbf{e}' | \mathbf{X}) \\ &= \mathbb{E}(\tilde{\mathbf{D}} | \mathbf{X}) \end{aligned}$$

where  $\tilde{\mathbf{D}} = \text{diag}(e_1^2, \dots, e_n^2)$ . Thus  $\tilde{\mathbf{D}}$  is a conditionally unbiased estimator for  $\mathbf{D}$ . If the squared errors  $e_i^2$  were observable, we could construct an unbiased estimator for  $V_{\hat{\beta}}$  as

$$\begin{aligned} \hat{V}_{\hat{\beta}}^{\text{ideal}} &= (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \tilde{\mathbf{D}} \mathbf{X}) (\mathbf{X}' \mathbf{X})^{-1} \\ &= (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' e_i^2 \right) (\mathbf{X}' \mathbf{X})^{-1}. \end{aligned}$$

Indeed,

$$\begin{aligned} \mathbb{E}(\hat{V}_{\hat{\beta}}^{\text{ideal}} | \mathbf{X}) &= (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \mathbb{E}(e_i^2 | \mathbf{X}) \right) (\mathbf{X}' \mathbf{X})^{-1} \\ &= (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \sigma_i^2 \right) (\mathbf{X}' \mathbf{X})^{-1} \\ &= (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{D} \mathbf{X}) (\mathbf{X}' \mathbf{X})^{-1} \\ &= V_{\hat{\beta}} \end{aligned}$$

verifying that  $\hat{V}_{\hat{\beta}}^{\text{ideal}}$  is unbiased for  $V_{\hat{\beta}}$ .

Since the errors  $e_i^2$  are unobserved,  $\hat{V}_{\hat{\beta}}^{\text{ideal}}$  is not a feasible estimator. However, we can replace the errors  $e_i$  with the least-squares residuals  $\hat{e}_i$ . Making this substitution we obtain the estimator

$$\hat{V}_{\hat{\beta}}^{\text{HC0}} = (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \hat{e}_i^2 \right) (\mathbf{X}' \mathbf{X})^{-1}. \quad (4.31)$$

The label “HC” refers to “heteroskedasticity-consistent”. The label “HC0” refers to this being the baseline heteroskedasticity-consistent covariance matrix estimator.

We know, however, that  $\hat{e}_i^2$  is biased towards zero (recall equation (4.22)). To estimate the variance  $\sigma^2$  the unbiased estimator  $s^2$  scales the moment estimator  $\hat{\sigma}^2$  by  $n/(n - k)$ . Making the same adjustment we obtain the estimator

$$\hat{V}_{\hat{\beta}}^{\text{HC1}} = \left( \frac{n}{n - k} \right) (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \hat{e}_i^2 \right) (\mathbf{X}' \mathbf{X})^{-1}. \quad (4.32)$$

While the scaling by  $n/(n - k)$  is *ad hoc*, HC1 is often recommended over the unscaled HC0 estimator.

Alternatively, we could use the standardized residuals  $\bar{e}_i$  or the prediction errors  $\tilde{e}_i$ , yielding the estimators

$$\begin{aligned} \hat{V}_{\hat{\beta}}^{\text{HC2}} &= (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \bar{e}_i^2 \right) (\mathbf{X}' \mathbf{X})^{-1} \\ &= (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^n (1 - h_{ii})^{-1} \mathbf{x}_i \mathbf{x}'_i \hat{e}_i^2 \right) (\mathbf{X}' \mathbf{X})^{-1} \end{aligned} \quad (4.33)$$

and

$$\begin{aligned} \hat{V}_{\hat{\beta}}^{\text{HC3}} &= (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \tilde{e}_i^2 \right) (\mathbf{X}' \mathbf{X})^{-1} \\ &= (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^n (1 - h_{ii})^{-2} \mathbf{x}_i \mathbf{x}'_i \hat{e}_i^2 \right) (\mathbf{X}' \mathbf{X})^{-1}. \end{aligned} \quad (4.34)$$

These are often called the “HC2” and “HC3” estimators, as labeled.

The four estimators HC0, HC1, HC2 and HC3 are collectively called **robust**, **heteroskedasticity-consistent**, or **heteroskedasticity-robust** covariance matrix estimators. The HC0 estimator was first developed by Eicker (1963) and introduced to econometrics by White (1980), and is sometimes called the **Eicker-White** or **White** covariance matrix estimator. The degree-of-freedom adjustment in HC1 was recommended by Hinkley (1977), and is the default robust covariance matrix estimator implemented in Stata. It is implemented by the “r” option, for example by a regression executed with the command “reg y x, r”. In applied econometric practice, this is the currently most popular covariance matrix estimator. The HC2 estimator was introduced by Horn, Horn and Duncan (1975) (and is implemented using the vce(hc2) option in Stata). The HC3 estimator was derived by MacKinnon and White (1985) from the jackknife principle (see Section 10.3), and by Andrews (1991a) based on the principle of leave-one-out cross-validation (and is implemented using the vce(hc3) option in Stata).

Since  $(1 - h_{ii})^{-2} > (1 - h_{ii})^{-1} > 1$  it is straightforward to show that

$$\hat{V}_{\hat{\beta}}^{\text{HC0}} < \hat{V}_{\hat{\beta}}^{\text{HC2}} < \hat{V}_{\hat{\beta}}^{\text{HC3}} \quad (4.35)$$

(See Exercise 4.10). The inequality  $A < B$  when applied to matrices means that the matrix  $B - A$  is positive definite.

In general, the bias of the covariance matrix estimators is quite complicated, but they greatly simplify under the assumption of homoskedasticity (4.3). For example, using (4.22),

$$\begin{aligned}\mathbb{E}(\widehat{V}_{\hat{\beta}}^{\text{HC0}} | \mathbf{X}) &= (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \mathbb{E}(\tilde{e}_i^2 | \mathbf{X}) \right) (\mathbf{X}' \mathbf{X})^{-1} \\ &= (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' (1 - h_{ii}) \sigma^2 \right) (\mathbf{X}' \mathbf{X})^{-1} \\ &= (\mathbf{X}' \mathbf{X})^{-1} \sigma^2 - (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' h_{ii} \right) (\mathbf{X}' \mathbf{X})^{-1} \sigma^2 \\ &< (\mathbf{X}' \mathbf{X})^{-1} \sigma^2 \\ &= V_{\hat{\beta}}.\end{aligned}$$

This calculation shows that  $\widehat{V}_{\hat{\beta}}^{\text{HC0}}$  is biased towards zero.

By a similar calculation (again under homoskedasticity) we can calculate that the HC2 estimator is unbiased

$$\mathbb{E}(\widehat{V}_{\hat{\beta}}^{\text{HC2}} | \mathbf{X}) = (\mathbf{X}' \mathbf{X})^{-1} \sigma^2. \quad (4.36)$$

(See Exercise 4.11.)

It might seem rather odd to compare the bias of heteroskedasticity-robust estimators under the assumption of homoskedasticity, but it does give us a baseline for comparison.

Another interesting calculation shows that in general (that is, without assuming homoskedasticity) the HC3 estimator is biased away from zero. Indeed, using the definition of the prediction errors (3.45)

$$\tilde{e}_i = y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{(-i)} = e_i - \mathbf{x}_i' (\hat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta})$$

so

$$\tilde{e}_i^2 = e_i^2 - 2\mathbf{x}_i' (\hat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}) e_i + (\mathbf{x}_i' (\hat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}))^2.$$

Note that  $e_i$  and  $\hat{\boldsymbol{\beta}}_{(-i)}$  are functions of non-overlapping observations and are thus independent. Hence  $\mathbb{E}((\hat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}) e_i | \mathbf{X}) = 0$  and

$$\begin{aligned}\mathbb{E}(\tilde{e}_i^2 | \mathbf{X}) &= \mathbb{E}(e_i^2 | \mathbf{X}) - 2\mathbf{x}_i' \mathbb{E}((\hat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}) e_i | \mathbf{X}) + \mathbb{E}((\mathbf{x}_i' (\hat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}))^2 | \mathbf{X}) \\ &= \sigma_i^2 + \mathbb{E}((\mathbf{x}_i' (\hat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}))^2 | \mathbf{X}) \\ &\geq \sigma_i^2.\end{aligned}$$

It follows that

$$\begin{aligned}\mathbb{E}(\widehat{V}_{\hat{\beta}}^{\text{HC3}} | \mathbf{X}) &= (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \mathbb{E}(\tilde{e}_i^2 | \mathbf{X}) \right) (\mathbf{X}' \mathbf{X})^{-1} \\ &\geq (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \sigma_i^2 \right) (\mathbf{X}' \mathbf{X})^{-1} \\ &= V_{\hat{\beta}}.\end{aligned}$$

This means that the HC3 estimator is conservative in the sense that it is weakly larger (in expectation) than the correct variance for any realization of  $\mathbf{X}$ .

We have introduced five covariance matrix estimators, including the homoskedastic estimator  $\widehat{V}_{\hat{\beta}}^0$  and the four HC estimators. Which should you use? The classic estimator  $\widehat{V}_{\hat{\beta}}^0$  is typically a poor choice, as it is only valid under the unlikely homoskedasticity restriction. For this reason it is not typically used in contemporary econometric research. Unfortunately, standard regression packages set their default choice as  $\widehat{V}_{\hat{\beta}}^0$ , so users must intentionally select a robust covariance matrix estimator.

Of the four robust estimators, HC1 is the most commonly used as it is the default robust covariance matrix option in Stata. However, HC2 and HC3 are preferred. HC2 is unbiased (under homoskedasticity) and HC3 is conservative for any  $\mathbf{X}$ . In most applications HC1, HC2 and HC3 will be very similar so this choice will not matter. The context where the estimators can differ substantially is when the sample has a large leverage value  $h_{ii}$  for some observation (or multiple large leverage values). You can see this by comparing the formulas (4.32), (4.33) and (4.34), and noting that the only difference is the scaling by the leverage values  $h_{ii}$ . If there is an observation with  $h_{ii}$  close to one, then  $(1 - h_{ii})^{-1}$  and  $(1 - h_{ii})^{-2}$  will be large, giving this observation much greater weight for construction of the covariance matrix estimator.

### Halbert L. White

Hal White (1950-2012) of the United States was an influential econometrician of recent years. His 1980 paper on heteroskedasticity-consistent covariance matrix estimation for many years was the most cited paper in economics. His research was central to the movement to view econometric models as approximations, and to the drive for increased mathematical rigor in the discipline. In addition to being a highly prolific and influential scholar, he also co-founded the economic consulting firm Bates White.

## 4.15 Standard Errors

A variance estimator such as  $\hat{V}_{\hat{\beta}}$  is an estimator of the variance of the distribution of  $\hat{\beta}$ . A more easily interpretable measure of spread is its square root – the standard deviation. This is so important when discussing the distribution of parameter estimators, we have a special name for estimates of their standard deviation.

**Definition 4.2** A **standard error**  $s(\hat{\beta})$  for a real-valued estimator  $\hat{\beta}$  is an estimator of the standard deviation of the distribution of  $\hat{\beta}$ .

When  $\beta$  is a vector with estimator  $\hat{\beta}$  and covariance matrix estimator  $\hat{V}_{\hat{\beta}}$ , standard errors for individual elements are the square roots of the diagonal elements of  $\hat{V}_{\hat{\beta}}$ . That is,

$$s(\hat{\beta}_j) = \sqrt{\hat{V}_{\hat{\beta}_j}} = \sqrt{[\hat{V}_{\hat{\beta}}]_{jj}}.$$

When the classical covariance matrix estimator (4.30) is used, the standard error takes the particularly simple form

$$s(\hat{\beta}_j) = s \sqrt{[(\mathbf{X}' \mathbf{X})^{-1}]_{jj}}. \quad (4.37)$$

As we discussed in the previous section, there are multiple possible covariance matrix estimators, so standard errors are not unique. It is therefore important to understand what formula and method is used by an author when studying their work. It is also important to understand that a particular standard error may be relevant under one set of model assumptions, but not under another set of assumptions.

To illustrate, we return to the log wage regression (3.13) of Section 3.7. We calculate that  $s^2 = 0.160$ . Therefore the homoskedastic covariance matrix estimate is

$$\hat{V}_{\hat{\beta}}^0 = \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix}^{-1} 0.160 = \begin{pmatrix} 0.002 & -0.031 \\ -0.031 & 0.499 \end{pmatrix}.$$

We also calculate that

$$\sum_{i=1}^n (1 - h_{ii})^{-1} \mathbf{x}_i \mathbf{x}'_i \hat{e}_i^2 = \begin{pmatrix} 763.26 & 48.513 \\ 48.513 & 3.1078 \end{pmatrix}.$$

Therefore the HC2 covariance matrix estimate is

$$\begin{aligned} \widehat{\mathbf{V}}_{\hat{\beta}}^{\text{HC2}} &= \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix}^{-1} \begin{pmatrix} 763.26 & 48.513 \\ 48.513 & 3.1078 \end{pmatrix} \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} 0.001 & -0.015 \\ -0.015 & 0.243 \end{pmatrix}. \end{aligned} \quad (4.38)$$

The standard errors are the square roots of the diagonal elements of these matrices. A conventional format to write the estimated equation with standard errors is

$$\widehat{\log(Wage)} = \begin{matrix} 0.155 \\ (0.031) \end{matrix} Education + \begin{matrix} 0.698 \\ (0.493) \end{matrix}. \quad (4.39)$$

Alternatively, standard errors could be calculated using the other formulae. We report the different standard errors in the following table.

Table 4.1: Standard Errors

	Education	Intercept
Homoskedastic (4.30)	0.045	0.707
HC0 (4.31)	0.029	0.461
HC1 (4.32)	0.030	0.486
HC2 (4.33)	0.031	0.493
HC3 (4.34)	0.033	0.527

The homoskedastic standard errors are noticeably different (larger, in this case) than the others. The robust standard errors are reasonably close to one another, though the jackknife standard errors are meaningfully larger than the others.

## 4.16 Covariance Matrix Estimation with Sparse Dummy Variables

The heteroskedasticity-robust covariance matrix estimators can be quite imprecise in some contexts. One is in the presence of **sparse dummy variables** – when a dummy variable only takes the value 1 or 0 for very few observations. In these contexts one component of the variance matrix is estimated on just those few observations and will be imprecise. This is effectively hidden from the user.

To see the problem, let  $d_{1i}$  be a dummy variable (takes on the values 1 and 0) and consider the dummy variable regression

$$y_i = \beta_1 d_{1i} + \beta_2 + e_i. \quad (4.40)$$

The number of observations for which  $d_{1i} = 1$  is  $n_1 = \sum_{i=1}^n d_{1i}$ . The number of observations for which  $d_{1i} = 0$  is  $n_2 = n - n_1$ . We say the design is **sparse** if  $n_1$  is small.

To simplify our analysis, we take the most extreme case where  $n_1 = 1$ . The ideas extend to the case of  $n_1 > 1$  but small, though with less dramatic effects.

In the regression model (4.40), we can calculate that the true covariance matrix of the least-squares estimator for the coefficients in (4.40) under the simplifying assumption of conditional homoskedasticity is

$$\mathbf{V}_{\hat{\beta}} = \sigma^2 (\mathbf{X}' \mathbf{X})^{-1} = \sigma^2 \begin{pmatrix} 1 & 1 \\ 1 & n \end{pmatrix}^{-1} = \frac{\sigma^2}{n-1} \begin{pmatrix} n & -1 \\ -1 & 1 \end{pmatrix}.$$

In particular, the variance of the estimator for the coefficient on the dummy variable is

$$V_{\hat{\beta}_1} = \sigma^2 \frac{n}{n-1}.$$

Essentially, the coefficient  $\beta_1$  is estimated from a single observation so its variance is roughly unaffected by sample size.

Now let's examine the standard HC1 covariance matrix estimator (4.32). The regression has perfect fit for the observation for which  $d_i = 1$  so the corresponding residual is  $\hat{e}_i = 0$ . It follows that  $d_i \hat{e}_i = 0$  for all  $i$  (either  $d_i = 0$  or  $\hat{e}_i = 0$ ). Hence

$$\sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \hat{e}_i^2 = \begin{pmatrix} 0 & 0 \\ 0 & \sum_{i=1}^n \hat{e}_i^2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & (n-2)s^2 \end{pmatrix}$$

where  $s^2 = (n-2)^{-1} \sum_{i=1}^n \hat{e}_i^2$  is the bias-corrected estimator of  $\sigma^2$ . Together we find that

$$\begin{aligned} \hat{V}_{\hat{\beta}}^{\text{HC1}} &= \left( \frac{n}{n-2} \right) \frac{1}{(n-1)^2} \begin{pmatrix} n & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & (n-2)s^2 \end{pmatrix} \begin{pmatrix} n & -1 \\ -1 & 1 \end{pmatrix} \\ &= s^2 \frac{n}{(n-1)^2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \end{aligned}$$

In particular, the estimator for  $V_{\hat{\beta}_1}$  is

$$\hat{V}_{\hat{\beta}_1}^{\text{HC1}} = s^2 \frac{n}{(n-1)^2}.$$

It has expectation

$$\mathbb{E}(\hat{V}_{\hat{\beta}_1}^{\text{HC1}}) = \sigma^2 \frac{n}{(n-1)^2} = \frac{V_{\hat{\beta}_1}}{n-1} << V_{\hat{\beta}_1}.$$

The variance estimator  $\hat{V}_{\hat{\beta}_1}^{\text{HC1}}$  is extremely biased for  $V_{\hat{\beta}_1}$ . It is too small by a multiple of  $n!$  The reported variance – and standard error – is misleadingly small. The variance estimate erroneously mis-states the precision of  $\hat{\beta}_1$ .

The fact that  $\hat{V}_{\hat{\beta}_1}^{\text{HC1}}$  is biased is unlikely to be noticed by the applied researcher. Nothing in the reported output will alert a researcher to the problem.

Another way to see the issue is to consider the estimator  $\hat{\theta} = \hat{\beta}_1 + \hat{\beta}_2$  for the sum of the coefficients  $\theta = \beta_1 + \beta_2$ . This estimator has true variance  $\sigma^2$ . The variance estimator, however is  $\hat{V}_{\hat{\theta}}^{\text{HC1}} = 0!$  (It equals the sum of the four elements in  $\hat{V}_{\hat{\beta}}^{\text{HC1}}$ ). Clearly, the estimator “0” is biased for the true value  $\sigma^2$ .

Another insight is to examine the leverage values. For the observation with  $d_i = 1$  we can calculate that

$$h_{ii} = \frac{1}{n-1} \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} n & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 1.$$

This is an extreme leverage value.

The general solution is to replace the biased covariance matrix estimator  $\hat{V}_{\hat{\beta}_1}^{\text{HC1}}$  with the unbiased estimator  $\hat{V}_{\hat{\beta}_1}^{\text{HC2}}$  (unbiased under homoskedasticity) or the conservative estimator  $\hat{V}_{\hat{\beta}_1}^{\text{HC3}}$ . This excludes the extreme sparse case  $n_1 = 1$  (for  $\hat{V}_{\hat{\beta}_1}^{\text{HC2}}$  and  $\hat{V}_{\hat{\beta}_1}^{\text{HC3}}$  cannot be calculated if  $h_{ii} = 1$  for any observation) but applies otherwise. When  $h_{ii} = 1$  for some observation, then  $\hat{V}_{\hat{\beta}_1}^{\text{HC2}}$  and  $\hat{V}_{\hat{\beta}_1}^{\text{HC3}}$  cannot be calculated. In this case, unbiased covariance matrix estimation appears to be impossible.

## 4.17 Computation

We illustrate methods to compute standard errors for equation (3.14) extending the code of Section 3.25.

**Stata do File (continued)**

```

*      Homoskedastic formula (4.30):
reg wage education experience exp2 if (mnwf == 1)
*      HC1 formula (4.32):
reg wage education experience exp2 if (mnwf == 1), r
*      HC2 formula (4.33):
reg wage education experience exp2 if (mnwf == 1), vce(hc2)
*      HC3 formula (4.34):
reg wage education experience exp2 if (mnwf == 1), vce(hc3)

```

**R Program File (continued)**

```

n <- nrow(y)
k <- ncol(x)
a <- n/(n-k)
sig2 <- (t(e) %*% e)/(n-k)
u1 <- x*(e%*%matrix(1,1,k))
u2 <- x*((e/sqrt(1-leverage))%*%matrix(1,1,k))
u3 <- x*((e/(1-leverage))%*%matrix(1,1,k))
v0 <- xx*sig2
xx <- solve(t(x)%*%x)
v1 <- xx %*% (t(u1)%*%u1) %*% xx
v1a <- a * xx %*% (t(u1)%*%u1) %*% xx
v2 <- xx %*% (t(u2)%*%u2) %*% xx
v3 <- xx %*% (t(u3)%*%u3) %*% xx
s0 <- sqrt(diag(v0))      # Homoskedastic formula
s1 <- sqrt(diag(v1))      # HC0
s1a <- sqrt(diag(v1a))    # HC1
s2 <- sqrt(diag(v2))      # HC2
s3 <- sqrt(diag(v3))      # HC3

```

**MATLAB Program File (continued)**

```
[n,k]=size(x);
a=n/(n-k);
sig2=(e'*e)/(n-k);
u1=x.*((e'*ones(1,k));u2=x.*((e./sqrt(1-leverage))*ones(1,k));
u3=x.*((e./sqrt(1-leverage))*ones(1,k));xx=inv(x'*x);
v0=xx*sig2;
v1=xx*(u1'*u1)*xx;
v1a=a*xx*(u1'*u1)*xx;
v2=xx*(u2'*u2)*xx;
v3=xx*(u3'*u3)*xx;
s0=sqrt(diag(v0));           # Homoskedastic formula
s1=sqrt(diag(v1));           # HC0 formula
s1a=sqrt(diag(v1a));         # HC1 formula
s2=sqrt(diag(v2));           # HC2 formula
s3=sqrt(diag(v3));           # HC3 formula
```

## 4.18 Measures of Fit

As we described in the previous chapter, a commonly reported measure of regression fit is the regression  $R^2$  defined as

$$R^2 = 1 - \frac{\sum_{i=1}^n \hat{e}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2}.$$

where  $\hat{\sigma}_y^2 = n^{-1} \sum_{i=1}^n (y_i - \bar{y})^2$ .  $R^2$  can be viewed as an estimator of the population parameter

$$\rho^2 = \frac{\text{var}(\mathbf{x}'_i \boldsymbol{\beta})}{\text{var}(y_i)} = 1 - \frac{\sigma^2}{\sigma_y^2}.$$

However,  $\hat{\sigma}^2$  and  $\hat{\sigma}_y^2$  are biased estimators. Theil (1961) proposed replacing these by the unbiased versions  $s^2$  and  $\tilde{\sigma}_y^2 = (n-1)^{-1} \sum_{i=1}^n (y_i - \bar{y})^2$  yielding what is known as **R-bar-squared** or **adjusted R-squared**:

$$\bar{R}^2 = 1 - \frac{s^2}{\tilde{\sigma}_y^2} = 1 - \frac{(n-1) \sum_{i=1}^n \hat{e}_i^2}{(n-k) \sum_{i=1}^n (y_i - \bar{y})^2}.$$

While  $\bar{R}^2$  is an improvement on  $R^2$ , a much better improvement is

$$\tilde{R}^2 = 1 - \frac{\sum_{i=1}^n \tilde{e}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\tilde{\sigma}^2}{\hat{\sigma}_y^2}$$

where  $\tilde{e}_i$  are the prediction errors (3.45) and  $\tilde{\sigma}^2$  is the MSPE from (3.47). As described in Section (4.12),  $\tilde{\sigma}^2$  is a good estimator of the out-of-sample mean-squared forecast error, so  $\tilde{R}^2$  is a good estimator of the percentage of the forecast variance which is explained by the regression forecast. In this sense,  $\tilde{R}^2$  is a good measure of fit.

One problem with  $R^2$ , which is partially corrected by  $\bar{R}^2$  and fully corrected by  $\tilde{R}^2$ , is that  $R^2$  necessarily increases when regressors are added to a regression model. This occurs because  $R^2$  is a negative function of the sum of squared residuals which cannot increase when a regressor is added. In contrast,  $\bar{R}^2$  and  $\tilde{R}^2$  are non-monotonic in the number of regressors.  $\tilde{R}^2$  can even be negative, which occurs when an estimated model predicts worse than a constant-only model.

In the statistical literature the MSPE  $\tilde{\sigma}^2$  is known as the **leave-one-out cross validation** criterion, and is popular for model comparison and selection, especially in high-dimensional (non-parametric) contexts. It is equivalent to use  $\tilde{R}^2$  or  $\tilde{\sigma}^2$  to compare and select models. Models with high  $\tilde{R}^2$  (or low  $\tilde{\sigma}^2$ ) are better models in terms of expected out of sample squared error. In contrast,  $R^2$  cannot be used for model selection, as it necessarily increases when regressors are added to a regression model.  $\bar{R}^2$  is also an inappropriate choice for model selection (it tends to select models with too many parameters), though a justification of this assertion requires a study of the theory of model selection. Unfortunately,  $\bar{R}^2$  is routinely used by some economists, possibly as a hold-over from previous generations.

In summary, it is recommended to omit  $R^2$  and  $\bar{R}^2$ . If a measure of fit is desired, report  $\tilde{R}^2$  or  $\tilde{\sigma}^2$ .

### Henri Theil

Henri Theil (1924-2000) of the Netherlands invented  $\bar{R}^2$  and two-stage least squares, both of which are routinely seen in applied econometrics. He also wrote an early influential advanced textbook on econometrics (Theil, 1971).

## 4.19 Empirical Example

We again return to our wage equation, but use a much larger sample of all individuals with at least 12 years of education. For regressors we include years of education, potential work experience, experience squared, and dummy variable indicators for the following: female, female union member, male union member, married female<sup>2</sup>, married male, formerly married female<sup>3</sup>, formerly married male, Hispanic, black, American Indian, Asian, and mixed race<sup>4</sup>. The available sample is 46,943 so the parameter estimates are quite precise and reported in Table 4.2. For standard errors we use the unbiased Horn-Horn-Duncan formula.

Table 4.2 displays the parameter estimates in a standard tabular format. Parameter estimates and standard errors are reported for all coefficients. In addition to the coefficient estimates, the table also reports the estimated error standard deviation and the sample size. These are useful summary measures of fit which aid readers.

As a general rule, it is advisable to always report standard errors along with parameter estimates. This allows readers to assess the precision of the parameter estimates, and as we will discuss in later chapters, form confidence intervals and t-tests for individual coefficients if desired.

The results in Table 4.2 confirm our earlier findings that the return to a year of education is approximately 12%, the return to experience is concave, that single women earn approximately 10% less than single men, and blacks earn about 10% less than whites. In addition, we see that Hispanics earn about 11% less than whites, American Indians 14% less, and Asians and Mixed races about 4% less. We also see there are wage premiums for men who are members of a labor union (about 10%), married (about 21%) or formerly married (about 8%), but no similar premiums are apparent for women.

## 4.20 Multicollinearity

As discussed in Section 3.24, if  $\mathbf{X}'\mathbf{X}$  is singular, then  $(\mathbf{X}'\mathbf{X})^{-1}$  and  $\hat{\boldsymbol{\beta}}$  are not defined. This situation is called **strict multicollinearity**, as the columns of  $\mathbf{X}$  are linearly dependent, i.e., there is some  $\boldsymbol{\alpha} \neq \mathbf{0}$  such that  $\mathbf{X}\boldsymbol{\alpha} = \mathbf{0}$ . Most commonly, this arises when sets of regressors are included which are identically

<sup>2</sup>Defining “married” as marital code 1, 2, or 3.

<sup>3</sup>Defining “formerly married” as marital code 4, 5, or 6.

<sup>4</sup>Race code 6 or higher.

Table 4.2: OLS Estimates of Linear Equation for Log(Wage)

	$\hat{\beta}$	$s(\hat{\beta})$
Education	0.117	0.001
Experience	0.033	0.001
Experience <sup>2</sup> /100	-0.056	0.002
Female	-0.098	0.011
Female Union Member	0.023	0.020
Male Union Member	0.095	0.020
Married Female	0.016	0.010
Married Male	0.211	0.010
Formerly Married Female	-0.006	0.012
Formerly Married Male	0.083	0.015
Hispanic	-0.108	0.008
Black	-0.096	0.008
American Indian	-0.137	0.027
Asian	-0.038	0.013
Mixed Race	-0.041	0.021
Intercept	0.909	0.021
$\hat{\sigma}$	0.565	
Sample Size	46,943	

Standard errors are heteroskedasticity-consistent (Horn-Horn-Duncan formula).

related. In Section 3.24 we discussed possible causes of strict multicollinearity, and discussed the related problem of ill-conditioning, which can cause numerical inaccuracies in severe cases.

A related common situation is **near multicollinearity**, which is often called “multicollinearity” for brevity. This is the situation when the regressors are highly correlated. An implication of near multicollinearity is that individual coefficient estimates will be imprecise. This is not necessarily a problem for econometric analysis as the imprecision will be reflected in the standard errors, but it is still important to understand how highly correlated regressors can result in a lack of precision of individual coefficient estimates.

We can see this most simply in a homoskedastic linear regression model with two regressors

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + e_i,$$

and

$$\frac{1}{n} \mathbf{X}' \mathbf{X} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

In this case

$$\text{var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) = \frac{\sigma^2}{n} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} = \frac{\sigma^2}{n(1-\rho^2)} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$

The correlation  $\rho$  indexes collinearity, since as  $\rho$  approaches 1 the matrix becomes singular. We can see the effect of collinearity on precision by observing that the variance of a coefficient estimate  $\sigma^2 [n(1-\rho^2)]^{-1}$  approaches infinity as  $\rho$  approaches 1. Thus the more “collinear” are the regressors, the worse the precision of the individual coefficient estimates.

What is happening is that when the regressors are highly dependent, it is statistically difficult to disentangle the impact of  $\beta_1$  from that of  $\beta_2$ . As a consequence, the precision of individual estimates are reduced. The imprecision, however, will be reflected by large standard errors, so there is no distortion in inference.

Some earlier textbooks overemphasized a concern about multicollinearity. A very amusing parody of these texts appeared in Chapter 23.3 of Goldberger's *A Course in Econometrics* (1991), which is reprinted below. To understand his basic point, you should notice how the estimation variance  $\sigma^2 [n(1 - \rho^2)]^{-1}$  depends equally and symmetrically on the correlation  $\rho$  and the sample size  $n$ . Goldberger was pointing out that the only statistical implication of multicollinearity is a lack of precision, and low sample sizes have the exact same implication.

**Arthur S. Goldberger**

Art Goldberger (1930-2009) was one of the most distinguished members of the Department of Economics at the University of Wisconsin. His PhD thesis developed an early macroeconomic forecasting model (known as the Klein-Goldberger model) but most of his career focused on microeconomic issues. He was the leading pioneer of what has been called the Wisconsin Tradition of empirical work – a combination of formal econometric theory with a careful critical analysis of empirical work. Goldberger wrote a series of highly regarded and influential graduate econometric textbooks, including *Econometric Theory* (1964), *Topics in Regression Analysis* (1968), and *A Course in Econometrics* (1991).

**Micronumerosity**

Arthur S. Goldberger

*A Course in Econometrics* (1991), Chapter 23.3

Econometrics texts devote many pages to the problem of multicollinearity in multiple regression, but they say little about the closely analogous problem of small sample size in estimating a univariate mean. Perhaps that imbalance is attributable to the lack of an exotic polysyllabic name for “small sample size.” If so, we can remove that impediment by introducing the term *micronumerosity*.

Suppose an econometrician set out to write a chapter about small sample size in sampling from a univariate population. Judging from what is now written about multicollinearity, the chapter might look like this:

1. *Micronumerosity*

The extreme case, “exact micronumerosity,” arises when  $n = 0$ , in which case the sample estimate of  $\mu$  is not unique. (Technically, there is a violation of the rank condition  $n > 0$ : the matrix 0 is singular.) The extreme case is easy enough to recognize. “Near micronumerosity” is more subtle, and yet very serious. It arises when the rank condition  $n > 0$  is barely satisfied. Near micronumerosity is very prevalent in empirical economics.

2. *Consequences of micronumerosity*

The consequences of micronumerosity are serious. Precision of estimation is reduced. There are two aspects of this reduction: estimates of  $\mu$  may have large errors, and not only that, but  $V_{\bar{y}}$  will be large.

Investigators will sometimes be led to accept the hypothesis  $\mu = 0$  because  $\bar{y}/\hat{\sigma}_{\bar{y}}$  is small, even though the true situation may be not that  $\mu = 0$  but simply that the sample data have not enabled us to pick  $\mu$  up.

The estimate of  $\mu$  will be very sensitive to sample data, and the addition of a few more observations can sometimes produce drastic shifts in the sample mean.

The true  $\mu$  may be sufficiently large for the null hypothesis  $\mu = 0$  to be rejected, even though  $V_{\bar{y}} = \sigma^2/n$  is large because of micronumerosity. But if the true  $\mu$  is small (although nonzero) the hypothesis  $\mu = 0$  may mistakenly be accepted.

*3. Testing for micronumerosity*

Tests for the presence of micronumerosity require the judicious use of various fingers. Some researchers prefer a single finger, others use their toes, still others let their thumbs rule.

A generally reliable guide may be obtained by counting the number of observations. Most of the time in econometric analysis, when  $n$  is close to zero, it is also far from infinity.

Several test procedures develop critical values  $n^*$ , such that micronumerosity is a problem only if  $n$  is smaller than  $n^*$ . But those procedures are questionable.

*4. Remedies for micronumerosity*

If micronumerosity proves serious in the sense that the estimate of  $\mu$  has an unsatisfactorily low degree of precision, we are in the statistical position of not being able to make bricks without straw.

## 4.21 Clustered Sampling

In Section 4.2 we briefly mentioned clustered sampling as an alternative to the assumption of random sampling. We now introduce the framework in more detail and extend the primary results of this chapter to encompass clustered dependence.

It might be easiest to understand the idea of clusters by considering a concrete example. Duflo, Dupas and Kremer (2011) investigate the impact of tracking (assigning students based on initial test score) on educational attainment in a randomized experiment. An extract of their data set is available on the textbook webpage in the file DDK2011.

In 2005, 140 primary schools in Kenya received funding to hire an extra first grade teacher to reduce class sizes. In half of the schools (selected randomly), students were assigned to classrooms based on an initial test score (“tracking”); in the remaining schools the students were randomly assigned to classrooms. For their analysis, the authors restricted attention to the 121 schools which initially had a single first-grade class.

The key regression<sup>5</sup> in the paper is

$$TestScore_{ig} = -0.071 + 0.138 Tracking_g + e_{ig} \quad (4.41)$$

where  $TestScore_{ig}$  is the standardized test score (normalized to have mean 0 and variance 1) of student  $i$  in school  $g$ , and  $Tracking_g$  is a dummy equal to 1 if school  $g$  was tracking. The OLS estimates indicate that schools which tracked the students had an overall increase in test scores by about 0.14 standard deviations, which is quite meaningful. More general versions of this regression are estimated, many of which take the form

$$TestScore_{ig} = \alpha + \gamma Tracking_g + \mathbf{x}'_{ig} \boldsymbol{\beta} + e_{ig} \quad (4.42)$$

where  $\mathbf{x}_{ig}$  is a set of controls specific to the student (including age, sex and initial test score).

A difficulty with applying the classical regression framework is that student achievement is likely to be correlated within a given school. Student achievement may be affected by local demographics, individual teachers, and classmates, all of which imply dependence. These concerns, however, do not

---

<sup>5</sup>Table 2, column (1). Duflo, Dupas and Kremer (2011) report a coefficient estimate of 0.139, perhaps due to a slightly different calculation to standardize the test score.

suggest that achievement will be correlated across schools, so it seems reasonable to model achievement across schools as mutually independent.

In clustering contexts it is convenient to double index the observations as  $(y_{ig}, \mathbf{x}_{ig})$  where  $g = 1, \dots, G$  indexes the cluster and  $i = 1, \dots, n_g$  indexes the individual within the  $g^{th}$  cluster. The number of observations per cluster  $n_g$  may vary across clusters. The number of clusters is  $G$ . The total number of observations is  $n = \sum_{g=1}^G n_g$ . In the Kenyan schooling example, the number of clusters (schools) in the estimation sample is  $G = 121$ , the number of students per school varies from 19 to 62, and the total number of observations is  $n = 5795$ .

While it is typical to write the observations using the double index notation  $(y_{ig}, \mathbf{x}_{ig})$ , it is also useful to use cluster-level notation. Let  $\mathbf{y}_g = (y_{1g}, \dots, y_{n_g g})'$  and  $\mathbf{X}_g = (\mathbf{x}_{1g}, \dots, \mathbf{x}_{n_g g})'$  denote the  $n_g \times 1$  vector of dependent variables and  $n_g \times k$  matrix of regressors for the  $g^{th}$  cluster. A linear regression model can be written for the individual observations as

$$y_{ig} = \mathbf{x}'_{ig} \boldsymbol{\beta} + e_{ig}$$

and using cluster notation as

$$\mathbf{y}_g = \mathbf{X}_g \boldsymbol{\beta} + \mathbf{e}_g \quad (4.43)$$

where  $\mathbf{e}_g = (e_{1g}, \dots, e_{n_g g})'$  is a  $n_g \times 1$  error vector. We can also stack the observations into full sample matrices and write the model as

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{e}.$$

Using this notation we can write the sums over the observations using the double sum  $\sum_{g=1}^G \sum_{i=1}^{n_g}$ . This is the sum across clusters of the sum across observations within each cluster. The OLS estimator can be written as

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left( \sum_{g=1}^G \sum_{i=1}^{n_g} \mathbf{x}_{ig} \mathbf{x}'_{ig} \right)^{-1} \left( \sum_{g=1}^G \sum_{i=1}^{n_g} \mathbf{x}_{ig} y_{ig} \right) \\ &= \left( \sum_{g=1}^G \mathbf{X}'_g \mathbf{X}_g \right)^{-1} \left( \sum_{g=1}^G \mathbf{X}'_g \mathbf{y}_g \right) \\ &= (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{y}). \end{aligned} \quad (4.44)$$

The OLS residuals are  $\hat{e}_{ig} = y_{ig} - \mathbf{x}'_{ig} \hat{\boldsymbol{\beta}}$  in individual level notation and  $\hat{\mathbf{e}}_g = \mathbf{y}_g - \mathbf{X}_g \hat{\boldsymbol{\beta}}$  in cluster level notation.

The standard clustering assumption is that the clusters are known to the researcher and that the observations are independent across clusters.

**Assumption 4.4** The clusters  $(\mathbf{y}_g, \mathbf{X}_g)$  are mutually independent across clusters  $g$ .

In our example, clusters are schools. In other common applications, cluster dependence has been assumed within individual classrooms, families, villages, regions, and within larger units such as industries and states. This choice is up to the researcher, though the justification will depend on the context, the nature of the data, and will reflect information and assumptions on the dependence structure across observations.

The model is a linear regression under the assumption

$$\mathbb{E}(\mathbf{e}_g | \mathbf{X}_g) = 0. \quad (4.45)$$

This is the same as assuming that the individual errors are conditionally mean zero

$$\mathbb{E}(e_{ig} | \mathbf{X}_g) = 0$$

or that the conditional mean of  $\mathbf{y}_g$  given  $\mathbf{X}_g$  is linear. As in the independent case, equation (4.45) means that the linear regression model is correctly specified. In the clustered regression model, this requires that all interaction effects within clusters have been accounted for in the specification of the individual regressors  $\mathbf{x}_{ig}$ .

In the regression (4.41), the conditional mean is necessarily linear and satisfies (4.45) since the regressor  $Tracking_g$  is a dummy variable at the cluster level. In the regression (4.42) with individual controls, (4.45) requires that the achievement of any student is unaffected by the individual controls (e.g. age, sex and initial test score) of other students within the same school.

Given (4.45), we can calculate the mean of the OLS estimator. Substituting (4.43) into (4.44) we find

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left( \sum_{g=1}^G \mathbf{X}'_g \mathbf{X}_g \right)^{-1} \left( \sum_{g=1}^G \mathbf{X}'_g \mathbf{e}_g \right).$$

The mean of  $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$  conditioning on all the regressors is

$$\begin{aligned} \mathbb{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} | \mathbf{X}) &= \left( \sum_{g=1}^G \mathbf{X}'_g \mathbf{X}_g \right)^{-1} \left( \sum_{g=1}^G \mathbf{X}'_g \mathbb{E}(\mathbf{e}_g | \mathbf{X}) \right) \\ &= \left( \sum_{g=1}^G \mathbf{X}'_g \mathbf{X}_g \right)^{-1} \left( \sum_{g=1}^G \mathbf{X}'_g \mathbb{E}(\mathbf{e}_g | \mathbf{X}_g) \right) \\ &= \mathbf{0}. \end{aligned}$$

The first equality holds by linearity, the second by Assumption 4.4 and the third by (4.45).

This shows that OLS is unbiased under clustering if the conditional mean is linear.

**Theorem 4.7** In the clustered linear regression model (Assumption 4.4 and (4.45))

$$\mathbb{E}(\hat{\boldsymbol{\beta}} | \mathbf{X}) = \boldsymbol{\beta}.$$

Now consider the covariance matrix of  $\hat{\boldsymbol{\beta}}$ . Let

$$\Sigma_g = \mathbb{E}(\mathbf{e}_g \mathbf{e}'_g | \mathbf{X}_g)$$

denote the  $n_g \times n_g$  conditional covariance matrix of the errors within the  $g^{th}$  cluster. Since the observations are independent across clusters,

$$\begin{aligned} \text{var}\left(\left(\sum_{g=1}^G \mathbf{X}'_g \mathbf{e}_g\right) | \mathbf{X}\right) &= \sum_{g=1}^G \text{var}(\mathbf{X}'_g \mathbf{e}_g | \mathbf{X}_g) \\ &= \sum_{g=1}^G \mathbf{X}'_g \mathbb{E}(\mathbf{e}_g \mathbf{e}'_g | \mathbf{X}_g) \mathbf{X}_g \\ &= \sum_{g=1}^G \mathbf{X}'_g \Sigma_g \mathbf{X}_g \\ &\stackrel{\text{def}}{=} \boldsymbol{\Omega}_n. \end{aligned} \tag{4.46}$$

It follows that

$$\begin{aligned} V_{\hat{\boldsymbol{\beta}}} &= \text{var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) \\ &= (\mathbf{X}' \mathbf{X})^{-1} \boldsymbol{\Omega}_n (\mathbf{X}' \mathbf{X})^{-1}. \end{aligned} \tag{4.47}$$

This differs from the formula in the independent case due to the correlation between observations within clusters. The magnitude of the difference depends on the degree of correlation between observations within clusters and the number of observations within clusters. To see this, suppose that all clusters have the same number of observations  $n_g = N$ ,  $\mathbb{E}(e_{ig}^2 | \mathbf{x}_g) = \sigma^2$ ,  $\mathbb{E}(e_{ig} e_{\ell g} | \mathbf{x}_g) = \sigma^2 \rho$  for  $i \neq \ell$ , and the regressors  $\mathbf{x}_{ig}$  do not vary within a cluster. In this case the exact variance of the OLS estimator equals<sup>6</sup> (after some calculations)

$$\mathbf{V}_{\hat{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} \sigma^2 (1 + \rho(N - 1)). \quad (4.48)$$

If  $\rho > 0$ , this shows that the actual variance is appropriately a multiple  $\rho N$  of the conventional formula. In the Kenyan school example, the average cluster size is 48, so if the correlation between students is  $\rho = 0.25$  the actual variance exceeds the conventional formula by a factor of about twelve. In this case the correct standard errors (the square root of the variance) should be a multiple of about three times the conventional formula. This is a substantial difference, and should not be neglected.

The solution proposed by Arellano (1987) which is now standard is to use a covariance matrix estimate which extends the robust White formula to allow for general correlation within clusters. Recall that the insight of the White covariance estimator is that the squared error  $e_i^2$  is unbiased for  $\mathbb{E}(e_i^2 | \mathbf{x}_i) = \sigma_i^2$ . Similarly with cluster dependence the matrix  $\mathbf{e}_g \mathbf{e}_g'$  is unbiased for  $\mathbb{E}(\mathbf{e}_g \mathbf{e}_g' | \mathbf{X}_g) = \Sigma_g$ . This means that an unbiased estimate for (4.46) is  $\tilde{\Omega}_n = \sum_{g=1}^G \mathbf{X}'_g \mathbf{e}_g \mathbf{e}_g' \mathbf{X}_g$ . This is not feasible, but we can replace the unknown errors by the OLS residuals to obtain Arellano's estimator

$$\begin{aligned} \hat{\Omega}_n &= \sum_{g=1}^G \mathbf{X}'_g \hat{\mathbf{e}}_g \hat{\mathbf{e}}'_g \mathbf{X}_g \\ &= \sum_{g=1}^G \sum_{i=1}^{n_g} \sum_{\ell=1}^{n_g} \mathbf{x}_{ig} \mathbf{x}'_{\ell g} \hat{e}_{ig} \hat{e}'_{\ell g} \\ &= \sum_{g=1}^G \left( \sum_{i=1}^{n_g} \mathbf{x}_{ig} \hat{e}_{ig} \right) \left( \sum_{\ell=1}^{n_g} \mathbf{x}_{\ell g} \hat{e}_{\ell g} \right)' . \end{aligned} \quad (4.49)$$

The three expressions in (4.49) give three equivalent formula which could be used to calculate  $\hat{\Omega}_n$ . The final expression writes  $\hat{\Omega}_n$  in terms of the cluster sums  $\sum_{\ell=1}^{n_g} \mathbf{x}_{\ell g} \hat{e}_{\ell g}$  which is basis for our example R and MATLAB codes shown below.

Given the expressions (4.46)-(4.47), a natural cluster covariance matrix estimator takes the form

$$\hat{\mathbf{V}}_{\hat{\beta}} = a_n (\mathbf{X}' \mathbf{X})^{-1} \hat{\Omega}_n (\mathbf{X}' \mathbf{X})^{-1} \quad (4.50)$$

where the term  $a_n$  is a possible finite-sample adjustment. The Stata cluster command uses

$$a_n = \left( \frac{n-1}{n-k} \right) \left( \frac{G}{G-1} \right).$$

The factor  $G/(G-1)$  was derived by Chris Hansen (2007) in the context of equal-sized clusters to improve performance when the number of clusters  $G$  is small. The factor  $(n-1)/(n-k)$  is an ad hoc generalization which nests the adjustment used in (4.32), since when  $G = n$  we have the simplification  $a_n = n/(n-k)$ .

Alternative cluster-robust covariance matrix estimators can be constructed using cluster-level prediction errors such as

$$\tilde{\mathbf{e}}_g = \mathbf{y}_g - \mathbf{X}_g \hat{\beta}_{(-g)}$$

where  $\hat{\beta}_{(-g)}$  is the least-squares estimator omitting cluster  $g$ . Similarly as in Section 3.20, we can show that

$$\tilde{\mathbf{e}}_g = \left( \mathbf{I}_{n_g} - \mathbf{X}_g (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'_g \right)^{-1} \hat{\mathbf{e}}_g \quad (4.51)$$

and

$$\hat{\beta}_{(-g)} = \hat{\beta} - (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'_g \tilde{\mathbf{e}}_g. \quad (4.52)$$

---

<sup>6</sup>This formula is due to Moulton (1990).

We then have the robust covariance matrix estimator

$$\hat{V}_{\hat{\beta}}^{\text{CR3}} = (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{g=1}^G \mathbf{X}'_g \tilde{\mathbf{e}}_g \tilde{\mathbf{e}}'_g \mathbf{X}_g \right) (\mathbf{X}' \mathbf{X})^{-1}. \quad (4.53)$$

The label “CR” refers to “cluster-robust” and “CR3” refers to the analogous formula for the HC3 estimator.

Similarly to the heteroskedastic-robust case, you can show that CR3 is a conservative estimator for  $V_{\hat{\beta}}$  in the sense that the conditional expectation of  $\hat{V}_{\hat{\beta}}^{\text{CR3}}$  exceeds  $V_{\hat{\beta}}$ . This covariance matrix estimator may be more cumbersome to implement, however, as the cluster-level prediction errors (4.51) cannot be calculated in a simple linear operation, and appear to require a loop (across clusters) to calculate.

To illustrate in the context of the Kenyan schooling example, we present the regression of student test scores on the school-level tracking dummy, with two standard errors displayed. The first (in parenthesis) is the conventional robust standard error. The second [in square brackets] is the clustered standard error, where clustering is at the level of the school.

$$\begin{aligned} \text{TestScore}_{ig} = & -0.071 + 0.138 \text{ Tracking}_g + e_{ig}. \\ & (0.019) \quad (0.026) \\ & [0.054] \quad [0.078] \end{aligned} \quad (4.54)$$

We can see that the cluster-robust standard errors are roughly three times the conventional robust standard errors. Consequently, confidence intervals for the coefficients are greatly affected by the choice.

For illustration, we list here the commands needed to produce the regression results with clustered standard errors in Stata, R, and MATLAB.

#### Stata do File

```
* Load data:  
use "DDK2011.dta"  
* Standard the test score variable to have mean zero and unit variance:  
egen testscore = std(totalscore)  
* Regression with standard errors clustered at the school level:  
reg testscore tracking, cluster(schoolid)
```

You can see that clustered standard errors are simple to calculate in Stata.

**R Program File**

```
# Load the data and create variables
data <- read.table("DDK2011.txt",header=TRUE,sep="\t")
y <- scale(as.matrix(data$totalscore))
n <- nrow(y)
x <- cbind(as.matrix(data$tracking),matrix(1,n,1))
schoolid <- as.matrix(data$schoolid)
k <- ncol(x)
xx <- t(x)%%x
invx <- solve(xx)
beta <- solve(xx,t(x)%%y)
xe <- x*rep(y-x%%beta,times=k)
# Clustered robust standard error
xe_sum <- rowsum(xe,schoolid)
G <- nrow(xe_sum)
omega <- t(xe_sum)%%xe_sum
scale <- G/(G-1)*(n-1)/(n-k)
V_clustered <- scale*invx%%omega%%invx
se_clustered <- sqrt(diag(V_clustered))
print(beta)
print(se_clustered)
```

Programming clustered standard errors in R is also relatively easy due to the convenient `rowsum` command, which sums variables within clusters.

**MATLAB Program File**

```
% Load the data and create variables
data = xlsread('DDK2011.xlsx');
schoolid = data(:,2);
tracking = data(:,7);
totalscore = data(:,62);
y = (totalscore - mean(totalscore))./std(totalscore);
x = [tracking,ones(size(y,1),1)];
[n,k] = size(x);
xx = x'*x;
invx = inv(xx);
beta = xx\ (x'*y);
e = y - x*beta;
% Clustered robust standard error
[schools,~,schoolidx] = unique(schoolid);
G = size(schools,1);
cluster_sums = zeros(G,k);
for j = 1:k
    cluster_sums(:,j) = accumarray(schoolidx,x(:,j).*e);
end
omega = cluster_sums'*cluster_sums;
scale = G/(G-1)*(n-1)/(n-k);
V_clustered = scale*invx*omega*invx;
se_clustered = sqrt(diag(V_clustered));
display(beta);
display(se_clustered);
```

Here we see that programming clustered standard errors in MATLAB is less convenient than the other packages, but still can be executed with just a few lines of code. This example uses the `accumarray` command, which is similar to the `rowsum` command in R, but only can be applied to vectors (hence the loop across the regressors) and works best if the *clusterid* variable are indices (which is why the original *schoolid* variable is transformed into indices in *schoolidx*. Application of these commands requires care and attention.

## 4.22 Inference with Clustered Samples

In this section we give some cautionary remarks and general advice about cluster-robust inference in econometric practice. There has been remarkably little theoretical research about the properties of cluster-robust methods – until quite recently – so these remarks may become dated rather quickly.

In many respects cluster-robust inference should be viewed similarly to heteroskedasticity-robust inference, where a “cluster” in the cluster-robust case is interpreted similarly to an “observation” in the heteroskedasticity-robust case. In particular, the effective sample size should be viewed as the number of clusters, not the “sample size”  $n$ . This is because the cluster-robust covariance matrix estimator effectively treats each cluster as a single observation, and estimates the covariance matrix based on the variation across cluster means. Hence if there are only  $G = 50$  clusters, inference should be viewed as (at best) similar to heteroskedasticity-robust inference with  $n = 50$  observations. This is a bit unsettling, for if the number of regressors is large (say  $k = 20$ ), then the covariance matrix will be estimated quite imprecisely.

Furthermore, most cluster-robust theory (for example, the work of Chris Hansen (2007)) assumes that the clusters are homogeneous, including the assumption that the cluster sizes are all identical. This turns out to be a very important simplification. When this is violated – when, for example, cluster sizes are highly heterogeneous – the regression should be viewed as roughly equivalent to the heteroskedasticity-robust case with an extremely high degree of heteroskedasticity. Cluster sums have variances which are proportional to the cluster sizes, so if the latter is heterogeneous so will be the variances of the cluster sums. This also has a large effect on finite sample inference. When clusters are heterogeneous then cluster-robust inference is similar to heteroskedasticity-robust inference with highly heteroskedastic observations.

Put together, if the number of clusters  $G$  is small and the number of observations per cluster is highly varied, then we should interpret inferential statements with a great degree of caution. Unfortunately, small  $G$  with heterogeneous cluster sizes is commonplace. Many empirical studies on U.S. data cluster at the “state” level, meaning that there are 50 or 51 clusters (the District of Columbia is typically treated as a state). The number of observations vary considerably across states since the populations are highly unequal. Thus when you read empirical papers with individual-level data but clustered at the “state” level you should be very cautious, and recognize that this is equivalent to inference with a small number of extremely heterogeneous observations.

A further complication occurs when we are interested in treatment, as in the tracking example given in the previous section. In many cases (including Duflo, Dupas and Kremer (2011)) the interest is in the effect of a treatment applied at the cluster level (e.g., schools). In many cases (not, however, Duflo, Dupas and Kremer (2011)), the number of treated clusters is small relative to the total number of clusters; in an extreme case there is just a single treated cluster. Based on the reasoning given above, these applications should be interpreted as equivalent to heteroskedasticity-robust inference with a sparse dummy variable as discussed in Section 4.16. As discussed there, standard error estimates can be erroneously small. In the extreme of a single treated cluster (in the example, if only a single school was tracked) then the estimated coefficient on *tracking* will be very imprecisely estimated, yet will have a misleadingly small cluster standard error. In general, reported standard errors will greatly underestimate the imprecision of parameter estimates.

## 4.23 At What Level to Cluster?

A practical question which arises in the context of cluster-robust inference is “At what level should we cluster?” In some examples you could cluster at a very fine level, such as families or classrooms, or at higher levels of aggregation, such as neighborhoods, schools, towns, counties, or states. What is the correct level at which to cluster? Rules of thumb have been advocated by practitioners, but at present there is little formal analysis to provide useful guidance. What do we know?

First, suppose cluster dependence is ignored or imposed at too fine a level (e.g. clustering by households instead of villages). Then variance estimators will be biased as they will omit covariance terms. As correlation is typically positive, this suggests that standard errors will be too small, giving rise to spurious indications of significance and precision.

Second, suppose cluster dependence is imposed at too aggregate a measure (e.g. clustering by states rather than villages). This does not cause bias. But the variance estimators will contain many extra components, so the precision of the covariance matrix estimator will be poor. This means that reported standard errors will be imprecise – more random – than if clustering had been less aggregate.

These considerations show that there is a trade-off between bias and variance in the estimation of the covariance matrix by cluster-robust methods. It is not at all clear – based on current theory – what to do. I state this emphatically. We really do not know what is the “correct” level at which to do cluster-robust inference. This is a very interesting question and should certainly be explored by econometric research.

One challenge is that in empirical practice, many people have observed: “Clustering is important. Standard errors change a lot whether or not we properly cluster. Therefore we should only report clustered standard errors.” The flaw in this reasoning is that we do not know why in a specific empirical

example the standard errors change under clustering. One possibility is that clustering reduces bias and thus is more accurate. The other possibility is that clustering adds sampling noise and is thus less accurate. In reality it is likely that both factors are present.

In any event a researcher should be aware of the number of clusters used in the reported calculations and should treat the number of clusters as the effective sample size for assessing inference. If the number of clusters is, say,  $G = 20$ , this should be treated as a very small sample.

To illustrate the thought experiment, consider the empirical example of Duflo, Dupas and Kremer (2011). They reported standard errors clustered at the school level, and the application uses 111 schools. Thus  $G = 111$  which we can treat as the effective sample size. The number of observations (students) ranges from 19 to 62, which is reasonably homogeneous. This seems like a well balanced application of clustered variance estimation. However, one could imagine clustering at a different level of aggregation. In some applications we might consider clustering at a less aggregate level such as the classroom level. This is not relevant in this particular application as there was only one classroom per school. We might consider consider clustering at a more aggregate level. The data set contains information on the school district, division, and zone. However, there are only 2 districts, 7 divisions, and 9 zones. Thus if we cluster by zone,  $G = 9$  is the effective sample size which would lead to imprecise standard errors. In this particular example, clustering at the school level (as done by the authors) is indeed the prudent choice.

## Exercises

**Exercise 4.1** For some integer  $k$ , set  $\mu_k = \mathbb{E}(y^k)$ .

- (a) Construct an estimator  $\hat{\mu}_k$  for  $\mu_k$ .
- (b) Show that  $\hat{\mu}_k$  is unbiased for  $\mu_k$ .
- (c) Calculate the variance of  $\hat{\mu}_k$ , say  $\text{var}(\hat{\mu}_k)$ . What assumption is needed for  $\text{var}(\hat{\mu}_k)$  to be finite?
- (d) Propose an estimator of  $\text{var}(\hat{\mu}_k)$ .

**Exercise 4.2** Calculate  $E((\bar{y} - \mu)^3)$ , the skewness of  $\bar{y}$ . Under what condition is it zero?

**Exercise 4.3** Explain the difference between  $\bar{y}$  and  $\mu$ . Explain the difference between  $n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i$  and  $\mathbb{E}(\mathbf{x}_i \mathbf{x}'_i)$ .

**Exercise 4.4** True or False. If  $y_i = x_i \beta + e_i$ ,  $x_i \in \mathbb{R}$ ,  $\mathbb{E}(e_i | x_i) = 0$ , and  $\hat{e}_i$  is the OLS residual from the regression of  $y_i$  on  $x_i$ , then  $\sum_{i=1}^n x_i^2 \hat{e}_i = 0$ .

**Exercise 4.5** Prove (4.15) and (4.16)

**Exercise 4.6** Prove Theorem 4.5.

**Exercise 4.7** Let  $\tilde{\beta}$  be the GLS estimator (4.17) under the assumptions (4.13) and (4.14). Assume that  $\Omega = c^2 \Sigma$  with  $\Sigma$  known and  $c^2$  unknown. Define the residual vector  $\tilde{\mathbf{e}} = \mathbf{y} - \mathbf{X}\tilde{\beta}$ , and an estimator for  $c^2$

$$\tilde{c}^2 = \frac{1}{n-k} \tilde{\mathbf{e}}' \Sigma^{-1} \tilde{\mathbf{e}}.$$

- (a) Show (4.18).
- (b) Show (4.19).
- (c) Prove that  $\tilde{\mathbf{e}} = \mathbf{M}_1 \mathbf{e}$ , where  $\mathbf{M}_1 = \mathbf{I} - \mathbf{X} (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma^{-1}$ .
- (d) Prove that  $\mathbf{M}_1' \Sigma^{-1} \mathbf{M}_1 = \Sigma^{-1} - \Sigma^{-1} \mathbf{X} (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma^{-1}$ .
- (e) Find  $\mathbb{E}(\tilde{c}^2 | \mathbf{X})$ .
- (f) Is  $\tilde{c}^2$  a reasonable estimator for  $c^2$ ?

**Exercise 4.8** Let  $(y_i, \mathbf{x}_i)$  be a random sample with  $\mathbb{E}(\mathbf{y} | \mathbf{X}) = \mathbf{X}\beta$ . Consider the **Weighted Least Squares** (WLS) estimator of  $\beta$

$$\tilde{\beta}_{\text{wls}} = (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} (\mathbf{X}' \mathbf{W} \mathbf{y})$$

where  $\mathbf{W} = \text{diag}(w_1, \dots, w_n)$  and  $w_i = x_{ji}^{-2}$ , where  $x_{ji}$  is one of the  $\mathbf{x}_i$ .

- (a) In which contexts would  $\tilde{\beta}_{\text{wls}}$  be a good estimator?
- (b) Using your intuition, in which situations would you expect that  $\tilde{\beta}_{\text{wls}}$  would perform better than OLS?

**Exercise 4.9** Show (4.27) in the homoskedastic regression model.

**Exercise 4.10** Prove (4.35).

**Exercise 4.11** Show (4.36) in the homoskedastic regression model.

**Exercise 4.12** Let  $\mu = \mathbb{E}(y_i)$ ,  $\sigma^2 = \mathbb{E}((y_i - \mu)^2)$  and  $\mu_3 = \mathbb{E}((y_i - \mu)^3)$  and consider the sample mean  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ . Find  $\mathbb{E}((\bar{y} - \mu)^3)$  as a function of  $\mu$ ,  $\sigma^2$ ,  $\mu_3$  and  $n$ .

**Exercise 4.13** Take the simple regression model  $y_i = x_i \beta + e_i$ ,  $x_i \in \mathbb{R}$ ,  $\mathbb{E}(e_i | x_i) = 0$ . Define  $\sigma_i^2 = \mathbb{E}(e_i^2 | x_i)$  and  $\mu_{3i} = \mathbb{E}(e_i^3 | x_i)$  and consider the OLS coefficient  $\hat{\beta}$ . Find  $\mathbb{E}((\hat{\beta} - \beta)^3 | \mathbf{X})$ .

**Exercise 4.14** Take a regression model with i.i.d. observations  $(y_i, x_i)$  and scalar  $x_i$

$$y_i = x_i \beta + e_i$$

$$\mathbb{E}(e_i | x_i) = 0$$

The parameter of interest is  $\theta = \beta^2$ . Consider the OLS estimates  $\hat{\beta}$  and  $\hat{\theta} = \hat{\beta}^2$ .

- (a) Find  $\mathbb{E}(\hat{\theta} | \mathbf{X})$  using our knowledge of  $\mathbb{E}(\hat{\beta} | \mathbf{X})$  and  $V_{\hat{\beta}} = \text{var}(\hat{\beta} | \mathbf{X})$ . Is  $\hat{\theta}$  biased for  $\theta$ ?
- (b) Suggest an (approximate) biased-corrected estimator  $\hat{\theta}^*$  using an estimator  $\hat{V}_{\hat{\beta}}$  for  $V_{\hat{\beta}}$ .
- (c) For  $\hat{\theta}^*$  to be potentially unbiased, which estimator of  $V_{\hat{\beta}}$  is most appropriate?

Under which conditions is  $\hat{\theta}^*$  unbiased?

**Exercise 4.15** Consider an iid sample  $\{y_i, \mathbf{x}_i\} i = 1, \dots, n$  where  $\mathbf{x}_i$  is  $k \times 1$ . Assume the linear conditional expectation model

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(e_i | \mathbf{x}_i) &= 0 \end{aligned}$$

Assume that  $n^{-1} \mathbf{X}' \mathbf{X} = \mathbf{I}_k$  (orthonormal regressors). Consider the OLS estimator  $\hat{\boldsymbol{\beta}}$  for  $\boldsymbol{\beta}$ .

- (a) Find  $V_{\hat{\boldsymbol{\beta}}} = \text{var}(\hat{\boldsymbol{\beta}})$
- (b) In general, are  $\hat{\beta}_j$  and  $\hat{\beta}_\ell$  for  $j \neq \ell$  correlated or uncorrelated?
- (c) Find a sufficient condition so that  $\hat{\beta}_j$  and  $\hat{\beta}_\ell$  for  $j \neq \ell$  are uncorrelated.

**Exercise 4.16** Take the linear homoskedastic CEF

$$\begin{aligned} y_i^* &= \mathbf{x}'_i \boldsymbol{\beta} + e_i & (4.55) \\ \mathbb{E}(e_i | \mathbf{x}_i) &= 0 \\ \mathbb{E}(e_i^2 | \mathbf{x}_i) &= \sigma^2 \end{aligned}$$

and suppose that  $y_i^*$  is measured with error. Instead of  $y_i^*$ , we observe  $y_i$  which satisfies

$$y_i = y_i^* + u_i$$

where  $u_i$  is measurement error. Suppose that  $e_i$  and  $u_i$  are independent and

$$\begin{aligned} \mathbb{E}(u_i | \mathbf{x}_i) &= 0 \\ \mathbb{E}(u_i^2 | \mathbf{x}_i) &= \sigma_u^2 \end{aligned}$$

- (a) Derive an equation for  $y_i$  as a function of  $\mathbf{x}_i$ . Be explicit to write the error term as a function of the structural errors  $e_i$  and  $u_i$ . What is the effect of this measurement error on the model (4.55)?

- (b) Describe the effect of this measurement error on OLS estimation of  $\beta$  in the feasible regression of the observed  $y_i$  on  $\mathbf{x}_i$ .
- (c) Describe the effect (if any) of this measurement error on appropriate standard error calculation for  $\hat{\beta}$ .

**Exercise 4.17** Suppose that for a pair of observables  $(y_i, x_i)$  with  $x_i > 0$  that an economic model implies

$$\mathbb{E}(y_i | x_i) = (\gamma + \theta x_i)^{1/2}. \quad (4.56)$$

A friend suggests that (given an iid sample) you estimate  $\gamma$  and  $\theta$  by the linear regression of  $y_i^2$  on  $x_i$ , that is, to estimate the equation

$$y_i^2 = \alpha + \beta x_i + e_i. \quad (4.57)$$

- (a) Investigate your friend's suggestion. Define  $u_i = y_i - (\gamma + \theta x_i)^{1/2}$ . Show that  $\mathbb{E}(u_i | x_i) = 0$  is implied by (4.56).
- (b) Use  $y_i = (\gamma + \theta x_i)^{1/2} + u_i$  to calculate  $\mathbb{E}(y_i^2 | x_i)$ . What does this tell you about the implied equation (4.57)?
- (c) Can you recover either  $\gamma$  and/or  $\theta$  from estimation of (4.57)? Are additional assumptions required?
- (d) Is this a reasonable suggestion?

**Exercise 4.18** Take the model

$$\begin{aligned} y_i &= \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + e_i \\ \mathbb{E}(e_i | \mathbf{x}_i) &= 0 \\ \mathbb{E}(e_i^2 | \mathbf{x}_i) &= \sigma^2 \end{aligned}$$

where  $\mathbf{x}_i = (\mathbf{x}_{1i}, \mathbf{x}_{2i})$ , with  $\mathbf{x}_{1i}$   $k_1 \times 1$  and  $\mathbf{x}_{2i}$   $k_2 \times 1$ . Consider the short regression

$$y_i = \mathbf{x}'_{1i} \hat{\boldsymbol{\beta}}_1 + \hat{e}_i$$

and define the error variance estimator

$$s^2 = \frac{1}{n - k_1} \sum_{i=1}^n \hat{e}_i^2.$$

Find  $\mathbb{E}(s^2 | \mathbf{X})$

**Exercise 4.19** Let  $\mathbf{y}$  be  $n \times 1$ ,  $\mathbf{X}$  be  $n \times k$ , and  $\mathbf{X}^* = \mathbf{X}\mathbf{C}$  where  $\mathbf{C}$  is  $k \times k$  and full-rank. Let  $\hat{\boldsymbol{\beta}}$  be the least-squares estimator from the regression of  $\mathbf{y}$  on  $\mathbf{X}$ , and let  $\hat{V}$  be the estimate of its asymptotic covariance matrix. Let  $\hat{\boldsymbol{\beta}}^*$  and  $\hat{V}^*$  be those from the regression of  $\mathbf{y}$  on  $\mathbf{X}^*$ . Derive an expression for  $\hat{V}^*$  as a function of  $\hat{V}$ .

**Exercise 4.20** Take the model

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \\ \mathbb{E}(\mathbf{e} | \mathbf{X}) &= \mathbf{0} \\ \mathbb{E}(\mathbf{e}\mathbf{e}' | \mathbf{X}) &= \boldsymbol{\Omega} \end{aligned}$$

Assume for simplicity that  $\boldsymbol{\Omega}$  is known. Consider the OLS and GLS estimators  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y})$  and  $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y})$ . Compute the (conditional) covariance between  $\hat{\boldsymbol{\beta}}$  and  $\tilde{\boldsymbol{\beta}}$ :

$$\mathbb{E}((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})' | \mathbf{X})$$

Find the (conditional) covariance matrix for  $\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}$ :

$$\mathbb{E}((\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})' | \mathbf{X})$$

**Exercise 4.21** The model is

$$\begin{aligned}y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(e_i | \mathbf{x}_i) &= 0 \\ \mathbb{E}(e_i^2 | \mathbf{x}_i) &= \sigma_i^2 \\ \boldsymbol{\Omega} &= \text{diag}\{\sigma_1^2, \dots, \sigma_n^2\}.\end{aligned}$$

The parameter  $\beta$  is estimated both by OLS  $\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}$  and GLS  $\tilde{\boldsymbol{\beta}} = (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{y}$ . Let  $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}$  and  $\tilde{\mathbf{e}} = \mathbf{y} - \mathbf{X} \tilde{\boldsymbol{\beta}}$  denote the residuals. Let  $\hat{R}^2 = 1 - \hat{\mathbf{e}}' \hat{\mathbf{e}} / (\mathbf{y}' \mathbf{y})$  and  $\tilde{R}^2 = 1 - \tilde{\mathbf{e}}' \tilde{\mathbf{e}} / (\mathbf{y}' \mathbf{y})$  denote the equation  $R^2$  where  $\mathbf{y}^* = \mathbf{y} - \bar{y}$ . If the error  $e_i$  is truly heteroskedastic will  $\hat{R}^2$  or  $\tilde{R}^2$  be smaller?

**Exercise 4.22** An economist friend tells you that the assumption that the observations  $(y_i, \mathbf{x}_i)$  are i.i.d. implies that the regression  $y_i = \mathbf{x}'_i \boldsymbol{\beta} + e_i$  is homoskedastic. Do you agree with your friend? How would you explain your position?

**Exercise 4.23** Take the linear regression model with  $\mathbb{E}(\mathbf{y} | \mathbf{X}) = \mathbf{X}\beta$ . Define the *ridge regression* estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{X} + \mathbf{I}_k \lambda)^{-1} \mathbf{X}' \mathbf{y}$$

where  $\lambda > 0$  is a fixed constant. Find  $E(\hat{\boldsymbol{\beta}} | \mathbf{X})$ . Is  $\hat{\boldsymbol{\beta}}$  biased for  $\boldsymbol{\beta}$ ?

**Exercise 4.24** Continue the empirical analysis in Exercise 3.24.

- (a) Calculate standard errors using the homoskedasticity formula and using the four covariance matrices from Section 4.14.
- (b) Repeat in your second programming language. Are they identical?

**Exercise 4.25** Continue the empirical analysis in Exercise 3.26. Calculate standard errors using the HC3 method. Repeat in your second programming language. Are they identical?

**Exercise 4.26** Extend the empirical analysis reported in Section 4.21. Do a regression of standardized test score (*totalscore* normalized to have zero mean and variance 1) on tracking, age, sex, being assigned to the contract teacher, and student's percentile in the initial distribution. (The sample size will be smaller as some observations have missing variables.) Calculate standard errors using both the conventional robust formula, and clustering based on the school.

- (a) Compare the two sets of standard errors. Which standard error changes the most by clustering? Which changes the least?
- (b) How does the coefficient on *tracking* change by inclusion of the individual controls (in comparison to the results from (4.54))?

# Chapter 5

## Normal Regression and Maximum Likelihood

### 5.1 Introduction

This chapter introduces the normal regression model and the method of maximum likelihood. The normal regression model is a special case of the linear regression model. It is important as normality allows precise distributional characterizations and sharp inferences. It also provides a baseline for comparison with alternative inference methods, such as asymptotic approximations and the bootstrap.

The method of maximum likelihood is a powerful statistical method for parametric models (such as the normal regression model) and is widely used in econometric practice.

### 5.2 The Normal Distribution

We say that a random variable  $X$  has the **standard normal distribution**, or **Gaussian**, written  $X \sim N(0, 1)$ , if it has the density

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \quad -\infty < x < \infty.$$

The standard normal density is typically written with the symbol  $\phi(x)$  and the corresponding distribution function by  $\Phi(x)$ . It is a valid density function by the following result.

**Theorem 5.1**

$$\int_0^\infty \exp(-x^2/2) dx = \sqrt{\frac{\pi}{2}}. \quad (5.1)$$

The proof is presented in Section 5.20.

Plots of the standard normal density function  $\phi(x)$  and distribution function  $\Phi(x)$  are displayed in Figure 5.1.

All moments of the normal distribution are finite. Since the density is symmetric about zero all odd moments are zero. By integration by parts, you can show (see Exercises 5.2 and 5.3) that  $\mathbb{E}(X^2) = 1$  and  $\mathbb{E}(X^4) = 3$ . In fact, for any positive integer  $m$ ,

$$\mathbb{E}(X^{2m}) = (2m-1)!! = (2m-1) \cdot (2m-3) \cdots 1.$$

The notation  $k!! = k \cdot (k-2) \cdots 1$  is known as the **double factorial**. For example,  $\mathbb{E}(X^6) = 15$ ,  $\mathbb{E}(X^8) = 105$ , and  $\mathbb{E}(X^{10}) = 945$ .

The absolute moments are also straightforward to calculate.

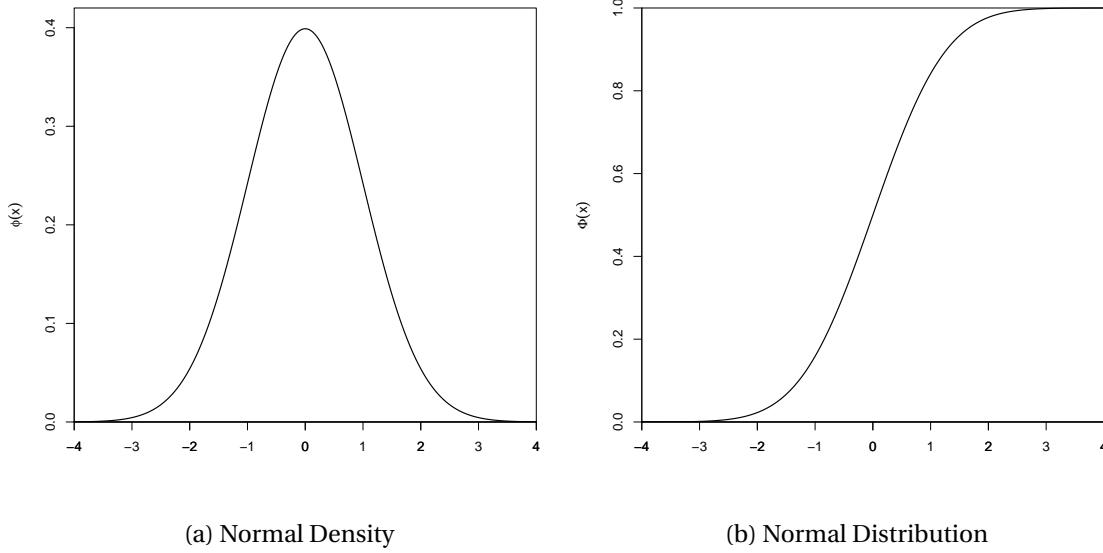


Figure 5.1: Standard Normal Density and Distribution

**Theorem 5.2** If  $X \sim N(0, 1)$  then for any  $r > 0$

$$\mathbb{E}|X|^r = \frac{2^{r/2}}{\sqrt{\pi}} \Gamma\left(\frac{r+1}{2}\right)$$

where  $\Gamma(t) = \int_0^\infty u^{t-1} e^{-u} du$  is the gamma function (Section 5.19).

The proof is presented in Section 5.20.

We say that  $X$  has a **univariate normal distribution**, written  $X \sim N(\mu, \sigma^2)$ , for  $\mu \in \mathbb{R}$  and  $\sigma^2 > 0$ , if it has the density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty.$$

The mean and variance of  $X$  are  $\mu$  and  $\sigma^2$ , respectively.

We say that the  $k$ -vector  $\mathbf{X}$  has a **multivariate standard normal distribution**, written  $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_k)$ , if it has the joint density

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{\mathbf{x}'\mathbf{x}}{2}\right), \quad \mathbf{x} \in \mathbb{R}^k.$$

The mean and covariance matrix of  $\mathbf{X}$  are  $\mathbf{0}$  and  $\mathbf{I}_k$ , respectively. Since this joint density factors, you can check that the elements of  $\mathbf{X}$  are independent standard normal random variables.

We say that the  $k$ -vector  $\mathbf{X}$  has a **multivariate normal distribution**, written  $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , for  $\boldsymbol{\mu} \in \mathbb{R}^k$  and  $\boldsymbol{\Sigma} > 0$ , if it has the joint density

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right), \quad \mathbf{x} \in \mathbb{R}^k.$$

The mean and covariance matrix of  $\mathbf{X}$  are  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$ , respectively. By setting  $k = 1$  you can check that the multivariate normal simplifies to the univariate normal.

For technical purposes it is useful to know the form of the moment generating and characteristic functions.

**Theorem 5.3** If  $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  then its moment generating function is

$$M(\mathbf{t}) = \mathbb{E}(\exp(\mathbf{t}' \mathbf{X})) = \exp\left(\mathbf{t}' \boldsymbol{\mu} + \frac{1}{2} \mathbf{t}' \boldsymbol{\Sigma} \mathbf{t}\right)$$

(see Exercise 5.9) and its characteristic function is

$$C(\mathbf{t}) = \mathbb{E}(\exp(i\mathbf{t}' \mathbf{X})) = \exp\left(i\boldsymbol{\mu}' \boldsymbol{\lambda} - \frac{1}{2} \mathbf{t}' \boldsymbol{\Sigma} \mathbf{t}\right)$$

(see Exercise 5.10).

Our definitions of the univariate and multivariate normal distributions require non-singularity ( $\sigma^2 > 0$  and  $\boldsymbol{\Sigma} > 0$ ) but in some cases it is useful for the definitions to be extended to the singular case. For example, if  $\sigma^2 = 0$  then  $X \sim N(\boldsymbol{\mu}, 0) = \boldsymbol{\mu}$  with probability one. This extension can be made easily by re-defining the multivariate normal distribution by the moment generating function  $M(\mathbf{t}) = \exp(\mathbf{t}' \boldsymbol{\mu} + \frac{1}{2} \mathbf{t}' \boldsymbol{\Sigma} \mathbf{t})$ . This allows for both non-singular and singular covariance matrices.

An important property of normal random vectors is that affine functions are also multivariate normal.

**Theorem 5.4** If  $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  and  $\mathbf{Y} = \mathbf{a} + \mathbf{B}X$ , then  $\mathbf{Y} \sim N(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}')$ .

The proof is presented in Section 5.20.

One simple implication of Theorem 5.4 is that if  $X$  is multivariate normal, then each component of  $X$  is univariate normal.

Another useful property of the multivariate normal distribution is that uncorrelatedness is the same as independence. That is, if a vector is multivariate normal, subsets of variables are independent if and only if they are uncorrelated.

**Theorem 5.5** If  $\mathbf{X} = (X'_1, X'_2)'$  is multivariate normal,  $X_1$  and  $X_2$  are uncorrelated if and only if they are independent.

The proof is presented in Section 5.20.

The normal distribution is frequently used for inference to calculate critical values and p-values. This involves evaluating the normal cdf  $\Phi(x)$  and its inverse. Since the cdf  $\Phi(x)$  is not available in closed form, statistical textbooks have traditionally provided tables for this purpose. Such tables are not used currently as now these calculations are embedded in statistical software. For convenience, we list the appropriate commands in MATLAB, R, and Stata to compute the cumulative distribution function of commonly used statistical distributions.

Numerical Cumulative Distribution Function To calculate $\mathbb{P}(X \leq x)$ for given $x$			
	MATLAB	R	Stata
$N(0, 1)$	<code>normcdf(x)</code>	<code>pnorm(x)</code>	<code>normal(x)</code>
$\chi^2_r$	<code>chi2cdf(x, r)</code>	<code>pchisq(x, r)</code>	<code>chi2(r, x)</code>
$t_r$	<code>tcdf(x, r)</code>	<code>pt(x, r)</code>	<code>1-ttail(r, x)</code>
$F_{r,k}$	<code>fcdf(x, r, k)</code>	<code>pf(x, r, k)</code>	<code>F(r, k, x)</code>
$\chi^2_r(d)$	<code>ncx2cdf(x, r, d)</code>	<code>pchisq(x, r, d)</code>	<code>nchi2(r, d, x)</code>
$F_{r,k}(d)$	<code>ncfcdf(x, r, k, d)</code>	<code>pf(x, r, k, d)</code>	<code>1-nFtail(r, k, d, x)</code>

Here we list the appropriate commands to compute the inverse probabilities (quantiles) of the same distributions.

Numerical Quantile Function To calculate $x$ which solves $p = \mathbb{P}(X \leq x)$ for given $p$			
	MATLAB	R	Stata
$N(0, 1)$	<code>norminv(p)</code>	<code>qnorm(p)</code>	<code>invnormal(p)</code>
$\chi^2_r$	<code>chi2inv(p, r)</code>	<code>qchisq(p, r)</code>	<code>invchi2(r, p)</code>
$t_r$	<code>tinv(p, r)</code>	<code>qt(p, r)</code>	<code>invttail(r, 1-p)</code>
$F_{r,k}$	<code>finv(p, r, k)</code>	<code>qf(p, r, k)</code>	<code>invF(r, k, p)</code>
$\chi^2_r(d)$	<code>ncx2inv(p, r, d)</code>	<code>qchisq(p, r, d)</code>	<code>invnchi2(r, d, p)</code>
$F_{r,k}(d)$	<code>ncfinv(p, r, k, d)</code>	<code>qf(p, r, k, d)</code>	<code>invnFtail(r, k, d, 1-p)</code>

### 5.3 Chi-Square Distribution

Many important distributions can be derived as transformation of multivariate normal random vectors, including the chi-square, the student  $t$ , and the  $F$ . In this section we introduce the chi-square distribution.

Let  $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_r)$  be multivariate standard normal and define  $Q = \mathbf{X}'\mathbf{X}$ . The distribution of  $Q$  is called **chi-square** with  $r$  degrees of freedom, written as  $Q \sim \chi^2_r$ .

The mean and variance of  $Q \sim \chi^2_r$  are  $r$  and  $2r$ , respectively. (See Exercise 5.11.)

The chi-square distribution function is frequently used for inference (critical values and p-values). In practice these calculations are performed numerically by statistical software, but for completeness we provide the density function.

**Theorem 5.6** The density of  $\chi^2_r$  is

$$f(x) = \frac{1}{2^{r/2}\Gamma\left(\frac{r}{2}\right)} x^{r/2-1} e^{-x/2}, \quad x > 0. \quad (5.2)$$

The proof is presented in Section 5.20.

Plots of the chi-square density function for  $r = 2, 3, 4$ , and  $6$  are displayed in Figure 5.2

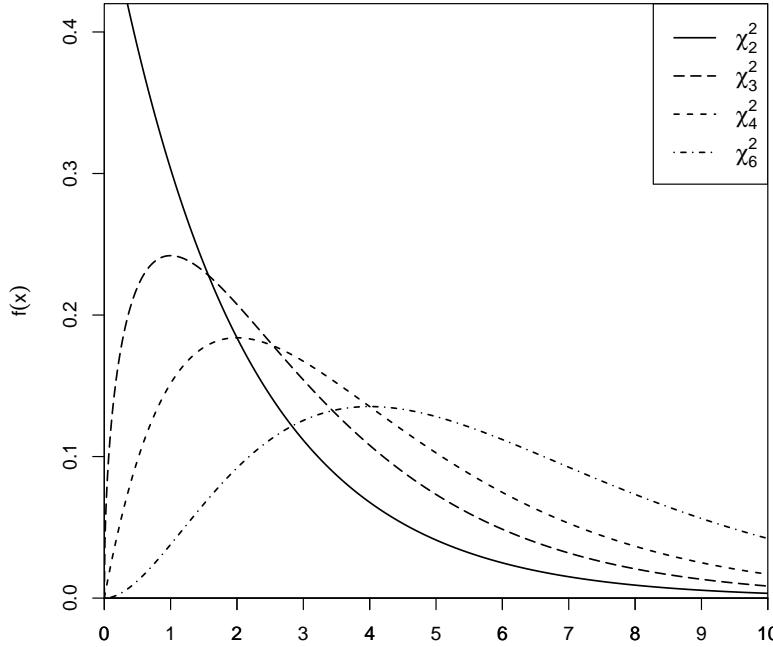


Figure 5.2: Chi-Square Density

## 5.4 Student t Distribution

Let  $Z \sim N(0, 1)$  and  $Q \sim \chi_r^2$  be independent, and define  $T = Z/\sqrt{Q/r}$ . The distribution of  $T$  is called the **student  $t$**  with  $r$  degrees of freedom, and is written  $T \sim t_r$ . Like the chi-square, the distribution only depends on the degree of freedom parameter  $r$ .

**Theorem 5.7** The density of  $T$  is

$$f(x) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)} \left(1 + \frac{x^2}{r}\right)^{-\left(\frac{r+1}{2}\right)}, \quad -\infty < x < \infty.$$

The proof is presented in Section 5.20.

Plots of the student  $t$  density function are displayed in Figure 5.3 for  $r = 1, 2, 5$  and  $\infty$ . The density function of the student  $t$  is bell-shaped like the normal density function, but the  $t$  has thicker tails. The  $t$  distribution has the property that moments below  $r$  are finite, but absolute moments greater than or equal to  $r$  are infinite.

The student  $t$  can also be seen as a generalization of the standard normal, for the latter is obtained as the limiting case where  $r$  is taken to infinity.

**Theorem 5.8** Let  $f_r(x)$  be the student  $t$  density. As  $r \rightarrow \infty$ ,  $f_r(x) \rightarrow \phi(x)$ .

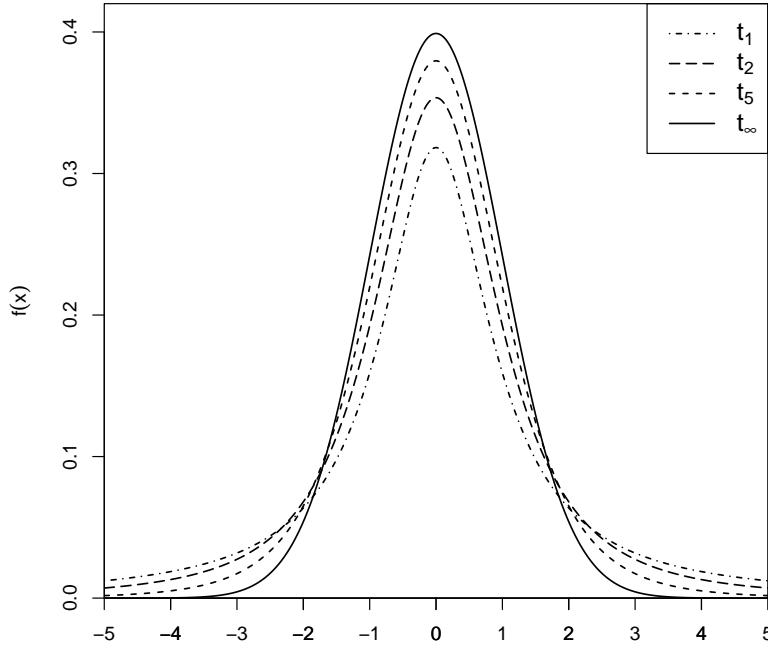


Figure 5.3: Student t Density

The proof is presented in Section 5.20.

This means that the  $t_\infty$  distribution equals the standard normal distribution.

Another special case of the student  $t$  distribution occurs when  $r = 1$  and is known as the **Cauchy** distribution. The Cauchy density function is

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty.$$

A Cauchy random variable  $T = Z_1/Z_2$  can also be derived as the ratio of two independent  $N(0, 1)$  variables. The Cauchy has the property that it has no finite integer moments.

### William Gosset

William S. Gosset (1876-1937) of England is most famous for his derivation of the student's  $t$  distribution, published in the paper "The probable error of a mean" in 1908. At the time, Gosset worked at Guiness Brewery, which prohibited its employees from publishing in order to prevent the possible loss of trade secrets. To circumvent this barrier, Gosset published under the pseudonym "Student". Consequently, this famous distribution is known as the student  $t$  rather than Gosset's  $t$ !

## 5.5 F Distribution

Let  $Q_m \sim \chi_m^2$  and  $Q_r \sim \chi_r^2$  be independent. The distribution of  $F = (Q_m/m) / (Q_r/r)$  is called the  $F$  distribution with degree of freedom parameters  $m$  and  $r$ , and we write  $F \sim F_{m,r}$ .

**Theorem 5.9** The density of  $F_{m,r}$  is

$$f(x) = \frac{\left(\frac{m}{r}\right)^{m/2} x^{m/2-1} \Gamma\left(\frac{m+r}{2}\right)}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{r}{2}\right) \left(1 + \frac{m}{r}x\right)^{(m+r)/2}}, \quad x > 0.$$

The proof is presented in Section 5.20.

Plots of the  $F_{m,r}$  density for  $m = 2, 3, 6, 8$ , and  $r = 10$  are displayed in Figure 5.4.

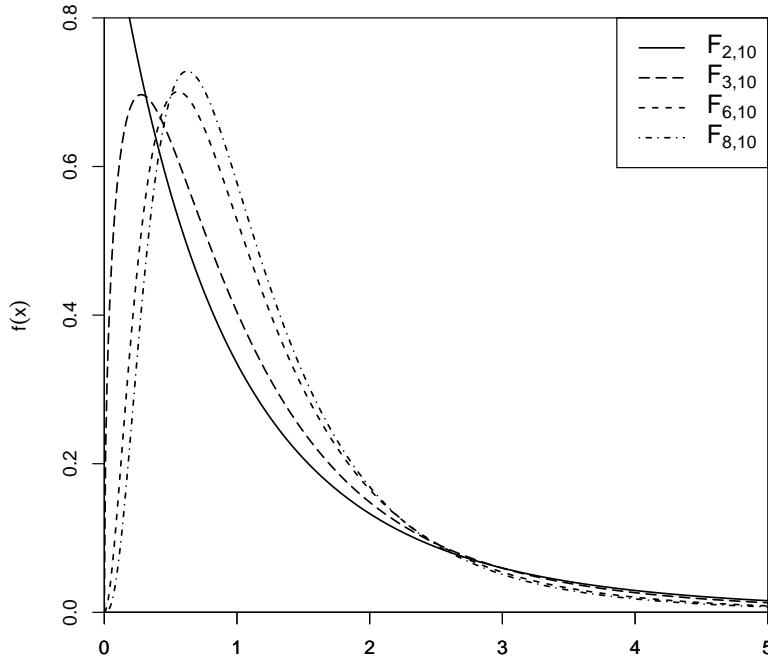


Figure 5.4: F Density

If  $m = 1$  then we can write  $Q_1 = Z^2$  where  $Z \sim N(0, 1)$ , and  $F = Z^2 / (Q_r/r) = (Z/\sqrt{Q_r/r})^2 = T^2$ , the square of a student  $t$  with  $r$  degree of freedom. Thus the  $F$  distribution with  $m = 1$  is equal to the squared student  $t$  distribution. In this sense the  $F$  distribution is a generalization of the student  $t$ .

As a limiting case, as  $r \rightarrow \infty$  the  $F$  distribution simplifies to  $F \rightarrow Q_m/m$ , a normalized  $\chi_m^2$ . Thus the  $F$  distribution is also a generalization of the  $\chi_m^2$  distribution.

**Theorem 5.10** Let  $f_{m,r}(x)$  be the density of  $mF$ . As  $r \rightarrow \infty$ ,  $f_{m,r}(x) \rightarrow f_m(x)$ , the density of  $\chi_m^2$ .

The proof is presented in Section 5.20.

The  $F$  distribution was tabulated by Snedecor (1934). He introduced the notation  $F$  as the distribution is related to Sir Ronald Fisher's work on the analysis of variance.

## 5.6 Non-Central Chi-Square and F Distributions

For some theoretical applications, including the study of the power of statistical tests, it is useful to define a non-central version of the chi-square distribution. When  $X \sim N(\boldsymbol{\mu}, I_r)$  is multivariate normal, we say that  $Q = X'X$  has a **non-central chi-square** distribution, with  $r$  degrees of freedom and non-centrality parameter  $\lambda = \boldsymbol{\mu}'\boldsymbol{\mu}$ , and is written as  $Q \sim \chi_r^2(\lambda)$ . The non-central chi-square simplifies to the central (conventional) chi-square when  $\lambda = 0$ , so that  $\chi_r^2(0) = \chi_r^2$ .

**Theorem 5.11** The density of  $\chi_r^2(\lambda)$  is

$$f(x) = \sum_{i=0}^{\infty} \frac{e^{-\lambda/2}}{i!} \left(\frac{\lambda}{2}\right)^i f_{r+2i}(x), \quad x > 0 \quad (5.3)$$

where  $f_{r+2i}(x)$  is the  $\chi_{r+2i}^2$  density function (5.2).

The proof is presented in Section 5.20.

Plots of the  $\chi_3^2(\lambda)$  density for  $\lambda = 0, 2, 4, \text{ and } 6$  are displayed in Figure 5.5.

Interestingly, as can be seen from the formula (5.3), the distribution of  $\chi_r^2(\lambda)$  only depends on the scalar non-centrality parameter  $\lambda$ , not the entire mean vector  $\boldsymbol{\mu}$ .

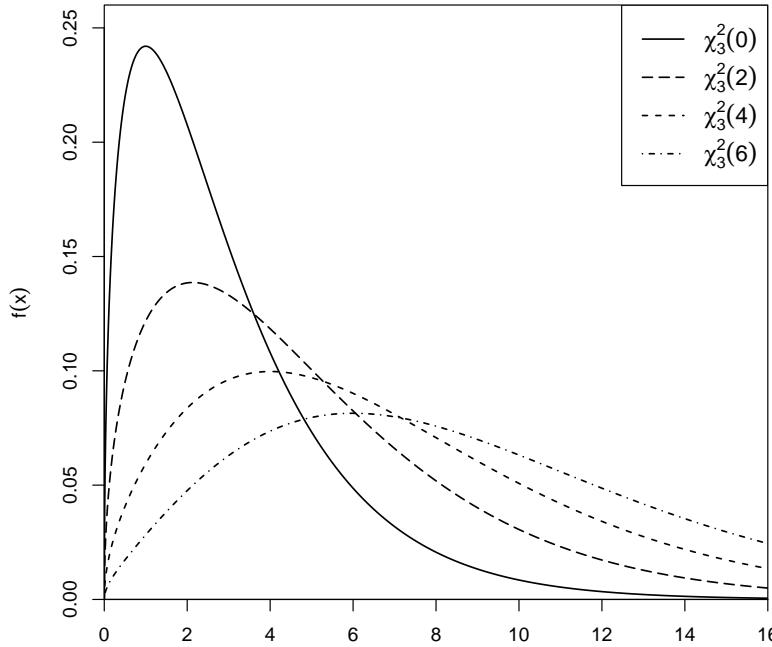


Figure 5.5: Non-Central Chi-Square Density

A useful fact about the central and non-central chi-square distributions is that they also can be derived from multivariate normal distributions with general covariance matrices.

**Theorem 5.12** If  $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A})$  with  $\mathbf{A} > 0$ ,  $r \times r$ , then  $\mathbf{X}'\mathbf{A}^{-1}\mathbf{X} \sim \chi_r^2(\lambda)$ , where  $\lambda = \boldsymbol{\mu}'\mathbf{A}^{-1}\boldsymbol{\mu}$ .

The proof is presented in Section 5.20.

In particular, Theorem 5.12 applies to the central chi-squared distribution, so if  $\mathbf{X} \sim \mathcal{N}(0, \mathbf{A})$  then  $\mathbf{X}'\mathbf{A}^{-1}\mathbf{X} \sim \chi_r^2$ .

Similarly with the non-central chi-square we define the non-central  $F$  distribution. If  $Q_m \sim \chi_m^2(\lambda)$  and  $Q_r \sim \chi_r^2$  are independent, then  $F = (Q_m/m) / (Q_r/r)$  is called a **non-central  $F$**  with degree of freedom parameters  $m$  and  $r$  and non-centrality parameter  $\lambda$ .

## 5.7 Joint Normality and Linear Regression

Suppose the variables  $(y, \mathbf{x})$  are jointly normally distributed. Consider the best linear predictor of  $y$  given  $\mathbf{x}$

$$y = \mathbf{x}'\boldsymbol{\beta} + \alpha + e.$$

By the properties of the best linear predictor,  $\mathbb{E}(\mathbf{x}e) = 0$  and  $\mathbb{E}(e) = 0$ , so  $\mathbf{x}$  and  $e$  are uncorrelated. Since  $(e, \mathbf{x})$  is an affine transformation of the normal vector  $(y, \mathbf{x})$ , it follows that  $(e, \mathbf{x})$  is jointly normal (Theorem 5.4). Since  $(e, \mathbf{x})$  is jointly normal and uncorrelated they are independent (Theorem 5.5). Independence implies that

$$\mathbb{E}(e | \mathbf{x}) = \mathbb{E}(e) = 0$$

and

$$\mathbb{E}(e^2 | \mathbf{x}) = \mathbb{E}(e^2) = \sigma^2$$

which are properties of a homoskedastic linear CEF.

We have shown that when  $(y, \mathbf{x})$  are jointly normally distributed, they satisfy a normal linear CEF

$$y = \mathbf{x}'\boldsymbol{\beta} + \alpha + e$$

where

$$e \sim \mathcal{N}(0, \sigma^2)$$

is independent of  $\mathbf{x}$ .

This is a classical motivation for the linear regression model.

## 5.8 Normal Regression Model

The normal regression model is the linear regression model with an independent normal error

$$\begin{aligned} y &= \mathbf{x}'\boldsymbol{\beta} + e \\ e &\sim \mathcal{N}(0, \sigma^2). \end{aligned} \tag{5.4}$$

As we learned in Section 5.7, the normal regression model holds when  $(y, \mathbf{x})$  are jointly normally distributed. Normal regression, however, does not require joint normality. All that is required is that the conditional distribution of  $y$  given  $\mathbf{x}$  is normal (the marginal distribution of  $\mathbf{x}$  is unrestricted). In this sense the normal regression model is broader than joint normality. Notice that for notational convenience we have written (5.4) so that  $\mathbf{x}$  contains the intercept.

Normal regression is a parametric model, where likelihood methods can be used for estimation, testing, and distribution theory. The **likelihood** is the name for the joint probability density of the data, evaluated at the observed sample, and viewed as a function of the parameters. The maximum likelihood

estimator is the value which maximizes this likelihood function. Let us now derive the likelihood of the normal regression model.

First, observe that model (5.4) is equivalent to the statement that the conditional density of  $y$  given  $\mathbf{x}$  takes the form

$$f(y | \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(y - \mathbf{x}'\boldsymbol{\beta})^2\right).$$

Under the assumption that the observations are mutually independent, this implies that the conditional density of  $(y_1, \dots, y_n)$  given  $(\mathbf{x}_1, \dots, \mathbf{x}_n)$  is

$$\begin{aligned} f(y_1, \dots, y_n | \mathbf{x}_1, \dots, \mathbf{x}_n) &= \prod_{i=1}^n f(y_i | \mathbf{x}_i) \\ &= \prod_{i=1}^n \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mathbf{x}'_i\boldsymbol{\beta})^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}'_i\boldsymbol{\beta})^2\right) \\ &\stackrel{def}{=} L(\boldsymbol{\beta}, \sigma^2) \end{aligned}$$

and is called the **likelihood function**.

For convenience, it is typical to work with the natural logarithm

$$\begin{aligned} \log f(y_1, \dots, y_n | \mathbf{x}_1, \dots, \mathbf{x}_n) &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}'_i\boldsymbol{\beta})^2 \\ &\stackrel{def}{=} \log L(\boldsymbol{\beta}, \sigma^2) \end{aligned} \quad (5.5)$$

which is called the **log-likelihood function**.

The **maximum likelihood estimator (MLE)**  $(\hat{\boldsymbol{\beta}}_{\text{mle}}, \hat{\sigma}_{\text{mle}}^2)$  is the value which maximizes the log-likelihood. (It is equivalent to maximize the likelihood or the log-likelihood. See Exercise 5.16.) We can write the maximization problem as

$$(\hat{\boldsymbol{\beta}}_{\text{mle}}, \hat{\sigma}_{\text{mle}}^2) = \underset{\boldsymbol{\beta} \in \mathbb{R}^k, \sigma^2 > 0}{\operatorname{argmax}} \log L(\boldsymbol{\beta}, \sigma^2). \quad (5.6)$$

In most applications of maximum likelihood, the MLE must be found by numerical methods. However, in the case of the normal regression model we can find an explicit expression for  $\hat{\boldsymbol{\beta}}_{\text{mle}}$  and  $\hat{\sigma}_{\text{mle}}^2$  as functions of the data.

The maximizers  $(\hat{\boldsymbol{\beta}}_{\text{mle}}, \hat{\sigma}_{\text{mle}}^2)$  of (5.6) jointly solve the first-order conditions (FOC)

$$0 = \frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta}, \sigma^2) \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\text{mle}}, \sigma^2=\hat{\sigma}_{\text{mle}}^2} = \frac{1}{\hat{\sigma}_{\text{mle}}^2} \sum_{i=1}^n \mathbf{x}_i (y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{\text{mle}}) \quad (5.7)$$

$$0 = \frac{\partial}{\partial \sigma^2} \log L(\boldsymbol{\beta}, \sigma^2) \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\text{mle}}, \sigma^2=\hat{\sigma}_{\text{mle}}^2} = -\frac{n}{2\hat{\sigma}_{\text{mle}}^2} + \frac{1}{\hat{\sigma}_{\text{mle}}^4} \sum_{i=1}^n (y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{\text{mle}})^2. \quad (5.8)$$

The first FOC (5.7) is proportional to the first-order conditions for the least-squares minimization problem of Section 3.6. It follows that the MLE satisfies

$$\hat{\boldsymbol{\beta}}_{\text{mle}} = \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{x}_i y_i \right) = \hat{\boldsymbol{\beta}}_{\text{ols}}.$$

That is, the MLE for  $\boldsymbol{\beta}$  is algebraically identical to the OLS estimator.

Solving the second FOC (5.8) for  $\hat{\sigma}_{\text{mle}}^2$  we find

$$\hat{\sigma}_{\text{mle}}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{\text{mle}})^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{\text{ols}})^2 = \frac{1}{n} \sum_{i=1}^n \hat{\sigma}_i^2 = \hat{\sigma}_{\text{ols}}^2.$$

Thus the MLE for  $\sigma^2$  is identical to the OLS/moment estimator from (3.27).

Since the OLS estimator and MLE under normality are equivalent,  $\hat{\beta}$  is described by some authors as the maximum likelihood estimator, and by other authors as the least-squares estimator. It is important to remember, however, that  $\hat{\beta}$  is only the MLE when the error  $e$  has a known normal distribution, and not otherwise.

Plugging the estimators into (5.5) we obtain the maximized log-likelihood

$$\log L(\hat{\beta}_{\text{mle}}, \hat{\sigma}_{\text{mle}}^2) = -\frac{n}{2} \log(2\pi\hat{\sigma}_{\text{mle}}^2) - \frac{n}{2}. \quad (5.9)$$

The log-likelihood is typically reported as a measure of fit.

It may seem surprising that the MLE  $\hat{\beta}_{\text{mle}}$  is numerically equal to the OLS estimator, despite emerging from quite different motivations. It is not completely accidental. The least-squares estimator minimizes a particular sample loss function – the sum of squared error criterion – and most loss functions are equivalent to the likelihood of a specific parametric distribution, in this case the normal regression model. In this sense it is not surprising that the least-squares estimator can be motivated as either the minimizer of a sample loss function or as the maximizer of a likelihood function.

### Carl Friedrich Gauss

The mathematician Carl Friedrich Gauss (1777-1855) proposed the normal regression model, and derived the least squares estimator as the maximum likelihood estimator for this model. He claimed to have discovered the method in 1795 at the age of eighteen, but did not publish the result until 1809. Interest in Gauss's approach was reinforced by Laplace's simultaneous discovery of the central limit theorem, which provided a justification for viewing random disturbances as approximately normal.

## 5.9 Distribution of OLS Coefficient Vector

In the normal linear regression model we can derive exact sampling distributions for the OLS/MLE estimator, residuals, and variance estimator. In this section we derive the distribution of the OLS coefficient estimator.

The normality assumption  $e_i | \mathbf{x}_i \sim N(0, \sigma^2)$  combined with independence of the observations has the multivariate implication

$$\mathbf{e} | \mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_n \sigma^2).$$

That is, the error vector  $\mathbf{e}$  is independent of  $\mathbf{X}$  and is normally distributed.

Recall that the OLS estimator satisfies

$$\hat{\beta} - \beta = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{e}$$

which is a linear function of  $\mathbf{e}$ . Since linear functions of normals are also normal (Theorem 5.4), this implies that conditional on  $\mathbf{X}$ ,

$$\begin{aligned} \hat{\beta} - \beta |_{\mathbf{X}} &\sim (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' N(0, \mathbf{I}_n \sigma^2) \\ &\sim N\left(0, \sigma^2 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1}\right) \\ &= N\left(0, \sigma^2 (\mathbf{X}' \mathbf{X})^{-1}\right). \end{aligned}$$

An alternative way of writing this is

$$\hat{\boldsymbol{\beta}}|_{\mathbf{X}} \sim N(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}' \mathbf{X})^{-1}).$$

This shows that under the assumption of normal errors, the OLS estimator has an exact normal distribution.

**Theorem 5.13** In the linear regression model,

$$\hat{\boldsymbol{\beta}}|_{\mathbf{X}} \sim N(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}' \mathbf{X})^{-1}).$$

Theorems 5.4 and 5.13 imply that any affine function of the OLS estimator is also normally distributed, including individual components. Letting  $\beta_j$  and  $\hat{\beta}_j$  denote the  $j^{th}$  elements of  $\boldsymbol{\beta}$  and  $\hat{\boldsymbol{\beta}}$ , we have

$$\hat{\beta}_j|_{\mathbf{X}} \sim N(\beta_j, \sigma^2 [(\mathbf{X}' \mathbf{X})^{-1}]_{jj}). \quad (5.10)$$

Theorem 5.13 is a statement about the conditional distribution. What about the unconditional distribution? In Section 4.7 we presented Kinal's theorem about the existence of moments for the joint normal regression model. We re-state the result here.

**Theorem 5.14** (Kinal, 1980) If  $\mathbf{y}, \mathbf{x}$  are jointly normal, then for any  $r$ ,  $\mathbb{E} \|\hat{\boldsymbol{\beta}}\|^r < \infty$  if and only if  $r < n - k + 1$ .

## 5.10 Distribution of OLS Residual Vector

Now consider the OLS residual vector. Recall from (3.25) that  $\hat{\mathbf{e}} = \mathbf{M}\mathbf{e}$  where  $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ . This shows that  $\hat{\mathbf{e}}$  is linear in  $\mathbf{e}$ . So conditional on  $\mathbf{X}$ ,

$$\hat{\mathbf{e}} = \mathbf{M}\mathbf{e}|_{\mathbf{X}} \sim N(0, \sigma^2 \mathbf{M}\mathbf{M}') = N(0, \sigma^2 \mathbf{M})$$

the final equality since  $\mathbf{M}$  is idempotent (see Section 3.12). This shows that the residual vector has an exact normal distribution.

Furthermore, it is useful to understand the joint distribution of  $\hat{\boldsymbol{\beta}}$  and  $\hat{\mathbf{e}}$ . This is easiest done by writing the two as a stacked linear function of the error  $\mathbf{e}$ . Indeed,

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\mathbf{e}} \end{pmatrix} = \begin{pmatrix} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} \\ \mathbf{M}\mathbf{e} \end{pmatrix} = \begin{pmatrix} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ \mathbf{M} \end{pmatrix} \mathbf{e}$$

which is a linear function of  $\mathbf{e}$ . The vector thus has a joint normal distribution with covariance matrix

$$\begin{pmatrix} \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} & 0 \\ 0 & \sigma^2 \mathbf{M} \end{pmatrix}.$$

The off-diagonal block is zero because  $\mathbf{X}'\mathbf{M} = 0$  from (3.22). Since this is zero, it follows that  $\hat{\boldsymbol{\beta}}$  and  $\hat{\mathbf{e}}$  are statistically independent (Theorem 5.5).

**Theorem 5.15** In the linear regression model,  $\hat{\mathbf{e}}|_{\mathbf{X}} \sim N(0, \sigma^2 \mathbf{M})$  and is independent of  $\hat{\boldsymbol{\beta}}$ .

The fact that  $\hat{\boldsymbol{\beta}}$  and  $\hat{\mathbf{e}}$  are independent implies that  $\hat{\boldsymbol{\beta}}$  is independent of any function of the residual vector, including individual residuals  $\hat{e}_i$  and the variance estimate  $s^2$  and  $\hat{s}^2$ .

## 5.11 Distribution of Variance Estimator

Next, consider the variance estimator  $s^2$  from (4.26). Using (3.29), it satisfies  $(n - k) s^2 = \hat{\mathbf{e}}' \hat{\mathbf{e}} = \mathbf{e}' \mathbf{M} \mathbf{e}$ . The spectral decomposition of  $\mathbf{M}$  (see equation (A.4)) is  $\mathbf{M} = \mathbf{H} \Lambda \mathbf{H}'$  where  $\mathbf{H}' \mathbf{H} = \mathbf{I}_n$  and  $\Lambda$  is diagonal with the eigenvalues of  $\mathbf{M}$  on the diagonal. Since  $\mathbf{M}$  is idempotent with rank  $n - k$  (see Section 3.12) it has  $n - k$  eigenvalues equalling 1 and  $k$  eigenvalues equalling 0, so

$$\Lambda = \begin{bmatrix} \mathbf{I}_{n-k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_k \end{bmatrix}.$$

Let  $\mathbf{u} = \mathbf{H}' \mathbf{e} \sim N(\mathbf{0}, \mathbf{I}_n \sigma^2)$  (see Exercise 5.14) and partition  $\mathbf{u} = (\mathbf{u}'_1, \mathbf{u}'_2)'$  where  $\mathbf{u}_1 \sim N(\mathbf{0}, \mathbf{I}_{n-k} \sigma^2)$ . Then

$$\begin{aligned} (n - k) s^2 &= \mathbf{e}' \mathbf{M} \mathbf{e} \\ &= \mathbf{e}' \mathbf{H} \begin{bmatrix} \mathbf{I}_{n-k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{H}' \mathbf{e} \\ &= \mathbf{u}' \begin{bmatrix} \mathbf{I}_{n-k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{u} \\ &= \mathbf{u}'_1 \mathbf{u}_1 \\ &\sim \sigma^2 \chi^2_{n-k}. \end{aligned}$$

We see that in the normal regression model the exact distribution of  $s^2$  is a scaled chi-square.

Since  $\hat{\mathbf{e}}$  is independent of  $\hat{\boldsymbol{\beta}}$  it follows that  $s^2$  is independent of  $\hat{\boldsymbol{\beta}}$  as well.

**Theorem 5.16** In the linear regression model,

$$\frac{(n - k) s^2}{\sigma^2} \sim \chi^2_{n-k}$$

and is independent of  $\hat{\boldsymbol{\beta}}$ .

## 5.12 t-statistic

An alternative way of writing (5.10) is

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 [(\mathbf{X}' \mathbf{X})^{-1}]_{jj}}} \sim N(0, 1).$$

This is sometimes called a **standardized** statistic, as the distribution is the standard normal.

Now take the standardized statistic and replace the unknown variance  $\sigma^2$  with its estimator  $s^2$ . We call this a **t-ratio** or **t-statistic**

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2[(X'X)^{-1}]_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)}$$

where  $s(\hat{\beta}_j)$  is the classical (homoskedastic) standard error for  $\hat{\beta}_j$  from (4.37). We will sometimes write the t-statistic as  $T(\beta_j)$  to explicitly indicate its dependence on the parameter value  $\beta_j$ , and sometimes will simplify notation and write the t-statistic as  $T$  when the dependence is clear from the context.

By some algebraic re-scaling we can write the t-statistic as the ratio of the standardized statistic and the square root of the scaled variance estimator. Since the distributions of these two components are normal and chi-square, respectively, and independent, then we can deduce that the t-statistic has the distribution

$$\begin{aligned} T &= \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2[(X'X)^{-1}]_{jj}}} \Bigg/ \sqrt{\frac{(n-k)s^2}{\sigma^2}} \Bigg/ \sqrt{(n-k)} \\ &\sim \frac{N(0, 1)}{\sqrt{\chi^2_{n-k}/(n-k)}} \\ &\sim t_{n-k} \end{aligned}$$

a student  $t$  distribution with  $n - k$  degrees of freedom.

This derivation shows that the t-ratio has a sampling distribution which depends only on the quantity  $n - k$ . The distribution does not depend on any other features of the data. In this context, we say that the distribution of the t-ratio is **pivotal**, meaning that it does not depend on unknowns.

The trick behind this result is scaling the centered coefficient by its standard error, and recognizing that each depends on the unknown  $\sigma$  only through scale. Thus the ratio of the two does not depend on  $\sigma$ . This trick (scaling to eliminate dependence on unknowns) is known as **studentization**.

**Theorem 5.17** In the normal regression model,  $T \sim t_{n-k}$ .

An important caveat about Theorem 5.17 is that it only applies to the t-statistic constructed with the homoskedastic (old-fashioned) standard error estimator. It does not apply to a t-statistic constructed with any of the robust standard error estimators. In fact, the robust t-statistics can have finite sample distributions which deviate considerably from  $t_{n-k}$  even when the regression errors are independent  $N(0, \sigma^2)$ . Thus the distributional result in Theorem 5.17, and the use of the t distribution in finite samples, should only be applied to classical t-statistics.

## 5.13 Confidence Intervals for Regression Coefficients

The OLS estimator  $\hat{\beta}$  is a **point estimator** for a coefficient  $\beta$ . A broader concept is a **set or interval estimator** which takes the form  $\hat{C} = [\hat{L}, \hat{U}]$ . The goal of an interval estimator  $\hat{C}$  is to contain the true value, e.g.  $\beta \in \hat{C}$ , with high probability.

The interval estimator  $\hat{C}$  is a function of the data and hence is random.

An interval estimator  $\hat{C}$  is called a  $1 - \alpha$  **confidence interval** when  $\mathbb{P}(\beta \in \hat{C}) = 1 - \alpha$  for a selected value of  $\alpha$ . The value  $1 - \alpha$  is called the **coverage probability**. Typical choices for the coverage probability  $1 - \alpha$  are 0.95 or 0.90.

The probability calculation  $\mathbb{P}(\beta \in \hat{C})$  is easily mis-interpreted as treating  $\beta$  as random and  $\hat{C}$  as fixed. (The probability that  $\beta$  is in  $\hat{C}$ .) This is not the appropriate interpretation. Instead, the correct interpretation is that the probability  $\mathbb{P}(\beta \in \hat{C})$  treats the point  $\beta$  as fixed and the set  $\hat{C}$  as random. It is the probability that the random set  $\hat{C}$  covers (or contains) the fixed true coefficient  $\beta$ .

There is not a unique method to construct confidence intervals. For example, one simple (yet silly) interval is

$$\hat{C} = \begin{cases} \mathbb{R} & \text{with probability } 1 - \alpha \\ \{\hat{\beta}\} & \text{with probability } \alpha \end{cases}.$$

If  $\hat{\beta}$  has a continuous distribution, then by construction  $\mathbb{P}(\beta \in \hat{C}) = 1 - \alpha$ , so this confidence interval has perfect coverage. However,  $\hat{C}$  is uninformative about  $\hat{\beta}$  and is therefore not useful.

Instead, a good choice for a confidence interval for the regression coefficient  $\beta$  is obtained by adding and subtracting from the estimator  $\hat{\beta}$  a fixed multiple of its standard error:

$$\hat{C} = [\hat{\beta} - c \cdot s(\hat{\beta}), \quad \hat{\beta} + c \cdot s(\hat{\beta})] \quad (5.11)$$

where  $c > 0$  is a pre-specified constant. This confidence interval is symmetric about the point estimator  $\hat{\beta}$ , and its length is proportional to the standard error  $s(\hat{\beta})$ .

Equivalently,  $\hat{C}$  is the set of parameter values for  $\beta$  such that the t-statistic  $T(\beta)$  is smaller (in absolute value) than  $c$ , that is

$$\hat{C} = \{\beta : |T(\beta)| \leq c\} = \left\{ \beta : -c \leq \frac{\hat{\beta} - \beta}{s(\hat{\beta})} \leq c \right\}.$$

The coverage probability of this confidence interval is

$$\begin{aligned} \mathbb{P}(\beta \in \hat{C}) &= \mathbb{P}(|T(\beta)| \leq c) \\ &= \mathbb{P}(-c \leq T(\beta) \leq c). \end{aligned} \quad (5.12)$$

Since the t-statistic  $T(\beta)$  has the  $t_{n-k}$  distribution, (5.12) equals  $F(c) - F(-c)$ , where  $F(u)$  is the student  $t$  distribution function with  $n - k$  degrees of freedom. Since  $F(-c) = 1 - F(c)$  (see Exercise 5.20) we can write (5.12) as

$$\mathbb{P}(\beta \in \hat{C}) = 2F(c) - 1.$$

This is the **coverage probability** of the interval  $\hat{C}$ , and only depends on the constant  $c$ .

As we mentioned before, a confidence interval has the coverage probability  $1 - \alpha$ . This requires selecting the constant  $c$  so that  $F(c) = 1 - \alpha/2$ . This holds if  $c$  equals the  $1 - \alpha/2$  quantile of the  $t_{n-k}$  distribution. As there is no closed form expression for these quantiles, we compute their values numerically. For example, by `tinv(1-alpha/2, n-k)` in MATLAB. With this choice the confidence interval (5.11) has exact coverage probability  $1 - \alpha$ . By default, Stata reports 95% confidence intervals  $\hat{C}$  for each estimated regression coefficient using the same formula.

**Theorem 5.18** In the normal regression model, (5.11) with  $c = F^{-1}(1 - \alpha/2)$  has coverage probability  $\mathbb{P}(\beta \in \hat{C}) = 1 - \alpha$ .

When the degree of freedom is large the distinction between the student  $t$  and the normal distribution is negligible. In particular, for  $n - k \geq 61$  we have  $c \leq 2.00$  for a 95% interval. Using this value we obtain the most commonly used confidence interval in applied econometric practice:

$$\hat{C} = [\hat{\beta} - 2s(\hat{\beta}), \quad \hat{\beta} + 2s(\hat{\beta})]. \quad (5.13)$$

This is a useful rule-of-thumb. This 95% confidence interval  $\hat{C}$  is simple to compute and can be easily calculated from coefficient estimates and standard errors.

**Theorem 5.19** In the normal regression model, if  $n - k \geq 61$  then (5.13) has coverage probability  $\mathbb{P}(\beta \in \hat{C}) \geq 0.95$ .

Confidence intervals are a simple yet effective tool to assess estimation uncertainty. When reading a set of empirical results, look at the estimated coefficient estimates and the standard errors. For a parameter of interest, compute the confidence interval  $\hat{C}$  and consider the meaning of the spread of the suggested values. If the range of values in the confidence interval are too wide to learn about  $\beta$ , then do not jump to a conclusion about  $\beta$  based on the point estimate alone.

## 5.14 Confidence Intervals for Error Variance

We can also construct a confidence interval for the regression error variance  $\sigma^2$  using the sampling distribution of  $s^2$  from Theorem 5.16, which states that in the normal regression model

$$\frac{(n - k)s^2}{\sigma^2} \sim \chi_{n-k}^2. \quad (5.14)$$

Let  $F(u)$  denote the  $\chi_{n-k}^2$  distribution function, and for some  $\alpha$  set  $c_1 = F^{-1}(\alpha/2)$  and  $c_2 = F^{-1}(1 - \alpha/2)$  (the  $\alpha/2$  and  $1 - \alpha/2$  quantiles of the  $\chi_{n-k}^2$  distribution). Equation (5.14) implies that

$$\mathbb{P}\left(c_1 \leq \frac{(n - k)s^2}{\sigma^2} \leq c_2\right) = F(c_2) - F(c_1) = 1 - \alpha.$$

Rewriting the inequalities we find

$$\mathbb{P}\left((n - k)s^2/c_2 \leq \sigma^2 \leq (n - k)s^2/c_1\right) = 1 - \alpha.$$

This shows that an exact  $1 - \alpha$  confidence interval for  $\sigma^2$  is

$$C = \left[ \frac{(n - k)s^2}{c_2}, \quad \frac{(n - k)s^2}{c_1} \right]. \quad (5.15)$$

**Theorem 5.20** In the normal regression model, (5.15) has coverage probability  $\mathbb{P}(\sigma^2 \in C) = 1 - \alpha$ .

The confidence interval (5.15) for  $\sigma^2$  is asymmetric about the point estimate  $s^2$ , due to the latter's asymmetric sampling distribution.

## 5.15 t Test

A typical goal in an econometric exercise is to assess whether or not coefficient  $\beta$  equals a specific value  $\beta_0$ . Often the specific value to be tested is  $\beta_0 = 0$  but this is not essential. This is called **hypothesis testing**, a subject which will be explored in detail in Chapter 9. In this section and the following we give a short introduction specific to the normal regression model.

For simplicity write the coefficient to be tested as  $\beta$ . The **null hypothesis** is

$$\mathbb{H}_0 : \beta = \beta_0. \quad (5.16)$$

This states that the hypothesis is that the true value of the coefficient  $\beta$  equals the hypothesized value  $\beta_0$ .

The alternative hypothesis is the complement of  $\mathbb{H}_0$ , and is written as

$$\mathbb{H}_1 : \beta \neq \beta_0.$$

This states that the true value of  $\beta$  does not equal the hypothesized value.

We are interested in testing  $\mathbb{H}_0$  against  $\mathbb{H}_1$ . The method is to design a statistic which is informative about  $\mathbb{H}_1$ . If the observed value of the statistic is consistent with random variation under the assumption that  $\mathbb{H}_0$  is true, then we deduce that there is no evidence against  $\mathbb{H}_0$  and consequently do not reject  $\mathbb{H}_0$ . However, if the statistic takes a value which is unlikely to occur under the assumption that  $\mathbb{H}_0$  is true, then we deduce that there is evidence against  $\mathbb{H}_0$ , and consequently we reject  $\mathbb{H}_0$  in favor of  $\mathbb{H}_1$ . The main steps are to design a test statistic and to characterize its sampling distribution.

The standard statistic to test  $\mathbb{H}_0$  against  $\mathbb{H}_1$  is the absolute value of the t-statistic

$$|T| = \left| \frac{\hat{\beta} - \beta_0}{s(\hat{\beta})} \right|. \quad (5.17)$$

If  $\mathbb{H}_0$  is true, then we expect  $|T|$  to be small, but if  $\mathbb{H}_1$  is true then we would expect  $|T|$  to be large. Hence the standard rule is to reject  $\mathbb{H}_0$  in favor of  $\mathbb{H}_1$  for large values of the t-statistic  $|T|$ , and otherwise fail to reject  $\mathbb{H}_0$ . Thus the hypothesis test takes the form

Reject  $\mathbb{H}_0$  if  $|T| > c$ .

The constant  $c$  which appears in the statement of the test is called the **critical value**. Its value is selected to control the probability of false rejections. When the null hypothesis is true,  $|T|$  has an exact student  $t$  distribution (with  $n - k$  degrees of freedom) in the normal regression model. Thus for a given value of  $c$  the probability of false rejection is

$$\begin{aligned} \mathbb{P}(\text{Reject } \mathbb{H}_0 \mid \mathbb{H}_0) &= \mathbb{P}(|T| > c \mid \mathbb{H}_0) \\ &= \mathbb{P}(T > c \mid \mathbb{H}_0) + \mathbb{P}(T < -c \mid \mathbb{H}_0) \\ &= 1 - F(c) + F(-c) \\ &= 2(1 - F(c)) \end{aligned}$$

where  $F(u)$  is the  $t_{n-k}$  distribution function. This is the probability of false rejection, and is decreasing in the critical value  $c$ . We select the value  $c$  so that this probability equals a pre-selected value called the **significance level**, which is typically written as  $\alpha$ . It is conventional to set  $\alpha = 0.05$ , though this is not a hard rule. We then select  $c$  so that  $F(c) = 1 - \alpha/2$ , which means that  $c$  is the  $1 - \alpha/2$  quantile (inverse CDF) of the  $t_{n-k}$  distribution, the same as used for confidence intervals. With this choice, the decision rule “Reject  $\mathbb{H}_0$  if  $|T| > c$ ” has a significance level (false rejection probability) of  $\alpha$ .

**Theorem 5.21** In the normal regression model, if the null hypothesis (5.16) is true, then for  $|T|$  defined in (5.17),  $|T| \sim t_{n-k}$ . If  $c$  is set so that  $\mathbb{P}(|t_{n-k}| \geq c) = \alpha$ , then the test “Reject  $\mathbb{H}_0$  in favor of  $\mathbb{H}_1$  if  $|T| > c$ ” has significance level  $\alpha$ .

To report the result of a hypothesis test we need to pre-determine the significance level  $\alpha$  in order to calculate the critical value  $c$ . This can be inconvenient and arbitrary. A simplification is to report what is known as the **p-value** of the test. In general, when a test takes the form “Reject  $\mathbb{H}_0$  if  $S > c$ ” and  $S$  has null distribution  $G(u)$ , then the p-value of the test is  $p = 1 - G(S)$ . A test with significance level  $\alpha$  can be restated as “Reject  $\mathbb{H}_0$  if  $p < \alpha$ ”. It is sufficient to report the p-value  $p$ , and we can interpret the value

of  $p$  as indexing the test's strength of rejection of the null hypothesis. Thus a p-value of 0.07 might be interpreted as “nearly significant”, 0.05 as “borderline significant”, and 0.001 as “highly significant”. In the context of the normal regression model, the p-value of a t-statistic  $|T|$  is  $p = 2(1 - F_{n-k}(|T|))$  where  $F_{n-k}$  is the CDF of the student  $t$  with  $n-k$  degrees of freedom. For example, in MATLAB the calculation is  $2*(1-tcdf(abs(t), n-k))$ . In Stata, the default is that for any estimated regression, t-statistics for each estimated coefficient are reported along with their p-values calculated using this same formula. These t-statistics test the hypotheses that each coefficient is zero.

A p-value reports the strength of evidence against  $H_0$  but is not itself a probability. A common misunderstanding is that the p-value is the “probability that the null hypothesis is true”. This is an incorrect interpretation. It is a statistic, and is random, and is a measure of the evidence against  $H_0$ , nothing more.

## 5.16 Likelihood Ratio Test

In the previous section we described the t-test as the standard method to test a hypothesis on a single coefficient in a regression. In many contexts, however, we want to simultaneously assess a set of coefficients. In the normal regression model, this can be done by an  $F$  test, which can be derived from the likelihood ratio test.

Partition the regressors as  $\mathbf{x}_i = (\mathbf{x}'_{1i}, \mathbf{x}'_{2i})'$  and similarly partition the coefficient vector as  $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2)'$ . Then the regression model can be written as

$$y_i = \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + e_i. \quad (5.18)$$

Let  $k = \dim(\mathbf{x}_i)$ ,  $k_1 = \dim(\mathbf{x}_{1i})$ , and  $q = \dim(\mathbf{x}_{2i})$ , so that  $k = k_1 + q$ . Partition the variables so that the hypothesis is that the second set of coefficients are zero, or

$$H_0 : \boldsymbol{\beta}_2 = 0. \quad (5.19)$$

If  $H_0$  is true, then the regressors  $\mathbf{x}_{2i}$  can be omitted from the regression. In this case we can write (5.18) as

$$y_i = \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + e_i. \quad (5.20)$$

We call (5.20) the null model. The alternative hypothesis is that at least one element of  $\boldsymbol{\beta}_2$  is non-zero and is written as

$$H_1 : \boldsymbol{\beta}_2 \neq 0.$$

When models are estimated by maximum likelihood, a well-accepted testing procedure is to reject  $H_0$  in favor of  $H_1$  for large values of the Likelihood Ratio – the ratio of the maximized likelihood function under  $H_1$  and  $H_0$ , respectively. We now construct this statistic in the normal regression model. Recall from (5.9) that the maximized log-likelihood equals

$$\log L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = -\frac{n}{2} \log(2\pi\hat{\sigma}^2) - \frac{n}{2}.$$

We similarly need to calculate the maximized log-likelihood for the constrained model (5.20). By the same steps for derivation of the unconstrained MLE, we can find that the MLE for (5.20) is OLS of  $y_i$  on  $\mathbf{x}_{1i}$ . We can write this estimator as

$$\tilde{\boldsymbol{\beta}}_1 = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{y}$$

with residual

$$\tilde{e}_i = y_i - \mathbf{x}'_{1i} \tilde{\boldsymbol{\beta}}_1$$

and error variance estimate

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \tilde{e}_i^2.$$

We use the tildes “~” rather than the hats “^” above the constrained estimates to distinguish them from the unconstrained estimates. You can calculate similar to (5.9) that the maximized constrained log-likelihood is

$$\log L(\tilde{\boldsymbol{\beta}}_1, \tilde{\sigma}^2) = -\frac{n}{2} \log(2\pi\tilde{\sigma}^2) - \frac{n}{2}.$$

A classic testing procedure is to reject  $\mathbb{H}_0$  for large values of the ratio of the maximized likelihoods. Equivalently, the test rejects  $\mathbb{H}_0$  for large values of twice the difference in the log-likelihood functions. (Multiplying the likelihood difference by two turns out to be a useful scaling.) This equals

$$\begin{aligned} LR &= 2 \left( \left( -\frac{n}{2} \log(2\pi\hat{\sigma}^2) - \frac{n}{2} \right) - \left( -\frac{n}{2} \log(2\pi\tilde{\sigma}^2) - \frac{n}{2} \right) \right) \\ &= n \log \left( \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \right). \end{aligned} \quad (5.21)$$

The likelihood ratio test rejects for large values of  $LR$ , or equivalently (see Exercise 5.22), for large values of

$$F = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/q}{\hat{\sigma}^2/(n-k)}. \quad (5.22)$$

This is known as the  $F$  statistic for the test of hypothesis  $\mathbb{H}_0$  against  $\mathbb{H}_1$ .

To develop an appropriate critical value, we need the null distribution of  $F$ . Recall from (3.29) that  $n\hat{\sigma}^2 = \mathbf{e}' \mathbf{M} \mathbf{e}$  where  $\mathbf{M} = \mathbf{I}_n - \mathbf{P}$  with  $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ . Similarly, under  $\mathbb{H}_0$ ,  $n\tilde{\sigma}^2 = \mathbf{e}' \mathbf{M}_1 \mathbf{e}$  where  $\mathbf{M} = \mathbf{I}_n - \mathbf{P}_1$  with  $\mathbf{P}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$ . You can calculate that  $\mathbf{M}_1 - \mathbf{M} = \mathbf{P} - \mathbf{P}_1$  is idempotent with rank  $q$ . Furthermore,  $(\mathbf{M}_1 - \mathbf{M})\mathbf{M} = 0$ . It follows that  $\mathbf{e}'(\mathbf{M}_1 - \mathbf{M})\mathbf{e} \sim \chi_q^2$  and is independent of  $\mathbf{e}'\mathbf{M}\mathbf{e}$ . Hence

$$F = \frac{\mathbf{e}'(\mathbf{M}_1 - \mathbf{M})\mathbf{e}/q}{\mathbf{e}'\mathbf{M}\mathbf{e}/(n-k)} \sim \frac{\chi_q^2/q}{\chi_{n-k}^2/(n-k)} \sim F_{q,n-k}$$

an exact  $F$  distribution with degrees of freedom  $q$  and  $n - k$ , respectively. Thus under  $\mathbb{H}_0$ , the  $F$  statistic has an exact  $F$  distribution.

The critical values are selected from the upper tail of the  $F$  distribution. For a given significance level  $\alpha$  (typically  $\alpha = 0.05$ ) we select the critical value  $c$  so that  $\mathbb{P}(F_{q,n-k} \geq c) = \alpha$ . (For example, in MATLAB the expression is `finv(1-alpha, q, n-k)`.) The test rejects  $\mathbb{H}_0$  in favor of  $\mathbb{H}_1$  if  $F > c$  and does not reject  $\mathbb{H}_0$  otherwise. The p-value of the test is  $p = 1 - G_{q,n-k}(F)$  where  $G_{q,n-k}(u)$  is the  $F_{q,n-k}$  distribution function. (In MATLAB, the p-value is computed as `1 - fcdf(f, q, n-k)`.) It is equivalent to reject  $\mathbb{H}_0$  if  $F > c$  or  $p < \alpha$ .

In Stata, the command to test multiple coefficients takes the form ‘test X1 X2’ where X1 and X2 are the names of the variables whose coefficients are tested. Stata then reports the  $F$  statistic for the hypothesis that the coefficients are jointly zero along with the p-value calculated using the  $F$  distribution.

**Theorem 5.22** In the normal regression model, if the null hypothesis (5.19) is true, then for  $F$  defined in (5.22),  $F \sim F_{q,n-k}$ . If  $c$  is set so that  $\mathbb{P}(F_{q,n-k} \geq c) = \alpha$ , then the test “Reject  $\mathbb{H}_0$  in favor of  $\mathbb{H}_1$  if  $F > c$ ” has significance level  $\alpha$ .

Theorem 5.22 justifies the  $F$  test in the normal regression model with critical values taken from the  $F$  distribution.

## 5.17 Likelihood Properties

In this section we present some general properties of the likelihood which hold broadly – not just in normal regression.

Suppose that a random vector  $\mathbf{y}$  has the conditional density  $f(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})$  where the function  $f$  is known, and the parameter vector  $\boldsymbol{\theta}$  takes values in a parameter space  $\Theta$ . The log-likelihood function for a random sample  $\{\mathbf{y}_i | \mathbf{x}_i : i = 1, \dots, n\}$  takes the form

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta}).$$

A key property is that the expected log-likelihood is maximized at the true value of the parameter vector. At this point it is useful to make a notational distinction between a generic parameter value  $\boldsymbol{\theta}$  and its true value  $\boldsymbol{\theta}_0$ . Set  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ .

**Theorem 5.23**  $\boldsymbol{\theta}_0 = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \mathbb{E}(\log L(\boldsymbol{\theta}) | \mathbf{X})$

The proof is presented in Section 5.20.

This motivates estimating  $\boldsymbol{\theta}$  by finding the value which maximizes the log-likelihood function. This is the maximum likelihood estimator (MLE):

$$\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \log L(\boldsymbol{\theta}).$$

The **score** of the likelihood function is the vector of partial derivatives with respect to the parameters, evaluated at the true values,

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

The covariance matrix of the score is known as the **Fisher information**:

$$\mathcal{I} = \operatorname{var}\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}_0) | \mathbf{X}\right).$$

Some important properties of the score and information are now presented.

**Theorem 5.24** If  $\log f(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})$  is second differentiable and the support of  $\mathbf{y}$  does not depend on  $\boldsymbol{\theta}$  then

1.  $\mathbb{E}\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} | \mathbf{X}\right) = 0$
2. 
$$\begin{aligned} \mathcal{I} &= \sum_{i=1}^n \mathbb{E}\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta}_0) \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta}_0)' | \mathbf{x}_i\right) \\ &= -\mathbb{E}\left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log L(\boldsymbol{\theta}_0) | \mathbf{X}\right) \end{aligned}$$

The proof is presented in Section 5.20.

The first result says that the score is mean zero. The second result shows that the variance of the score equals the negative expectation of the second derivative matrix. This is known as the **Information Matrix Equality**.

We now establish the famous Cramér-Rao Lower Bound.

**Theorem 5.25** (Cramér-Rao) Under the assumptions of Theorem 5.24, if  $\tilde{\boldsymbol{\theta}}$  is an unbiased estimator of  $\boldsymbol{\theta}$ , then  $\text{var}(\tilde{\boldsymbol{\theta}} | \mathbf{X}) \geq \mathcal{J}^{-1}$ .

The proof is presented in Section 5.20.

Theorem 5.25 shows that the inverse of the information matrix is a lower bound for the covariance matrix of unbiased estimators. This result is similar to the Gauss-Markov Theorem which established a lower bound for unbiased estimators in homoskedastic linear regression.

### Sir Ronald A. Fisher

The British statistician Ronald Fisher (1890-1962) is one of the core founders of modern statistical theory. His contributions include p-values, the concept of Fisher information, and that of sufficient statistics.

## 5.18 Information Bound for Normal Regression

Recall the normal regression log-likelihood which has the parameters  $\boldsymbol{\beta}$  and  $\sigma^2$ . The likelihood scores for this model are

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta}, \sigma^2) &= \frac{1}{\sigma^2} \sum_{i=1}^n \mathbf{x}_i (y_i - \mathbf{x}'_i \boldsymbol{\beta}) \\ &= \frac{1}{\sigma^2} \sum_{i=1}^n \mathbf{x}_i e_i\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial \sigma^2} \log L(\boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 \\ &= \frac{1}{2\sigma^4} \sum_{i=1}^n (e_i^2 - \sigma^2).\end{aligned}$$

It follows that the information matrix is

$$\mathcal{J} = \text{var} \left( \begin{array}{c|c} \frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta}, \sigma^2) & \mathbf{X} \\ \hline \frac{\partial}{\partial \sigma^2} \log L(\boldsymbol{\beta}, \sigma^2) & \mathbf{X}' \end{array} \right) = \begin{pmatrix} \frac{1}{\sigma^2} \mathbf{X}' \mathbf{X} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \quad (5.23)$$

(see Exercise 5.23). The Cramér-Rao Lower Bound is

$$\mathcal{J}^{-1} = \begin{pmatrix} \sigma^2 (\mathbf{X}' \mathbf{X})^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}.$$

This shows that the lower bound for estimation of  $\boldsymbol{\beta}$  is  $\sigma^2 (\mathbf{X}' \mathbf{X})^{-1}$  and the lower bound for  $\sigma^2$  is  $2\sigma^4/n$ .

Since in the homoskedastic linear regression model the OLS estimator is unbiased and has variance  $\sigma^2 (\mathbf{X}' \mathbf{X})^{-1}$ , it follows that OLS is Cramér-Rao efficient in the normal regression model, in the sense that no unbiased estimator has a lower variance matrix. This expands on the Gauss-Markov theorem, which stated that no linear unbiased estimator has a lower variance matrix in the homoskedastic regression model. Notice that the results are complementary. Gauss-Markov efficiency concerns a more narrow class of estimators (linear) but allows a broader model class (linear homoskedastic rather than normal regression). The Cramér-Rao efficiency result is more powerful in that it does not restrict the class of

estimators (beyond unbiasedness) but is more restrictive in the class of models allowed (normal regression).

In contrast, the unbiased estimator  $s^2$  of  $\sigma^2$  has variance  $2\sigma^4/(n-k)$  (see Exercise 5.24) which is larger than the Cramér-Rao lower bound  $2\sigma^4/n$ . Thus in contrast to the coefficient estimator, the variance estimator is not Cramér-Rao efficient.

## 5.19 Gamma Function\*

The normal and related distributions make frequent use of the what is known as the **gamma function**. For  $\alpha > 0$  it is defined as

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x) dx. \quad (5.24)$$

While it appears quite simple, it has some advanced properties. One is that  $\Gamma(\alpha)$  does not have a close-form solution (except for special values of  $\alpha$ ). Thus it is typically represented using the symbol  $\Gamma(\alpha)$  and implemented computationally using numerical methods.

Special values include

$$\Gamma(1) = \int_0^\infty \exp(-x) dx = 1 \quad (5.25)$$

and

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

The latter holds by making the change of variables  $x = u^2/2$  in (5.24) and applying (5.1).

By integration by parts you can show that it satisfies the property

$$\Gamma(1 + \alpha) = \Gamma(\alpha)\alpha.$$

Combined with (5.25) we find that for positive integers  $n$ ,

$$\Gamma(n) = (n-1)!$$

This shows that the gamma function is a continuous version of the factorial.

A useful fact is

$$\int_0^\infty y^{\alpha-1} \exp(-by) dy = b^{-\alpha} \Gamma(\alpha) \quad (5.26)$$

which can be found by applying change-of-variables to the definition (5.24).

Another useful fact is for  $\alpha \in \mathbb{R}$

$$\lim_{n \rightarrow \infty} \frac{\Gamma(n+\alpha)}{\Gamma(n) n^\alpha} = 1. \quad (5.27)$$

## 5.20 Technical Proofs\*

**Proof of Theorem 5.1.** Squaring expression (5.1)

$$\begin{aligned} \left( \int_0^\infty \exp(-x^2/2) dx \right)^2 &= \int_0^\infty \exp(-x^2/2) dx \int_0^\infty \exp(-u^2/2) du \\ &= \int_0^\infty \int_0^\infty \exp(-(x^2+u^2)/2) dx du \\ &= \int_0^\infty \int_0^{\pi/2} r \exp(-r^2/2) d\theta dr \\ &= \frac{\pi}{2} \int_0^\infty r \exp(-r^2/2) dr \\ &= \frac{\pi}{2}. \end{aligned}$$

The third equality is the key. It makes the change-of-variables to polar coordinates  $x = r \cos\theta$  and  $u = r \sin\theta$  so that  $x^2 + u^2 = r^2$ . The Jacobian of this transformation is  $r$ . The region of integration in the  $(x, u)$  units is the positive orthant (upper-right region), which corresponds to integrating  $\theta$  from 0 to  $\pi/2$  in polar coordinates. The final two equalities are simple integration. Taking the square root we obtain (5.1). ■

### Proof of Theorem 5.2.

$$\begin{aligned}\mathbb{E}|X|^r &= \int_{-\infty}^{\infty} |x|^r \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx \\ &= \frac{\sqrt{2}}{\sqrt{\pi}} \int_0^{\infty} x^r \exp(-x^2/2) dx \\ &= \frac{2^{r/2}}{\sqrt{\pi}} \int_0^{\infty} u^{(r-1)/2} \exp(-u) dt \\ &= \frac{2^{r/2}}{\sqrt{\pi}} \Gamma\left(\frac{r+1}{2}\right)\end{aligned}$$

The third equality is the change-of-variables  $u = x^2/2$  and the final is the definition of the gamma function. ■

**Proof of Theorem 5.4.** Let  $M_x(\mathbf{t}) = \exp(\mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})$  be the moment generating function of  $\mathbf{X}$  by Theorem 5.3. Then the MGF of  $\mathbf{Y}$  is

$$\begin{aligned}\mathbb{E}(\exp(\mathbf{s}'\mathbf{Y})) &= \mathbb{E}\exp(\mathbf{s}'(\mathbf{a} + \mathbf{B}\mathbf{X})) \\ &= \exp(\mathbf{s}'\mathbf{a})\mathbb{E}\exp(\mathbf{s}'\mathbf{B}\mathbf{X}) \\ &= \exp(\mathbf{s}'\mathbf{a})M_x(\mathbf{B}'\mathbf{s}) \\ &= \exp(\mathbf{s}'\mathbf{a})\exp\left(\mathbf{s}'\mathbf{B}\boldsymbol{\mu} + \frac{1}{2}\mathbf{s}'\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}'\mathbf{s}\right) \\ &= \exp\left(\mathbf{s}'(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}) + \frac{1}{2}\mathbf{s}'(\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}')\mathbf{s}\right)\end{aligned}$$

which is the MGF of  $\mathbf{N}(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}')$ . Thus  $\mathbf{Y} \sim \mathbf{N}(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}')$  as claimed. ■

**Proof of Theorem 5.5.** Let  $k_1$  and  $k_2$  denote the dimensions of  $\mathbf{X}_1$  and  $\mathbf{X}_2$  and set  $k = k_1 + k_2$ . If the components are uncorrelated then the covariance matrix for  $\mathbf{X}$  takes the form

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & 0 \\ 0 & \boldsymbol{\Sigma}_2 \end{bmatrix}.$$

In this case the joint density function of  $\mathbf{X}$  equals

$$\begin{aligned}f(\mathbf{x}_1, \mathbf{x}_2) &= \frac{1}{(2\pi)^{k/2} (\det(\boldsymbol{\Sigma}_1) \det(\boldsymbol{\Sigma}_2))^{1/2}} \\ &\cdot \exp\left(-\frac{(\mathbf{x}_1 - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_1^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) + (\mathbf{x}_2 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2)}{2}\right) \\ &= \frac{1}{(2\pi)^{k_1/2} (\det(\boldsymbol{\Sigma}_1))^{1/2}} \exp\left(-\frac{(\mathbf{x}_1 - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_1^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1)}{2}\right) \\ &\cdot \frac{1}{(2\pi)^{k_2/2} (\det(\boldsymbol{\Sigma}_2))^{1/2}} \exp\left(-\frac{(\mathbf{x}_2 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2)}{2}\right).\end{aligned}$$

This is the product of two multivariate normal densities in  $\mathbf{x}_1$  and  $\mathbf{x}_2$ . Joint densities factor if (and only if) the components are independent. This shows that uncorrelatedness implies independence.

The converse (that independence implies uncorrelatedness) holds generally. ■

**Proof of Theorem 5.6.** We demonstrate that  $Q = \mathbf{X}'\mathbf{X}$  has density function (5.2) by verifying that both have the same moment generating function (MGF). First, the MGF of  $\mathbf{X}'\mathbf{X}$  is

$$\begin{aligned}\mathbb{E}(\exp(t\mathbf{X}'\mathbf{X})) &= \int_{\mathbb{R}^r} \exp(t\mathbf{x}'\mathbf{x}) \frac{1}{(2\pi)^{r/2}} \exp\left(-\frac{\mathbf{x}'\mathbf{x}}{2}\right) d\mathbf{x} \\ &= \int_{\mathbb{R}^r} \frac{1}{(2\pi)^{r/2}} \exp\left(-\frac{\mathbf{x}'\mathbf{x}}{2}(1-2t)\right) d\mathbf{x} \\ &= (1-2t)^{-r/2} \int_{\mathbb{R}^r} \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{\mathbf{u}'\mathbf{u}}{2}\right) d\mathbf{u} \\ &= (1-2t)^{-r/2}. \end{aligned} \tag{5.28}$$

The fourth equality uses the change of variables  $\mathbf{u} = (1-2t)^{1/2} \mathbf{x}$  and the final equality is the normal probability integral. Second, the MGF of the density (5.2) is

$$\begin{aligned}\int_0^\infty \exp(tq) f(q) dq &= \int_0^\infty \exp(tq) \frac{1}{\Gamma(\frac{r}{2}) 2^{r/2}} q^{r/2-1} \exp(-q/2) dq \\ &= \int_0^\infty \frac{1}{\Gamma(\frac{r}{2}) 2^{r/2}} q^{r/2-1} \exp(-q(1/2-t)) dq \\ &= \frac{1}{\Gamma(\frac{r}{2}) 2^{r/2}} (1/2-t)^{-r/2} \Gamma\left(\frac{r}{2}\right) \\ &= (1-2t)^{-r/2}, \end{aligned} \tag{5.29}$$

the third equality using the gamma integral (5.26). The MGFs (5.28) and (5.29) are equal, verifying that (5.2) is the density of  $Q$  as claimed. ■

**Proof of Theorem 5.7.** Using the simple law of iterated expectations (2.1),  $T$  has density

$$\begin{aligned}f(x) &= \frac{d}{dx} \mathbb{P}\left(\frac{Z}{\sqrt{Q/r}} \leq x\right) \\ &= \frac{d}{dx} \mathbb{E}\left\{Z \leq x\sqrt{\frac{Q}{r}}\right\} \\ &= \frac{d}{dx} \mathbb{E}\left[\mathbb{P}\left(Z \leq x\sqrt{\frac{Q}{r}} \mid Q\right)\right] \\ &= \mathbb{E}\frac{d}{dx} \Phi\left(x\sqrt{\frac{Q}{r}}\right) \\ &= \mathbb{E}\left(\phi\left(x\sqrt{\frac{Q}{r}}\right)\sqrt{\frac{Q}{r}}\right) \\ &= \int_0^\infty \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{qx^2}{2r}\right)\right) \sqrt{\frac{q}{r}} \left(\frac{1}{\Gamma(\frac{r}{2}) 2^{r/2}} q^{r/2-1} \exp(-q/2)\right) dq \\ &= \frac{\Gamma(\frac{r+1}{2})}{\sqrt{r\pi}\Gamma(\frac{r}{2})} \left(1 + \frac{x^2}{r}\right)^{-(\frac{r+1}{2})} \end{aligned}$$

using the gamma integral (5.26). ■

**Proof of Theorem 5.8.** Notice that for large  $r$ , by the properties of the logarithm

$$\log\left(\left(1 + \frac{x^2}{r}\right)^{-(\frac{r+1}{2})}\right) = -\left(\frac{r+1}{2}\right) \log\left(1 + \frac{x^2}{r}\right) \approx -\left(\frac{r+1}{2}\right) \frac{x^2}{r} \rightarrow -\frac{x^2}{2},$$

the limit as  $r \rightarrow \infty$ , and thus

$$\lim_{r \rightarrow \infty} \left(1 + \frac{x^2}{r}\right)^{-\left(\frac{r+1}{2}\right)} = \exp\left(-\frac{x^2}{2}\right). \quad (5.30)$$

Using a property of the gamma function (5.27)

$$\lim_{n \rightarrow \infty} \frac{\Gamma(n + \alpha)}{\Gamma(n) n^\alpha} = 1$$

with  $n = r/2$  and  $\alpha = 1/2$  we find

$$\lim_{r \rightarrow \infty} \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)} \left(1 + \frac{x^2}{r}\right)^{-\left(\frac{r+1}{2}\right)} = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) = \phi(x).$$

■

**Proof of Theorem 5.9.** Let  $U \sim \chi_m^2$  and  $V \sim \chi_r^2$  be independent and set  $S = U/V$ . Let  $f_m(u)$  be the  $\chi_m^2$  density. By a similar argument as in the proof of Theorem 5.7,  $S$  has the density function

$$\begin{aligned} f_S(s) &= \mathbb{E}(f_m(sV)V) \\ &= \int_0^\infty f_m(sv) v f_r(v) d v \\ &= \frac{1}{2^{(m+r)/2} \Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{r}{2}\right)} \int_0^\infty (sv)^{m/2-1} e^{-sv/2} v^{r/2} e^{-v/2} d v \\ &= \frac{s^{m/2-1}}{2^{(m+r)/2} \Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{r}{2}\right)} \int_0^\infty v^{(m+r)/2-1} e^{-(s+1)v/2} d v \\ &= \frac{s^{m/2-1}}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{r}{2}\right) (1+s)^{(m+r)/2}} \int_0^\infty t^{(m+r)/2-1} e^{-t} d t \\ &= \frac{s^{m/2-1} \Gamma\left(\frac{m+r}{2}\right)}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{r}{2}\right) (1+s)^{(m+r)/2}}. \end{aligned}$$

The fifth equality make the change-of variables  $v = 2t/(1+s)$ , and the sixth uses the definition of the Gamma function (5.24). Making the change-of-variables  $x = sr/m$ , we obtain the density as stated. ■

**Proof of Theorem 5.10.** Applying change-of-variables to the density in Theorem 5.9, the density of  $mF$  is

$$\frac{x^{m/2-1} \Gamma\left(\frac{m+r}{2}\right)}{r^{m/2} \Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{r}{2}\right) (1+\frac{x}{r})^{(m+r)/2}}. \quad (5.31)$$

Using (5.27) with  $n = r/2$  and  $\alpha = m/2$  we have

$$\lim_{r \rightarrow \infty} \frac{\Gamma\left(\frac{m+r}{2}\right)}{r^{m/2} \Gamma\left(\frac{r}{2}\right)} = 2^{-m/2}$$

and similarly to (5.30) we have

$$\lim_{r \rightarrow \infty} \left(1 + \frac{x}{r}\right)^{-\left(\frac{m+r}{2}\right)} = \exp\left(-\frac{x}{2}\right).$$

Together, (5.31) tends to

$$\frac{x^{m/2-1} \exp\left(-\frac{x}{2}\right)}{2^{m/2} \Gamma\left(\frac{m}{2}\right)}$$

which is the  $\chi_m^2$  density. ■

**Proof of Theorem 5.11.** As in the proof of Theorem 5.6 we verify that the MGF of  $Q = \mathbf{X}'\mathbf{X}$  when  $\mathbf{X} \sim N(\boldsymbol{\mu}, \mathbf{I}_r)$  is equal to the MGF of the density function (5.3).

First, we calculate the MGF of  $Q = \mathbf{X}'\mathbf{X}$  when  $\mathbf{X} \sim N(\boldsymbol{\mu}, \mathbf{I}_r)$ . Construct an orthogonal  $r \times r$  matrix  $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2]$  whose first column equals  $\mathbf{h}_1 = \boldsymbol{\mu}(\boldsymbol{\mu}'\boldsymbol{\mu})^{-1/2}$ . Note that  $\mathbf{h}_1'\boldsymbol{\mu} = \lambda^{1/2}$  and  $\mathbf{h}_2'\boldsymbol{\mu} = \mathbf{0}$ . Define  $\mathbf{Z} = \mathbf{H}'\mathbf{X} \sim N(\boldsymbol{\mu}^*, \mathbf{I}_q)$  where

$$\boldsymbol{\mu}^* = \mathbf{H}'\boldsymbol{\mu} = \begin{pmatrix} \mathbf{h}_1'\boldsymbol{\mu} \\ \mathbf{h}_2'\boldsymbol{\mu} \end{pmatrix} = \begin{pmatrix} \lambda^{1/2} \\ \mathbf{0} \end{pmatrix} \frac{1}{r-1}.$$

It follows that  $Q = \mathbf{X}'\mathbf{X} = \mathbf{Z}'\mathbf{Z} = Z_1^2 + Z_2'Z_2$  where  $Z_1 \sim N(\lambda^{1/2}, 1)$  and  $Z_2 \sim N(0, \mathbf{I}_{r-1})$  are independent. Notice that  $Z_2'Z_2 \sim \chi_{r-1}^2$  so has MGF  $(1-2t)^{-(r-1)/2}$  by (5.29). The MGF of  $Z_1^2$  is

$$\begin{aligned} \mathbb{E}(\exp(tZ_1^2)) &= \int_{-\infty}^{\infty} \exp(tx^2) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x - \sqrt{\lambda})^2\right) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x^2(1-2t) - 2x\sqrt{\lambda} + \lambda)\right) dx \\ &= (1-2t)^{-1/2} \exp\left(-\frac{\lambda}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(u^2 - 2u\sqrt{\frac{\lambda}{1-2t}}\right)\right) du \\ &= (1-2t)^{-1/2} \exp\left(-\frac{\lambda t}{1-2t}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(u - \sqrt{\frac{\lambda}{1-2t}}\right)^2\right) du \\ &= (1-2t)^{-1/2} \exp\left(-\frac{\lambda t}{1-2t}\right) \end{aligned}$$

where the third equality uses the change of variables  $u = (1-2t)^{1/2}x$ . Thus the MGF of  $Q = Z_1^2 + Z_2'Z_2$  is

$$\begin{aligned} \mathbb{E}(\exp(tQ)) &= \mathbb{E}(\exp(t(Z_1^2 + Z_2'Z_2))) \\ &= \mathbb{E}(\exp(tZ_1^2))\mathbb{E}(\exp(tZ_2'Z_2)) \\ &= (1-2t)^{-r/2} \exp\left(-\frac{\lambda t}{1-2t}\right). \end{aligned} \tag{5.32}$$

Second, we calculate the MGF of (5.3). It equals

$$\begin{aligned} &\int_0^\infty \exp(tx) \sum_{i=0}^\infty \frac{e^{-\lambda/2}}{i!} \left(\frac{\lambda}{2}\right)^i f_{r+2i}(x) dx \\ &= \sum_{i=0}^\infty \frac{e^{-\lambda/2}}{i!} \left(\frac{\lambda}{2}\right)^i \int_0^\infty \exp(tx) f_{r+2i}(x) dx \\ &= \sum_{i=0}^\infty \frac{e^{-\lambda/2}}{i!} \left(\frac{\lambda}{2}\right)^i (1-2t)^{-(r+2i)/2} \\ &= e^{-\lambda/2} (1-2t)^{-r/2} \sum_{i=0}^\infty \frac{1}{i!} \left(\frac{\lambda}{2(1-2t)}\right)^i \\ &= e^{-\lambda/2} (1-2t)^{-r/2} \exp\left(\frac{\lambda}{2(1-2t)}\right) \\ &= (1-2t)^{-r/2} \exp\left(\frac{\lambda t}{1-2t}\right) \end{aligned} \tag{5.33}$$

where the second equality uses (5.29), and the fourth uses  $\exp(x) = \sum_{i=0}^\infty \frac{a^i}{i!}$ . We can see that (5.32) equals (5.33), verifying that (5.3) is the density of  $Q$  as stated. ■

**Proof of Theorem 5.12.** The fact that  $\mathbf{A} > 0$  means that we can write  $\mathbf{A} = \mathbf{C}\mathbf{C}'$  where  $\mathbf{C}$  is non-singular (see Section A.10). Then  $\mathbf{A}^{-1} = \mathbf{C}^{-1}\mathbf{C}^{-1}'$  and by Theorem 5.4

$$\mathbf{C}^{-1}\mathbf{X} \sim N(\mathbf{C}^{-1}\boldsymbol{\mu}, \mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-1}') = N(\mathbf{C}^{-1}\boldsymbol{\mu}, \mathbf{C}^{-1}\mathbf{C}\mathbf{C}'\mathbf{C}^{-1}') = N(\boldsymbol{\mu}^*, \mathbf{I}_r)$$

where  $\boldsymbol{\mu}^* = \mathbf{C}^{-1}\boldsymbol{\mu}$ . Thus by the definition of the non-central chi-square

$$\mathbf{X}'\mathbf{A}^{-1}\mathbf{X} = \mathbf{X}'\mathbf{C}^{-1/2}\mathbf{C}^{-1}\mathbf{X} = (\mathbf{C}^{-1}\mathbf{X})'(\mathbf{C}^{-1}\mathbf{X}) \sim \chi_r^2(\boldsymbol{\mu}^{*\prime}\boldsymbol{\mu}^*).$$

Since

$$\boldsymbol{\mu}^{*\prime}\boldsymbol{\mu}^* = \boldsymbol{\mu}'\mathbf{C}^{-1/2}\mathbf{C}^{-1}\boldsymbol{\mu} = \boldsymbol{\mu}'\mathbf{A}^{-1}\boldsymbol{\mu} = \lambda,$$

this equals  $\chi_r^2(\lambda)$  as claimed. ■

**Proof of Theorem 5.23.** Since  $\log(u)$  is concave we apply Jensen's inequality (B.26), take expectations with respect to the true density  $f(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}_0)$ , and note that the density  $f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta})$  integrates to 1 for any  $\boldsymbol{\theta} \in \Theta$ , to find that

$$\begin{aligned} \mathbb{E}\left(\log \frac{L(\boldsymbol{\theta})}{L(\boldsymbol{\theta}_0)} | \mathbf{X}\right) &\leq \log \mathbb{E}\left(\frac{L(\boldsymbol{\theta})}{L(\boldsymbol{\theta}_0)} | \mathbf{X}\right) \\ &= \log \int \cdots \int \left( \frac{\prod_{i=1}^n f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta})}{\prod_{i=1}^n f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta}_0)} \right) \prod_{i=1}^n f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta}_0) d\mathbf{y}_1 \cdots d\mathbf{y}_n \\ &= \log \int \cdots \int \prod_{i=1}^n f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta}) d\mathbf{y}_1 \cdots d\mathbf{y}_n \\ &= \log 1 \\ &= 0. \end{aligned}$$

This implies for any  $\boldsymbol{\theta} \in \Theta$ ,  $\mathbb{E}(\log L(\boldsymbol{\theta})) \leq \mathbb{E}(\log L(\boldsymbol{\theta}_0))$ . Hence  $\boldsymbol{\theta}_0$  maximizes  $\mathbb{E}(\log L(\boldsymbol{\theta}))$  as claimed. ■

**Proof of Theorem 5.24.** For part 1, since the support of  $\mathbf{y}$  does not depend on  $\boldsymbol{\theta}$  we can exchange integration and differentiation:

$$\mathbb{E}\left(\left.\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} | \mathbf{X}\right) = \frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E}\left(\left.\log L(\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} | \mathbf{X}\right).$$

Theorem 5.23 showed that  $\mathbb{E}(\log L(\boldsymbol{\theta}))$  is maximized at  $\boldsymbol{\theta}_0$ , which has the first-order condition

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E}\left(\left.\log L(\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} | \mathbf{X}\right) = 0$$

as needed.

For part 2, using part 1 and the fact the observations are independent

$$\begin{aligned} \mathcal{J} &= \text{var}\left(\left.\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}_0)\right| \mathbf{X}\right) \\ &= \mathbb{E}\left(\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}_0)\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}_0)\right)' | \mathbf{X}\right) \\ &= \sum_{i=1}^n \mathbb{E}\left(\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta}_0)\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta}_0)\right)' | \mathbf{x}_i\right) \end{aligned}$$

which is the first equality.

For the second, observe that

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}) = \frac{\frac{\partial}{\partial \boldsymbol{\theta}} f(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})}{f(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})}$$

and

$$\begin{aligned}\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}) &= \frac{\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})}{f(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})} - \frac{\frac{\partial}{\partial \boldsymbol{\theta}} f(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}} f(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})'}{f(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})^2} \\ &= \frac{\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})}{f(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})} - \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})'.\end{aligned}$$

It follows that

$$\begin{aligned}\mathcal{J} &= \sum_{i=1}^n \mathbb{E} \left( \left( \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta}_0) \right) \left( \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta}_0) \right)' \mid \mathbf{x}_i \right) \\ &= - \sum_{i=1}^n \mathbb{E} \left( \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta}_0) \mid \mathbf{x}_i \right) + \sum_{i=1}^n \mathbb{E} \left( \frac{\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta}_0)}{f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta}_0)} \mid \mathbf{x}_i \right).\end{aligned}$$

However, by exchanging integration and differentiation we can check that the second term is zero:

$$\begin{aligned}\mathbb{E} \left( \frac{\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta}_0)}{f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta}_0)} \mid \mathbf{x}_i \right) &= \int \left( \frac{\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(\mathbf{y} | \mathbf{x}_i, \boldsymbol{\theta}_0) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}}{f(\mathbf{y} | \mathbf{x}_i, \boldsymbol{\theta}_0)} \right) f(\mathbf{y} | \boldsymbol{\theta}_0) d\mathbf{y} \\ &= \int \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(\mathbf{y} | \mathbf{x}_i, \boldsymbol{\theta}_0) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} d\mathbf{y} \\ &= \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \int f(\mathbf{y} | \mathbf{x}_i, \boldsymbol{\theta}_0) d\mathbf{y} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\ &= \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} 1 \\ &= 0\end{aligned}$$

This establishes the second inequality. ■

**Proof of Theorem 5.25** Let  $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$  be the sample, let  $f(\mathbf{Y}, \boldsymbol{\theta}) = \prod_{i=1}^n f(\mathbf{y}_i, \boldsymbol{\theta})$  denote the joint density of the sample, and note  $\log L(\boldsymbol{\theta}) = \log f(\mathbf{Y}, \boldsymbol{\theta})$ . Set

$$\mathbf{S} = \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}_0)$$

which by Theorem (5.24) has mean zero and variance  $\mathcal{J}$  conditional on  $\mathbf{X}$ . Write the estimator  $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\mathbf{Y})$  as a function of the data. Since  $\tilde{\boldsymbol{\theta}}$  is unbiased, for any  $\boldsymbol{\theta}$ ,

$$\boldsymbol{\theta} = \mathbb{E}(\tilde{\boldsymbol{\theta}} | \mathbf{X}) = \int \tilde{\boldsymbol{\theta}}(\mathbf{Y}) f(\mathbf{Y}, \boldsymbol{\theta}) d\mathbf{Y}.$$

Differentiating with respect to  $\boldsymbol{\theta}$

$$\begin{aligned}\mathbf{I}_n &= \int \tilde{\boldsymbol{\theta}}(\mathbf{Y}) \frac{\partial}{\partial \boldsymbol{\theta}'} f(\mathbf{Y}, \boldsymbol{\theta}) d\mathbf{Y} \\ &= \int \tilde{\boldsymbol{\theta}}(\mathbf{Y}) \frac{\partial}{\partial \boldsymbol{\theta}'} \log f(\mathbf{Y}, \boldsymbol{\theta}) f(\mathbf{Y}, \boldsymbol{\theta}) d\mathbf{Y}.\end{aligned}$$

Evaluating at  $\boldsymbol{\theta}_0$  yields

$$\mathbf{I}_n = \mathbb{E}(\tilde{\boldsymbol{\theta}} \mathbf{S}' | \mathbf{X}) = \mathbb{E}((\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \mathbf{S}' | \mathbf{X}) \quad (5.34)$$

the second equality since  $\mathbb{E}(\mathbf{S} | \mathbf{X}) = 0$ .

By the matrix Cauchy-Schwarz inequality (B.32), (5.34), and  $\text{var}(\mathbf{S} | \mathbf{X}) = \mathbb{E}(\mathbf{SS}' | \mathbf{X}) = \mathcal{J}$ ,

$$\begin{aligned}\text{var}(\tilde{\boldsymbol{\theta}} | \mathbf{X}) &= \mathbb{E}\left((\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' | \mathbf{X}\right) \\ &\geq \mathbb{E}((\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\mathbf{S}' | \mathbf{X})(\mathbb{E}(\mathbf{SS}' | \mathbf{X}))^{-1}\mathbb{E}(\mathbf{S}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' | \mathbf{X}) \\ &= (\mathbb{E}(\mathbf{SS}' | \mathbf{X}))^{-1} \\ &= \mathcal{J}^{-1}\end{aligned}$$

as stated. ■

## Exercises

**Exercise 5.1** For the standard normal density  $\phi(x)$ , show that  $\phi'(x) = -x\phi(x)$ .

**Exercise 5.2** Use the result in Exercise 5.1 and integration by parts to show that for  $X \sim N(0, 1)$ ,  $\mathbb{E}X^2 = 1$ .

**Exercise 5.3** Use the results in Exercises 5.1 and 5.2, plus integration by parts, to show that for  $X \sim N(0, 1)$ ,  $\mathbb{E}X^4 = 3$ .

**Exercise 5.4** Show that the moment generating function (mgf) of  $X \sim N(0, 1)$  is  $m(t) = \mathbb{E}(\exp(tX)) = \exp(t^2/2)$ . (For the definition of the mgf see Section 2.32).

**Exercise 5.5** Show that the moment generating function (mgf) of  $X \sim N(\mu, \sigma^2)$  is  $m(t) = \mathbb{E}(\exp(tX)) = \exp(t\mu + t^2\sigma^2/2)$ .

Hint: Write  $X = \mu + \sigma Z$  where  $Z \sim N(0, 1)$ .

**Exercise 5.6** Use the mgf from Exercise 5.4 to verify that for  $X \sim N(0, 1)$ ,  $\mathbb{E}(X^2) = m''(0) = 1$  and  $\mathbb{E}(X^4) = m^{(4)}(0) = 3$ .

**Exercise 5.7** Write the multivariate  $N(\mathbf{0}, \mathbf{I}_k)$  density as the product of  $N(0, 1)$  density functions. That is, show that

$$\frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{\mathbf{x}'\mathbf{x}}{2}\right) = \phi(x_1) \cdots \phi(x_k).$$

**Exercise 5.8** Show that the mgf of  $X \sim N(\mathbf{0}, \mathbf{I}_k)$  is  $\mathbb{E}(\exp(t'X)) = \exp(\frac{1}{2}t't)$ .

Hint: Use Exercise 5.4 and the fact that the elements of  $X$  are independent.

**Exercise 5.9** Show that the mgf of  $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  is

$$M(\mathbf{t}) = \mathbb{E}(\exp(t'X)) = \exp\left(\mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}\right).$$

Hint: Write  $X = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{Z}$  where  $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_k)$ .

**Exercise 5.10** Show that the characteristic function of  $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  is

$$C(\mathbf{t}) = \mathbb{E}(\exp(it'X)) = \exp\left(i\boldsymbol{\mu}'\boldsymbol{\lambda} - \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}\right).$$

For the definition of the characteristic function see Section 2.32.

Hint: For  $X \sim N(0, 1)$ , establish  $\mathbb{E}(\exp(itX)) = \exp(-\frac{1}{2}t^2)$  by integration. Then generalize to  $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  using the same steps as in Exercises 5.8 and 5.9.

**Exercise 5.11** Show that if  $Q \sim \chi_r^2$ , then  $\mathbb{E}(Q) = r$  and  $\text{var}(Q) = 2r$ .

Hint: Use the representation  $Q = \sum_{i=1}^n X_i^2$  with  $X_i$  independent  $N(0, 1)$ .

**Exercise 5.12** Show that if  $Q \sim \chi_k^2(\lambda)$ , then  $\mathbb{E}(Q) = k + \lambda$ .

**Exercise 5.13** Suppose  $X_i$  are independent  $N(\mu_i, \sigma_i^2)$ . Find the distribution of the weighted sum  $\sum_{i=1}^n w_i X_i$ .

**Exercise 5.14** Show that if  $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}_n \sigma^2)$  and  $\mathbf{H}'\mathbf{H} = \mathbf{I}_n$  then  $\mathbf{u} = \mathbf{H}'\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}_n \sigma^2)$ .

**Exercise 5.15** Show that if  $\mathbf{e} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$  and  $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}'$  then  $\mathbf{u} = \mathbf{A}^{-1}\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}_n)$ .

**Exercise 5.16** Show that  $\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \log L(\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta})$ .

**Exercise 5.17** For the regression in-sample predicted values  $\hat{y}_i$  show that  $\hat{y}_i|_X \sim N(\mathbf{x}'_i \boldsymbol{\beta}, \sigma^2 h_{ii})$  where  $h_{ii}$  are the leverage values (3.41).

**Exercise 5.18** In the normal regression model, show that the leave-one out prediction errors  $\tilde{e}_i$  and the standardized residuals  $\bar{e}_i$  are independent of  $\hat{\boldsymbol{\beta}}$ , conditional on  $X$ .

Hint: Use (3.46) and (4.24).

**Exercise 5.19** In the normal regression model, show that the robust covariance matrices  $\hat{V}_{\hat{\boldsymbol{\beta}}}^{HC0}$ ,  $\hat{V}_{\hat{\boldsymbol{\beta}}}^{HC1}$ ,  $\hat{V}_{\hat{\boldsymbol{\beta}}}^{HC2}$ , and  $\hat{V}_{\hat{\boldsymbol{\beta}}}^{HC3}$  are independent of the OLS estimator  $\hat{\boldsymbol{\beta}}$ , conditional on  $X$ .

**Exercise 5.20** Let  $F(u)$  be the distribution function of a random variable  $X$  whose density is symmetric about zero. (This includes the standard normal and the student  $t$ .) Show that  $F(-u) = 1 - F(u)$ .

**Exercise 5.21** Let  $C_{\beta} = [L, U]$  be a  $1 - \alpha$  confidence interval for  $\beta$ , and consider the transformation  $\theta = g(\beta)$  where  $g(\cdot)$  is monotonically increasing. Consider the confidence interval  $C_{\theta} = [g(L), g(U)]$  for  $\theta$ . Show that  $\mathbb{P}(\theta \in C_{\theta}) = \mathbb{P}(\beta \in C_{\beta})$ . Use this result to develop a confidence interval for  $\sigma$ .

**Exercise 5.22** Show that the test “Reject  $H_0$  if  $LR \geq c_1$ ” for  $LR$  defined in (5.21), and the test “Reject  $H_0$  if  $F \geq c_2$ ” for  $F$  defined in (5.22), yield the same decisions if  $c_2 = (\exp(c_1/n) - 1)(n - k)/q$ . Why does this mean that the two tests are *equivalent*?

**Exercise 5.23** Show (5.23).

**Exercise 5.24** In the normal regression model, let  $s^2$  be the unbiased estimator of the error variance  $\sigma^2$  from (4.26).

(a) Show that  $\text{var}(s^2) = 2\sigma^4/(n - k)$ .

(b) Show that  $\text{var}(s^2)$  is strictly larger than the Cramér-Rao Lower Bound for  $\sigma^2$ .

**Part II**

**Large Sample Methods**

# Chapter 6

## An Introduction to Large Sample Asymptotics

### 6.1 Introduction

For inference (confidence intervals and hypothesis testing) on unknown parameters we need sampling distributions, either exact or approximate, of estimators and other statistics.

In Chapter 4 we derived the mean and variance of the least-squares estimator in the context of the linear regression model, but this is not a complete description of the sampling distribution and is thus not sufficient for inference. Furthermore, the theory does not apply in the context of the linear projection model, which is more relevant for empirical applications.

In Chapter 5 we derived the exact sampling distribution of the OLS estimator, t-statistics, and F-statistics for the normal regression model, allowing for inference. But these results are narrowly confined to the normal regression model, which requires the unrealistic assumption that the regression error is normally distributed and independent of the regressors. Perhaps we can view these results as some sort of approximation to the sampling distributions without requiring the assumption of normality, but how can we be precise about this?

To illustrate the situation with an example, let  $y_i$  and  $x_i$  be drawn from the joint density

$$f(x, y) = \frac{1}{2\pi xy} \exp\left(-\frac{1}{2}(\log y - \log x)^2\right) \exp\left(-\frac{1}{2}(\log x)^2\right)$$

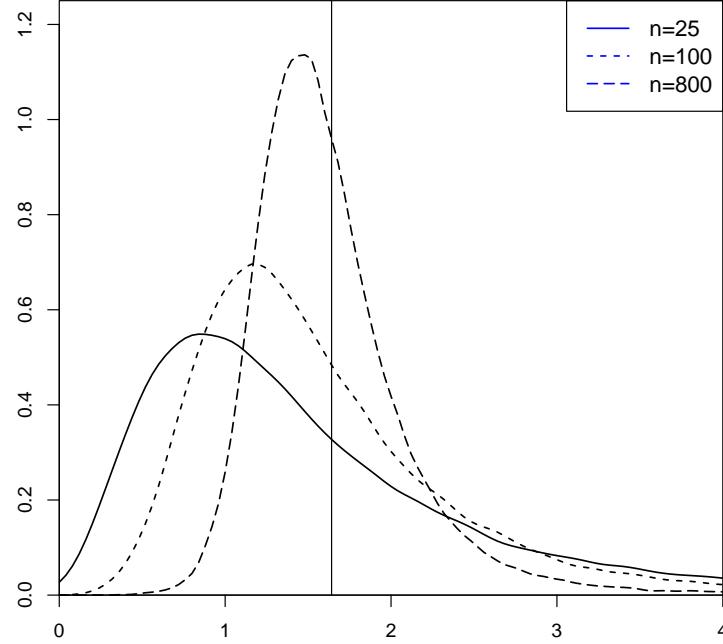
and let  $\hat{\beta}$  be the slope coefficient estimate from a least-squares regression of  $y_i$  on  $x_i$  and a constant. Using simulation methods, the density function of  $\hat{\beta}$  was computed and plotted in Figure 6.1 for sample sizes of  $n = 25$ ,  $n = 100$  and  $n = 800$ . The vertical line marks the true projection coefficient.

From the figure we can see that the density functions are dispersed and highly non-normal. As the sample size increases the density becomes more concentrated about the population coefficient. Is there a simple way to characterize the sampling distribution of  $\hat{\beta}$ ?

In principle the sampling distribution of  $\hat{\beta}$  is a function of the joint distribution of  $(y_i, x_i)$  and the sample size  $n$ , but in practice this function is extremely complicated so it is not feasible to analytically calculate the exact distribution of  $\hat{\beta}$  except in very special cases. Therefore we typically rely on approximation methods.

In this chapter we introduce asymptotic theory, which approximates by taking the limit of the finite sample distribution as the sample size  $n$  tends to infinity. It is important to understand that this is an approximation technique, as the asymptotic distributions are used to assess the finite sample distributions of our estimators in actual practical samples. The primary tools of asymptotic theory are the weak law of large numbers (WLLN), central limit theorem (CLT), and continuous mapping theorem (CMT). With these tools we can approximate the sampling distributions of most econometric estimators.

In this chapter we provide a concise summary. Some of the material is quite advanced, and provided for a complete reference.

Figure 6.1: Sampling Density of  $\hat{\beta}$ 

## 6.2 Asymptotic Limits

“Asymptotic analysis” is a method of approximation obtained by taking a suitable limit. There is more than one method to take limits, but the most common is to take the limit of the sequence of sampling distributions as the sample size tends to positive infinity, written “as  $n \rightarrow \infty$ .” It is not meant to be interpreted literally, but rather as an approximating device.

The first building block for asymptotic analysis is the concept of a limit of a sequence.

**Definition 6.1** A sequence  $a_n$  has the **limit**  $a$ , written  $a_n \rightarrow a$  as  $n \rightarrow \infty$ , or alternatively as  $\lim_{n \rightarrow \infty} a_n = a$ , if for all  $\delta > 0$  there is some  $n_\delta < \infty$  such that for all  $n \geq n_\delta$ ,  $|a_n - a| \leq \delta$ .

In words,  $a_n$  has the limit  $a$  if the sequence gets closer and closer to  $a$  as  $n$  gets larger. If a sequence has a limit, that limit is unique (a sequence cannot have two distinct limits). If  $a_n$  has the limit  $a$ , we also say that  $a_n$  **converges** to  $a$  as  $n \rightarrow \infty$ .

Not all sequences have limits. For example, the sequence  $\{1, 2, 1, 2, 1, 2, \dots\}$  does not have a limit. It is sometimes useful to have a more general definition of limits which always exist, and these are the limit superior and limit inferior of a sequence.

**Definition 6.2**  $\liminf_{n \rightarrow \infty} a_n \stackrel{\text{def}}{=} \lim_{n \rightarrow \infty} \inf_{m \geq n} a_m$

**Definition 6.3**  $\limsup_{n \rightarrow \infty} a_n \stackrel{\text{def}}{=} \lim_{n \rightarrow \infty} \sup_{m \geq n} a_m$

The limit inferior and limit superior always exist (including  $\pm\infty$  as possibilities), and equal when the limit exists. In the example given earlier, the limit inferior of  $\{1, 2, 1, 2, 1, 2, \dots\}$  is 1, and the limit superior is 2.

### 6.3 Convergence in Probability

A sequence of numbers may converge to a limit, but what about a sequence of random variables? For example, consider a sample mean  $\bar{y} = n^{-1} \sum_{i=1}^n y_i$  based on a random sample of  $n$  observations. As  $n$  increases, the distribution of  $\bar{y}$  changes. In what sense can we describe the “limit” of  $\bar{y}$ ? In what sense does it converge?

Since  $\bar{y}$  is a random variable, we cannot directly apply the deterministic concept of a sequence of numbers. Instead, we require a definition of convergence which is appropriate for random variables. There are more than one such definition, but the most commonly used is called convergence in probability.

**Definition 6.4** A random variable  $z_n \in \mathbb{R}$  **converges in probability** to  $z$  as  $n \rightarrow \infty$ , denoted  $z_n \xrightarrow{p} z$ , or alternatively  $\text{plim}_{n \rightarrow \infty} z_n = z$ , if for all  $\delta > 0$ ,

$$\lim_{n \rightarrow \infty} \mathbb{P}(|z_n - z| \leq \delta) = 1. \quad (6.1)$$

We call  $z$  the **probability limit** (or **plim**) of  $z_n$ .

The definition looks quite abstract, but it formalizes the concept of a sequence of random variables concentrating about a point. The event  $\{|z_n - z| \leq \delta\}$  occurs when  $z_n$  is within  $\delta$  of the point  $z$ .  $\mathbb{P}(|z_n - z| \leq \delta)$  is the probability of this event – that  $z_n$  is within  $\delta$  of the point  $z$ . Equation (6.1) states that this probability approaches 1 as the sample size  $n$  increases. The definition of convergence in probability requires that this holds for any  $\delta$ . So for any small interval about  $z$  the distribution of  $z_n$  concentrates within this interval for large  $n$ .

You may notice that the definition concerns the distribution of the random variables  $z_n$ , not their realizations. Furthermore, notice that the definition uses the concept of a conventional (deterministic) limit, but the latter is applied to a sequence of probabilities, not directly to the random variables  $z_n$  or their realizations.

Two comments about the notation are worth mentioning. First, it is conventional to write the convergence symbol as  $\xrightarrow{p}$  where the “ $p$ ” above the arrow indicates that the convergence is “in probability”. You should try and adhere to this notation, and not simply write  $z_n \rightarrow z$ . Second, it is important to include the phrase “as  $n \rightarrow \infty$ ” to be specific about how the limit is obtained.

A common mistake is to confuse convergence in probability with convergence in expectation:

$$\mathbb{E}(z_n) \rightarrow \mathbb{E}(z). \quad (6.2)$$

They are related but distinct concepts. Neither (6.1) nor (6.2) implies the other.

To see the distinction it might be helpful to think through a stylized example. Consider a discrete random variable  $z_n$  which takes the value 0 with probability  $1 - n^{-1}$  and the value  $a_n \neq 0$  with probability  $n^{-1}$ , or

$$\begin{aligned} \mathbb{P}(z_n = 0) &= 1 - \frac{1}{n} \\ \mathbb{P}(z_n = a_n) &= \frac{1}{n}. \end{aligned} \quad (6.3)$$

In this example the probability distribution of  $z_n$  concentrates at zero as  $n$  increases, regardless of the sequence  $a_n$ . You can check that  $z_n \xrightarrow{p} 0$  as  $n \rightarrow \infty$ .

In this example we can also calculate that the expectation of  $z_n$  is

$$\mathbb{E}(z_n) = \frac{a_n}{n}.$$

Despite the fact that  $z_n$  converges in probability to zero, its expectation will not decrease to zero unless  $a_n/n \rightarrow 0$ . If  $a_n$  diverges to infinity at a rate equal to  $n$  (or faster) then  $\mathbb{E}(z_n)$  will not converge to zero. For example, if  $a_n = n$ , then  $\mathbb{E}(z_n) = 1$  for all  $n$ , even though  $z_n \xrightarrow{p} 0$ .

This example might seem a bit artificial, but the point is that the concepts of convergence in probability and convergence in expectation are distinct, so it is important not to confuse one with the other.

Another common source of confusion with the notation surrounding probability limits is that the expression to the right of the arrow “ $\xrightarrow{p}$ ” must be free of dependence on the sample size  $n$ . Thus expressions of the form “ $z_n \xrightarrow{p} c_n$ ” are notationally meaningless and should not be used.

## 6.4 Weak Law of Large Numbers

In large samples we expect parameter estimates to be close to the population values. For example, in Section 4.3 we saw that the sample mean  $\bar{y}$  is unbiased for  $\mu = \mathbb{E}(y)$  and has variance  $\sigma^2/n$ . As  $n$  gets large its variance decreases and thus the distribution of  $\bar{y}$  concentrates about the population mean  $\mu$ . It turns out that this implies that the sample mean converges in probability to the population mean.

When  $y$  has a finite variance there is a fairly straightforward proof by applying Chebyshev's inequality.

**Theorem 6.1 Chebyshev's Inequality.** For any random variable  $z_n$  and constant  $\delta > 0$

$$\mathbb{P}(|z_n - \mathbb{E}(z_n)| \geq \delta) \leq \frac{\text{var}(z_n)}{\delta^2}.$$

Chebyshev's inequality is terrifically important in asymptotic theory. While its proof is a technical exercise in probability theory, it is quite simple so we discuss it forthwith. Let  $F_n(u)$  denote the distribution of  $z_n - \mathbb{E}(z_n)$ . Then

$$\mathbb{P}(|z_n - \mathbb{E}(z_n)| \geq \delta) = \mathbb{P}((z_n - \mathbb{E}(z_n))^2 \geq \delta^2) = \int_{\{u^2 \geq \delta^2\}} dF_n(u).$$

The integral is over the event  $\{u^2 \geq \delta^2\}$ , so that the inequality  $1 \leq \frac{u^2}{\delta^2}$  holds throughout. Thus

$$\int_{\{u^2 \geq \delta^2\}} dF_n(u) \leq \int_{\{u^2 \geq \delta^2\}} \frac{u^2}{\delta^2} dF_n(u) \leq \int \frac{u^2}{\delta^2} dF_n(u) = \frac{\mathbb{E}(z_n - \mathbb{E}(z_n))^2}{\delta^2} = \frac{\text{var}(z_n)}{\delta^2},$$

which establishes the desired inequality.

Applied to the sample mean  $\bar{y}$  which has variance  $\sigma^2/n$ , Chebyshev's inequality shows that for any  $\delta > 0$

$$\mathbb{P}(|\bar{y} - \mathbb{E}(\bar{y})| \geq \delta) \leq \frac{\sigma^2/n}{\delta^2}.$$

For fixed  $\sigma^2$  and  $\delta$ , the bound on the right-hand-side shrinks to zero as  $n \rightarrow \infty$ . (Specifically, for any  $\varepsilon > 0$  set  $n \geq \sigma^2 / (\delta^2 \varepsilon)$ . Then the right-hand-side is less than or equal to  $\varepsilon$ .) Thus the probability that  $\bar{y}$  is within  $\delta$  of  $\mathbb{E}(\bar{y}) = \mu$  approaches 1 as  $n$  gets large, or

$$\lim_{n \rightarrow \infty} \mathbb{P}(|\bar{y} - \mu| < \delta) = 1.$$

This means that  $\bar{y}$  converges in probability to  $\mu$  as  $n \rightarrow \infty$ .

This result is called the **weak law of large numbers**. Our derivation assumed that  $y$  has a finite variance, but with a more careful derivation all that is necessary is a finite mean.

**Theorem 6.2 Weak Law of Large Numbers (WLLN)**

If  $y_i$  are independent and identically distributed and  $\mathbb{E}|y| < \infty$ , then as  $n \rightarrow \infty$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \xrightarrow{p} \mathbb{E}(y).$$

The proof of Theorem 6.2 is presented in Section 6.28. Technically, the assumption that  $y_i$  are identically distributed can be replaced by the assumption that  $y_i$  is uniformly integrable – see Section 6.21 for the definition – but i.i.d. is sufficient for most applications.

The WLLN shows that the estimator  $\bar{y}$  converges in probability to the true population mean  $\mu$ . In general, an estimator which converges in probability to the population value is called **consistent**.

**Definition 6.5** An estimator  $\hat{\theta}$  of a parameter  $\theta$  is **consistent** if  $\hat{\theta} \xrightarrow{p} \theta$  as  $n \rightarrow \infty$ .

**Theorem 6.3** If  $y_i$  are independent and identically distributed and  $\mathbb{E}|y| < \infty$ , then  $\hat{\mu} = \bar{y}$  is consistent for the population mean  $\mu$ .

Consistency is a good property for an estimator to possess. It means that for any given data distribution, there is a sample size  $n$  sufficiently large such that the estimator  $\hat{\theta}$  will be arbitrarily close to the true value  $\theta$  with high probability. The theorem does not tell us, however, how large this  $n$  has to be. Thus the theorem does not give practical guidance for empirical practice. Still, it is a minimal property for an estimator to be considered a “good” estimator, and provides a foundation for more useful approximations.

## 6.5 Almost Sure Convergence and the Strong Law\*

Convergence in probability is sometimes called **weak convergence**. A related concept is **almost sure convergence**, also known as **strong convergence**. (In probability theory the term “almost sure” means “with probability equal to one”. An event which is random but occurs with probability equal to one is said to be **almost sure**.)

**Definition 6.6** A random variable  $z_n \in \mathbb{R}$  **converges almost surely** to  $z$  as  $n \rightarrow \infty$ , denoted  $z_n \xrightarrow{a.s.} z$ , if for every  $\delta > 0$

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} |z_n - z| \leq \delta\right) = 1. \quad (6.4)$$

The convergence (6.4) is stronger than (6.1) because it computes the probability of a limit rather than the limit of a probability. Almost sure convergence is stronger than convergence in probability in the sense that  $z_n \xrightarrow{a.s.} z$  implies  $z_n \xrightarrow{p} z$ .

In the example (6.3) of Section 6.3, the sequence  $z_n$  converges in probability to zero for any sequence  $a_n$ , but this is not sufficient for  $z_n$  to converge almost surely. In order for  $z_n$  to converge to zero almost surely, it is necessary that  $a_n \rightarrow 0$ .

In the random sampling context the sample mean can be shown to converge almost surely to the population mean. This is called the **strong law of large numbers**.

**Theorem 6.4 Strong Law of Large Numbers (SLLN)**

If  $y_i$  are independent and identically distributed and  $\mathbb{E}|y| < \infty$ , then as  $n \rightarrow \infty$ ,

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \xrightarrow{a.s.} \mathbb{E}(y).$$

The proof of the SLLN is technically advanced so is not presented here. For a proof see Billingsley (1995, Theorem 22.1) or Ash (1972, Theorem 7.2.5).

The WLLN is sufficient for most purposes in econometrics, so we will not use the SLLN in this text.

## 6.6 Vector-Valued Moments

Our preceding discussion focused on the case where  $y$  is real-valued (a scalar), but nothing important changes if we generalize to the case where  $\mathbf{y} \in \mathbb{R}^m$  is a vector. To fix notation, the elements of  $\mathbf{y}$  are

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}.$$

The population mean of  $\mathbf{y}$  is just the vector of marginal means

$$\boldsymbol{\mu} = \mathbb{E}(\mathbf{y}) = \begin{pmatrix} \mathbb{E}(y_1) \\ \mathbb{E}(y_2) \\ \vdots \\ \mathbb{E}(y_m) \end{pmatrix}.$$

When working with random vectors  $\mathbf{y}$  it is convenient to measure their magnitude by their Euclidean length or Euclidean norm

$$\|\mathbf{y}\| = (y_1^2 + \cdots + y_m^2)^{1/2}.$$

In vector notation we have

$$\|\mathbf{y}\|^2 = \mathbf{y}'\mathbf{y}.$$

It turns out that it is equivalent to describe finiteness of moments in terms of the Euclidean norm of a vector or all individual components.

**Theorem 6.5** For  $\mathbf{y} \in \mathbb{R}^m$ ,  $\mathbb{E}\|\mathbf{y}\| < \infty$  if and only if  $\mathbb{E}|y_j| < \infty$  for  $j = 1, \dots, m$ .

The  $m \times m$  variance matrix of  $\mathbf{y}$  is

$$\mathbf{V} = \text{var}(\mathbf{y}) = \mathbb{E}((\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})').$$

$\mathbf{V}$  is often called a variance-covariance matrix. You can show that the elements of  $\mathbf{V}$  are finite if  $\mathbb{E}(\|\mathbf{y}\|^2) < \infty$ .

A random sample  $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$  consists of  $n$  observations of independent and identically distributed draws from the distribution of  $\mathbf{y}$ . (Each draw is an  $m$ -vector.) The vector sample mean

$$\bar{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_m \end{pmatrix}$$

is the vector of sample means of the individual variables.

Convergence in probability of a vector can be defined as convergence in probability of all elements in the vector. Thus  $\bar{\mathbf{y}} \xrightarrow{p} \boldsymbol{\mu}$  if and only if  $\bar{y}_j \xrightarrow{p} \mu_j$  for  $j = 1, \dots, m$ . Since the latter holds if  $\mathbb{E}|y_j| < \infty$  for  $j = 1, \dots, m$ , or equivalently  $\mathbb{E}\|\mathbf{y}\| < \infty$ , we can state this formally as follows.

**Theorem 6.6 WLLN for random vectors**

If  $\mathbf{y}_i$  are independent and identically distributed and  $\mathbb{E}\|\mathbf{y}\| < \infty$ , then as  $n \rightarrow \infty$ ,

$$\bar{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i \xrightarrow{p} \mathbb{E}(\mathbf{y}).$$

## 6.7 Convergence in Distribution

The WLLN is a useful first step, but does not give an approximation to the distribution of an estimator. A **large-sample** or **asymptotic** approximation can be obtained using the concept of **convergence in distribution**.

We say that a sequence of random vectors  $\mathbf{z}_n$  converges in distribution if the sequence of distribution functions  $F_n(\mathbf{u}) = \mathbb{P}(\mathbf{z}_n \leq \mathbf{u})$  converges to a limit distribution function.

**Definition 6.7** Let  $\mathbf{z}_n$  be a random vector with distribution  $F_n(\mathbf{u}) = \mathbb{P}(\mathbf{z}_n \leq \mathbf{u})$ .

We say that  $\mathbf{z}_n$  **converges in distribution** to  $\mathbf{z}$  as  $n \rightarrow \infty$ , denoted  $\mathbf{z}_n \xrightarrow{d} \mathbf{z}$ , if for all  $\mathbf{u}$  at which  $F(\mathbf{u}) = \mathbb{P}(\mathbf{z} \leq \mathbf{u})$  is continuous,  $F_n(\mathbf{u}) \rightarrow F(\mathbf{u})$  as  $n \rightarrow \infty$ .

Under these conditions, it is also said that  $F_n$  **converges weakly** to  $F$ . It is common to refer to  $\mathbf{z}$  and its distribution  $F(\mathbf{u})$  as the **asymptotic distribution**, **large sample distribution**, or **limit distribution** of  $\mathbf{z}_n$ .

When the limit distribution  $\mathbf{z}$  is degenerate (that is,  $\mathbb{P}(\mathbf{z} = \mathbf{c}) = 1$  for some  $\mathbf{c}$ ) we can write the convergence as  $\mathbf{z}_n \xrightarrow{d} \mathbf{c}$ , which is equivalent to convergence in probability,  $\mathbf{z}_n \xrightarrow{p} \mathbf{c}$ .

The remainder of this Section is more technical than required for most students, and can be skipped if desired.

Our definition of convergence in distribution is pointwise, in the sense that it is stated for each  $\mathbf{u}$ . It turns out that the convergence is also uniform over  $\mathbf{u}$  when  $F(\mathbf{u})$  is continuous.

**Theorem 6.7** If  $z_n \xrightarrow{d} z$  and  $F(\mathbf{u})$  is continuous then

$$\sup_{-\infty < \mathbf{u} < \infty} |F_n(\mathbf{u}) - F(\mathbf{u})| \longrightarrow 0$$

as  $n \rightarrow \infty$ .

We present a proof in Section 6.28.

Technically, in most cases of interest it is difficult to establish the limit distributions of sample statistics  $\mathbf{z}_n$  by working directly with their distribution function. It turns out that in most cases it is easier to use alternative yet equivalent characterizations. One is the following.

**Theorem 6.8**  $\mathbf{z}_n \xrightarrow{d} z$  if and only if  $\mathbb{E}(f(z_n)) \rightarrow \mathbb{E}(f(z))$  for all bounded, continuous functions  $f$ .

The proof is rather technical so is not presented here. See Van der Vaart (1998) Lemma 2.2.

A further specialization of the above theorem focuses on the characteristic function  $C_n(\lambda) = \mathbb{E}(\exp(i\lambda' \mathbf{z}_n))$ , which is a transformation of the distribution. (See Section 2.32 for the definition.) The characteristic function  $C_n(t)$  completely describes the distribution of  $\mathbf{z}_n$ . It therefore seems reasonable to expect that if  $C_n(t)$  converges to a limit function  $C(t)$ , then the the distribution of  $\mathbf{z}_n$  converges as well. This turns out to be true, and is known as Lévy's continuity theorem.

**Theorem 6.9 Lévy's Continuity Theorem.**  $\mathbf{z}_n \xrightarrow{d} z$  if and only if  $\mathbb{E}(\exp(it' \mathbf{z}_n)) \rightarrow \mathbb{E}(\exp(it' z))$  for every  $t \in \mathbb{R}^k$ .

While this result seems quite intuitive, a rigorous proof is quite advanced and so is not presented here. See Van der Vaart (1998) Theorem 2.13.

Finally, we mention a standard trick which is commonly used to establish multivariate convergence results.

**Theorem 6.10 Cramér-Wold Device.**  $\mathbf{z}_n \xrightarrow{d} z$  if and only if  $\lambda' \mathbf{z}_n \xrightarrow{d} \lambda' z$  for every  $\lambda \in \mathbb{R}^k$  with  $\lambda' \lambda = 1$ .

We present a proof in Section 6.28 which is a simple application of Lévy's continuity theorem.

## 6.8 Central Limit Theorem

We would like to obtain a distributional approximation to the sample mean  $\bar{y}$ . We start under the random sampling assumption so that the observations are independent and identically distributed, and have a finite mean  $\mu = \mathbb{E}(y)$  and variance  $\sigma^2 = \text{var}(y)$ .

Let's start by finding the asymptotic distribution of  $\bar{y}$ , in the sense that  $\bar{y} \xrightarrow{d} z$  for some random variable  $z$ . From the WLLN we know that  $\bar{y} \xrightarrow{P} \mu$ . Since convergence in probability to a constant is the

same as convergence in distribution, this means that  $\bar{y} \xrightarrow{d} \mu$  as well. This is not a useful distributional result as the limit distribution is a constant. To obtain a non-degenerate distribution we need to rescale  $\bar{y}$ . Recall that  $\text{var}(\bar{y} - \mu) = \sigma^2/n$ , which means that  $\text{var}(\sqrt{n}(\bar{y} - \mu)) = \sigma^2$ . This suggests renormalizing the statistic as

$$z_n = \sqrt{n}(\bar{y} - \mu).$$

Notice that  $\mathbb{E}(z_n) = 0$  and  $\text{var}(z_n) = \sigma^2$ . This shows that the mean and variance have been stabilized. We now seek to determine the asymptotic distribution of  $z_n$ .

The answer is provided by the central limit theorem (CLT) which states that standardized sample averages converge in distribution to normal random vectors. There are several versions of the CLT. The most basic is the case where the observations are independent and identically distributed.

**Theorem 6.11 Lindeberg-Lévy Central Limit Theorem.** If  $y_i$  are independent and identically distributed and  $\mathbb{E}(y_i^2) < \infty$ , then as  $n \rightarrow \infty$

$$\sqrt{n}(\bar{y} - \mu) \xrightarrow{d} N(0, \sigma^2)$$

where  $\mu = \mathbb{E}(y)$  and  $\sigma^2 = \mathbb{E}(y_i - \mu)^2$ .

The proof of the CLT is rather technical (so is presented in Section 6.28) but at the core is a quadratic approximation of the log of the characteristic function.

As we discussed above, in finite samples the standardized sum  $z_n = \sqrt{n}(\bar{y}_n - \mu)$  has mean zero and variance  $\sigma^2$ . What the CLT adds is that  $z_n$  is also approximately normally distributed, and that the normal approximation improves as  $n$  increases.

The CLT is one of the most powerful and mysterious results in statistical theory. It shows that the simple process of averaging induces normality. The first version of the CLT (for the number of heads resulting from many tosses of a fair coin) was established by the French mathematician Abraham de Moivre in a private manuscript circulated in 1733. The most general statements are credited to work by the Russian mathematician Aleksandr Lyapunov (1901) and the Finnish mathematician Karl Waldemar Lindeberg. The above statement is known as the classic (or Lindeberg-Lévy) CLT due to contributions by Lindeberg and the French mathematician Paul Pierre Lévy.

The remainder of this Section is more technical than required for most students, and can be skipped if desired.

A more general version which allows heterogeneous distributions was provided by Lindeberg (1922). The following is the most general statement, known as the Lindeberg CLT or the Lindeberg-Feller CLT.

**Theorem 6.12 Lindeberg Central Limit Theorem.** Suppose for each  $n$ ,  $y_{ni}$ ,  $i = 1, \dots, r_n$  are independent but not necessarily identically distributed with means  $\mathbb{E}(y_{ni}) = 0$  and variances  $\sigma_{ni}^2 = \mathbb{E}(y_{ni}^2)$ . Set  $\bar{\sigma}_n^2 = \sum_{i=1}^{r_n} \sigma_{ni}^2$ . If  $\bar{\sigma}_n^2 > 0$  and for all  $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} \frac{1}{\bar{\sigma}_n^2} \sum_{i=1}^{r_n} \mathbb{E}(y_{ni}^2 \mathbf{1}(y_{ni}^2 \geq \varepsilon \bar{\sigma}_n^2)) = 0 \quad (6.5)$$

then as  $n \rightarrow \infty$

$$\frac{\sum_{i=1}^{r_n} y_{ni}}{\bar{\sigma}_n} \xrightarrow{d} N(0, 1).$$

The proof of the Lindeberg CLT is substantially more technical, so we do not present it here. See Billingsley (1995, Theorem 27.2).

The Lindeberg CLT is quite general as it puts minimal conditions on the sequence of means and variances. The key assumption is equation (6.5) which is known as **Lindeberg's Condition**. In its raw form it is difficult to interpret. The intuition for (6.5) is that it excludes any single observation from dominating the asymptotic distribution. Since (6.5) is quite abstract, in many contexts we use more elementary conditions which are simpler to interpret. All of the following assume  $r_n = n$ .

One such alternative is called **Lyapunov's condition**: For some  $\delta > 0$

$$\lim_{n \rightarrow \infty} \frac{1}{\bar{\sigma}_n^{2+\delta}} \sum_{i=1}^n \mathbb{E}(|y_{ni}|^{2+\delta}) = 0. \quad (6.6)$$

Lyapunov's condition implies Lindeberg's condition, and hence the CLT. Indeed, the left-side of (6.5) is bounded by

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{\bar{\sigma}_n^2} \sum_{i=1}^n \mathbb{E}\left(\frac{|y_{ni}|^{2+\delta}}{|y_{ni}|^\delta} \mathbf{1}(y_{ni}^2 \geq \varepsilon \bar{\sigma}_n^2)\right) \\ & \leq \lim_{n \rightarrow \infty} \frac{1}{\varepsilon^{\delta/2} \bar{\sigma}_n^{2+\delta}} \sum_{i=1}^n \mathbb{E}(|y_{ni}|^{2+\delta}) \\ & = 0 \end{aligned}$$

by (6.6).

Lyapunov's condition is still awkward to interpret. A still simpler condition is a uniform moment bound: For some  $\delta > 0$

$$\sup_{n,i} \mathbb{E} |y_{ni}|^{2+\delta} < \infty. \quad (6.7)$$

This is typically combined with the lower variance bound

$$\liminf_{n \rightarrow \infty} \frac{\bar{\sigma}_n^2}{n} > 0. \quad (6.8)$$

These bounds together imply Lyapunov's condition. To see this, (6.7) and (6.8) imply there is some  $C < \infty$  such that  $\sup_{n,i} \mathbb{E} |y_{ni}|^{2+\delta} \leq C$  and  $\liminf_{n \rightarrow \infty} n^{-1} \bar{\sigma}_n^2 \geq C^{-1}$ . Without loss of generality assume  $\mu_{ni} = 0$ . Then the left side of (6.6) is bounded by

$$\lim_{n \rightarrow \infty} \frac{C^{2+\delta/2}}{n^{\delta/2}} = 0,$$

so Lyapunov's condition holds and hence the CLT.

An alternative to (6.8) is to assume that the average variance converges to a constant, that is,

$$\frac{\bar{\sigma}_n^2}{n} = n^{-1} \sum_{i=1}^n \sigma_{ni}^2 \rightarrow \sigma^2 < \infty. \quad (6.9)$$

This assumption is reasonable in many applications.

We now state the simplest and most commonly used version of a heterogeneous CLT based on the Lindeberg CLT.

**Theorem 6.13** Suppose  $y_{ni}$  are independent but not necessarily identically distributed. If (6.7) and (6.9) hold, then as  $n \rightarrow \infty$

$$\sqrt{n}(\bar{y} - \mathbb{E}(\bar{y})) \xrightarrow{d} N(0, \sigma^2).$$

For a proof see Section 6.28.

One advantage of Theorem 6.13 is that it allows  $\sigma^2 = 0$  (unlike Theorem 6.12).

## 6.9 Higher Moments

As we discussed at the beginning of the previous section, the normalized sample mean  $z_n = \sqrt{n}(\bar{y} - \mu)$  has mean  $\mathbb{E}(z_n) = 0$  and second moment  $\mathbb{E}(z_n^2) = \sigma^2$  which are the same as those of  $Z \sim N(0, \sigma^2)$ . In this section we extend this analysis to higher moments. We find expressions for the finite sample third through sixth moments of  $z_n$ , and show that they converge to those of  $Z$  as  $n$  diverges. This can provide some intuition for the CLT, and can be useful for some other purposes as well. For these results we assume that any stated moment exists. Define the central moments  $\mu_r = \mathbb{E}(y_i - \mu)^r$ .

For simplicity and without loss of generality assume  $\mu = 0$ . The third moment of  $z_n$  is

$$\mathbb{E}(z_n^3) = \frac{1}{n^{3/2}} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \mathbb{E}(y_i y_j y_k).$$

Note that

$$\mathbb{E}(y_i y_j y_k) = \begin{cases} \mathbb{E}(y_i^3) = \mu_3 & \text{if } i = j = k, (\text{n instances}) \\ 0 & \text{otherwise.} \end{cases}$$

Thus

$$\mathbb{E}(z_n^3) = \frac{\mu_3}{n^{1/2}}. \quad (6.10)$$

This shows that the third moment of the normalized sample mean  $z_n$  is a scale of the third central moment of the observations. If  $y_i$  is skewed, then  $z_n$  will have skew in the same direction. However, the third moment of  $z_n$  is proportion to  $n^{-1/2}$ , so converges to zero as  $n \rightarrow \infty$ . This means that the skewness in the distribution of  $z_n$  diminishes with  $n$ .

The fourth moment of  $z_n$  (again assuming  $\mu = 0$ ) is

$$\mathbb{E}(z_n^4) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{\ell=1}^n \mathbb{E}(y_i y_j y_k y_\ell).$$

Note that

$$\mathbb{E}(y_i y_j y_k y_\ell) = \begin{cases} \mathbb{E}(y_i^4) = \mu_4 & \text{if } i = j = k = \ell, (\text{n instances}) \\ \mathbb{E}(y_i^2) \mathbb{E}(y_k^2) = \sigma^4 & \text{if } i = j \neq k = \ell, (\text{n(n-1) instances}) \\ \mathbb{E}(y_i^2) \mathbb{E}(y_j^2) = \sigma^4 & \text{if } i = k \neq j = \ell, (\text{n(n-1) instances}) \\ \mathbb{E}(y_i^2) \mathbb{E}(y_j^2) = \sigma^4 & \text{if } i = \ell \neq j = k, (\text{n(n-1) instances}) \\ 0 & \text{otherwise.} \end{cases}$$

Thus

$$\mathbb{E}(z_n^4) = \frac{\mu_4}{n} + 3\sigma^4 \left( \frac{n-1}{n} \right) = 3\sigma^4 + \frac{\kappa_4}{n} \quad (6.11)$$

where  $\kappa_4 = \mu_4 - 3\sigma^4$  is the 4<sup>th</sup> cumulant of the distribution of  $y_i$  (see Section 2.33 for the definition of the cumulants). Recall that the fourth central moment of  $Z \sim N(0, \sigma^2)$  is  $3\sigma^4$ . Thus the fourth moment of  $z_n$  is close to that of the normal distribution, with a deviation depending on the fourth cumulant of  $y_i$ . The deviation diminishes as  $n$  increases.

For higher moments we can make similar direct yet tedious calculations. A simpler though less intuitive method calculates the moments of  $z_n$  using the cumulant generating function  $K(t) = \log(M(t))$  where  $M(t)$  is the moment generating function of  $y_i$  (see Section 2.33). Since the observations are independent, the cumulant generating function of  $S_n = \sum_{i=1}^n y_i$  is  $\log(M(t)^n) = nK(t)$ . It follows that the  $r^{th}$  cumulant of  $S_n$  is  $nK^{(r)}(0) = n\kappa_r$  where  $\kappa_r = K^{(r)}(0)$  is the  $r^{th}$  cumulant of  $y_i$ . Rescaling, we find that the  $r^{th}$  cumulant of  $z_n = \sqrt{n}(\bar{y} - \mu)$  is  $\kappa_r/n^{r/2-1}$ . Using the relations between central moments and cumulants described in Section 2.33, we deduce that the 3<sup>rd</sup> through 6<sup>th</sup> moments of  $z_n$  are

$$\mathbb{E}(z_n^3) = \kappa_3/n^{1/2} \quad (6.12)$$

$$\mathbb{E}(z_n^4) = \kappa_4/n + 3\kappa_2^2$$

$$\mathbb{E}(z_n^5) = \kappa_5/n^{3/2} - 10\kappa_3\kappa_2/n^{1/2}$$

$$\mathbb{E}(z_n^6) = \kappa_6/n^2 + (15\kappa_4\kappa_2 + 10\kappa_3^2)/n + 15\kappa_2^3. \quad (6.13)$$

Since  $\kappa_2 = \sigma^2$  and  $\mu_3 = \kappa_3$ , the first two expressions are identical to (6.10) and (6.11). The last two also give the exact fifth and sixth moments of  $z_n$ , expressed in terms of the cumulants of  $y_i$  and the sample size  $n$ .

This technique can be used to calculate any non-negative integer moment of  $z_n$ . For odd  $r$  the moments take the form

$$\mathbb{E}(z_n^r) = \sum_{j=0}^{(r-3)/2} a_{rj} n^{-1/2-j}$$

and for even  $r$

$$\mathbb{E}(z_n^r) = (r-1)!! \sigma^{2r} + \sum_{j=1}^{(r-2)/2} b_{rj} n^{-j}$$

where  $a_{rj}$  and  $b_{rj}$  are the sum of constants multiplied by the products of cumulants whose indices sum to  $r$ . (Recall, the double factorial is  $k!! = k \cdot (k-2) \cdots 1$ .)

These are the exact (finite sample) moments of  $z_n$ . We can take the limit to find the asymptotic moments. As  $n \rightarrow \infty$ , for any odd  $r$  for which  $\mathbb{E}|y_i|^r < \infty$

$$\mathbb{E}(z_n^r) \rightarrow 0$$

and for any even  $r$  for which  $\mathbb{E}|y_i|^r < \infty$

$$\mathbb{E}(z_n^r) \rightarrow (r-1)!! \sigma^{2r}.$$

The limits are the moments of  $Z \sim N(0, \sigma^2)$ .

We have shown that when  $y_i$  has a finite  $r^{th}$  moment, the asymptotic  $r^{th}$  moment of the standardized mean matches that of the normal distribution. This may provide some intuition as to why the standardized mean converges to the normal distribution.

## 6.10 Multivariate Central Limit Theorem

Multivariate central limit theory applies when we consider vector-valued observations  $y_i$  and sample averages  $\bar{y}$ . In the i.i.d. case we know that the mean of  $\bar{y}$  is the mean vector  $\boldsymbol{\mu} = \mathbb{E}(y)$  and its variance is  $n^{-1}V$  where  $V = \mathbb{E}((y - \boldsymbol{\mu})(y - \boldsymbol{\mu})')$ . Again we wish to transform  $\bar{y}$  so that its mean and variance do not depend on  $n$ . We do this again by centering and scaling, by setting  $z_n = \sqrt{n}(\bar{y}_n - \boldsymbol{\mu})$ . This has mean  $\mathbf{0}$  and variance  $V$ , which are independent of  $n$  as desired.

To develop a distributional approximation for  $z_n$  we use a multivariate central limit theorem. We present three such results, corresponding to the three univariate results from the previous section. Each is derived from the univariate theory by the Cramér-Wold device (Theorem 6.10).

We first present the multivariate version of Theorem 6.11.

**Theorem 6.14 Multivariate Lindeberg–Lévy Central Limit Theorem.** If  $y_i \in \mathbb{R}^k$  are independent and identically distributed and  $\mathbb{E}\|y_i\|^2 < \infty$ , then as  $n \rightarrow \infty$

$$\sqrt{n}(\bar{y} - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, V)$$

where  $\boldsymbol{\mu} = \mathbb{E}(y)$  and  $V = \mathbb{E}((y - \boldsymbol{\mu})(y - \boldsymbol{\mu})')$ .

For a proof see Section 6.28.

We next present a multivariate version of Theorem 6.12.

**Theorem 6.15 Multivariate Lindeberg CLT.** Suppose that for all  $n$ ,  $\mathbf{y}_{ni} \in \mathbb{R}^k$ ,  $i = 1, \dots, r_n$ , are independent but not necessarily identically distributed with mean  $\mathbb{E}(\mathbf{y}_{ni}) = \mathbf{0}$  and variance matrices  $\mathbf{V}_{ni} = \mathbb{E}(\mathbf{y}_{ni}\mathbf{y}'_{ni})$ . Set  $\bar{\mathbf{V}}_n = \sum_{i=1}^n \mathbf{V}_{ni}$  and  $v_n^2 = \lambda_{\min}(\bar{\mathbf{V}}_n)$ . If  $v_n^2 > 0$  and for all  $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} \frac{1}{v_n^2} \sum_{i=1}^{r_n} \mathbb{E} \left( \|\mathbf{y}_{ni}\|^2 \mathbf{1} \left( \|\mathbf{y}_{ni}\|^2 \geq \varepsilon v_n^2 \right) \right) = 0 \quad (6.14)$$

then as  $n \rightarrow \infty$

$$\bar{\mathbf{V}}_n^{-1/2} \sum_{i=1}^{r_n} \mathbf{y}_{ni} \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_k).$$

For a proof see Section 6.28.

We finally present a multivariate version of Theorem 6.13.

**Theorem 6.16** Suppose  $\mathbf{y}_{ni} \in \mathbb{R}^k$  are independent but not necessarily identically distributed with means  $\mathbb{E}(\mathbf{y}_{ni}) = \mathbf{0}$  and variance matrices  $\mathbf{V}_{ni} = \mathbb{E}(\mathbf{y}_{ni}\mathbf{y}'_{ni})$ . Set  $\bar{\mathbf{V}}_n = n^{-1} \sum_{i=1}^n \mathbf{V}_{ni}$ . If

$$\bar{\mathbf{V}}_n \rightarrow \mathbf{V} > 0$$

and for some  $\delta > 0$

$$\sup_{n,i} \mathbb{E} \|\mathbf{y}_{ni}\|^{2+\delta} < \infty \quad (6.15)$$

then as  $n \rightarrow \infty$

$$\sqrt{n} \bar{\mathbf{y}} \xrightarrow{d} N(\mathbf{0}, \mathbf{V}).$$

For a proof see Section 6.28.

Similarly to Theorem 6.13, an advantage of Theorem 6.16 is that it allows the variance matrix  $\mathbf{V}$  to be singular.

## 6.11 Moments of Transformations

Often we want to estimate a parameter  $\boldsymbol{\mu}$  which is the expected value of a transformation of a random vector  $\mathbf{y}$ . That is,  $\boldsymbol{\mu}$  can be written as

$$\boldsymbol{\mu} = \mathbb{E}(\mathbf{h}(\mathbf{y}))$$

for some function  $\mathbf{h} : \mathbb{R}^m \rightarrow \mathbb{R}^k$ . For example, the second moment of  $y$  is  $\mathbb{E}(y^2)$ , the  $r^{th}$  is  $\mathbb{E}(y^r)$ , the moment generating function is  $\mathbb{E}(\exp(ty))$ , and the distribution function is  $\mathbb{E}(1\{y \leq x\})$ .

Estimating parameters of this form fits into our previous analysis by defining the random variable  $\mathbf{z} = \mathbf{h}(\mathbf{y})$ . Then  $\boldsymbol{\mu} = \mathbb{E}(\mathbf{z})$  is just a simple moment of  $\mathbf{z}$ . This suggests the moment estimator

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i = \frac{1}{n} \sum_{i=1}^n \mathbf{h}(\mathbf{y}_i).$$

For example, the moment estimator of  $\mathbb{E}(y^r)$  is  $n^{-1} \sum_{i=1}^n y_i^r$ , the moment estimator of the moment generating function is  $n^{-1} \sum_{i=1}^n \exp(ty_i)$ , and the estimator of the distribution function is  $n^{-1} \sum_{i=1}^n 1\{y_i \leq x\}$ .

Since  $\hat{\boldsymbol{\mu}}$  is a sample average, and transformations of i.i.d. variables are also i.i.d., the asymptotic results of the previous sections immediately apply.

**Theorem 6.17** If  $y_i$  are independent and identically distributed,  $\mu = \mathbb{E}(\mathbf{h}(y))$ , and  $\mathbb{E}\|\mathbf{h}(y)\| < \infty$ , then for  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \mathbf{h}(y_i)$ , as  $n \rightarrow \infty$ ,  $\hat{\mu} \xrightarrow{p} \mu$ .

**Theorem 6.18** If  $y_i$  are independent and identically distributed,  $\mu = \mathbb{E}(\mathbf{h}(y))$ , and  $\mathbb{E}(\|\mathbf{h}(y)\|^2) < \infty$ , then for  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \mathbf{h}(y_i)$ , as  $n \rightarrow \infty$ ,

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} N(\mathbf{0}, \mathbf{V})$$

where  $\mathbf{V} = \mathbb{E}((\mathbf{h}(y) - \mu)(\mathbf{h}(y) - \mu)')$ .

Theorems 6.17 and 6.18 show that the estimator  $\hat{\mu}$  is consistent for  $\mu$  and asymptotically normally distributed, so long as the stated moment conditions hold.

A word of caution. Theorems 6.17 and 6.18 give the impression that it is possible to estimate any moment of  $y$ . Technically this is the case so long as that moment is finite. What is hidden by the notation, however, is that estimates of high order moments can be quite imprecise. For example, consider the sample 8<sup>th</sup> moment  $\hat{\mu}_8 = \frac{1}{n} \sum_{i=1}^n y_i^8$ , and suppose for simplicity that  $y$  is  $N(0, 1)$ . Then we can calculate<sup>1</sup> that  $\text{var}(\hat{\mu}_8) = n^{-1} 2,016,000$ , which is immense, even for large  $n$ ! In general, higher-order moments are challenging to estimate because their variance depends upon even higher moments which can be quite large.

## 6.12 Smooth Function Model

We now expand our investigation and consider estimation of parameters which can be written as a continuous function of  $\mu = \mathbb{E}(\mathbf{h}(y))$ . That is, we consider cases where the parameter of interest can be written as

$$\theta = \mathbf{g}(\mu) = \mathbf{g}(\mathbb{E}(\mathbf{h}(y))) \quad (6.16)$$

for some functions  $\mathbf{g}: \mathbb{R}^k \rightarrow \mathbb{R}^\ell$  and  $\mathbf{h}: \mathbb{R}^m \rightarrow \mathbb{R}^k$ . This is generally known as the **smooth function model**, and encompasses a wide variety of econometric estimators.

As one example, the geometric mean of wages  $w$  is

$$\gamma = \exp(\mathbb{E}(\log(w))). \quad (6.17)$$

This is (6.16) with  $g(u) = \exp(u)$  and  $h(w) = \log(w)$ .

Another simple yet common example is the variance

$$\begin{aligned} \sigma^2 &= \mathbb{E}(w - \mathbb{E}(w))^2 \\ &= \mathbb{E}(w^2) - (\mathbb{E}(w))^2. \end{aligned}$$

This is (6.16) with

$$\mathbf{h}(w) = \begin{pmatrix} w \\ w^2 \end{pmatrix}$$

---

<sup>1</sup>By the formula for the variance of a mean  $\text{var}(\hat{\mu}_8) = n^{-1} (\mathbb{E}(y^{16}) - (\mathbb{E}(y^8))^2)$ . Since  $y$  is  $N(0, 1)$ ,  $\mathbb{E}(y^{16}) = 15!! = 2,027,025$  and  $\mathbb{E}(y^8) = 7!! = 105$  where  $k!!$  is the double factorial.

and

$$g(\mu_1, \mu_2) = \mu_2 - \mu_1^2.$$

Similarly, the skewness of the wage distribution is

$$sk = \frac{\mathbb{E}((w - \mathbb{E}(w))^3)}{(\mathbb{E}((w - \mathbb{E}(w))^2))^{3/2}}.$$

This is (6.16) with

$$\mathbf{h}(w) = \begin{pmatrix} w \\ w^2 \\ w^3 \end{pmatrix}$$

and

$$g(\mu_1, \mu_2, \mu_3) = \frac{\mu_3 - 3\mu_2\mu_1 + 2\mu_1^3}{(\mu_2 - \mu_1^2)^{3/2}}. \quad (6.18)$$

The parameter  $\boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\mu})$  is not a population moment, so it does not have a direct moment estimator. Instead, it is common to use a **plug-in estimator** formed by replacing the unknown  $\boldsymbol{\mu}$  with its point estimator  $\hat{\boldsymbol{\mu}}$  and then “plugging” this into the expression for  $\boldsymbol{\theta}$ . The first step is

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^n \mathbf{h}(y_i)$$

and the second step is

$$\hat{\boldsymbol{\theta}} = \mathbf{g}(\hat{\boldsymbol{\mu}}).$$

Again, the hat “ $\hat{\cdot}$ ” indicates that  $\hat{\boldsymbol{\theta}}$  is a sample estimator of  $\boldsymbol{\theta}$ .

For example, the plug-in estimate of the geometric mean  $\gamma$  of the wage distribution from (6.17) is

$$\hat{\gamma} = \exp(\hat{\mu})$$

with

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \log(w_i).$$

The plug-in estimator of the variance is

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n w_i^2 - \left( \frac{1}{n} \sum_{i=1}^n w_i \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^n (w_i - \bar{w})^2. \end{aligned}$$

The estimator for the skewness is

$$\begin{aligned} \hat{sk} &= \frac{\hat{\mu}_3 - 3\hat{\mu}_2\hat{\mu}_1 + 2\hat{\mu}_1^3}{(\hat{\mu}_2 - \hat{\mu}_1^2)^{3/2}} \\ &= \frac{\frac{1}{n} \sum_{i=1}^n (w_i - \bar{w})^3}{\left( \frac{1}{n} \sum_{i=1}^n (w_i - \bar{w})^2 \right)^{3/2}} \end{aligned}$$

where

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^n w_i^j.$$

In the next three sections we present a large-sample theory for the plug-in estimator for the smooth function model.

## 6.13 Continuous Mapping Theorem

Continuous functions are limit-preserving.

**Theorem 6.19 Continuous Mapping Theorem (CMT).** If  $z_n \xrightarrow{p} c$  as  $n \rightarrow \infty$  and  $\mathbf{g}(\cdot)$  is continuous at  $c$ , then  $\mathbf{g}(z_n) \xrightarrow{p} \mathbf{g}(c)$  as  $n \rightarrow \infty$ .

The proof of Theorem 6.19 is given in Section 6.28.

For example, if  $z_n \xrightarrow{p} c$  as  $n \rightarrow \infty$  then

$$\begin{aligned} z_n + a &\xrightarrow{p} c + a \\ az_n &\xrightarrow{p} ac \\ z_n^2 &\xrightarrow{p} c^2 \end{aligned}$$

as the functions  $g(u) = u + a$ ,  $g(u) = au$ , and  $g(u) = u^2$  are continuous. Also

$$\frac{a}{z_n} \xrightarrow{p} \frac{a}{c}$$

if  $c \neq 0$ . The condition  $c \neq 0$  is important as the function  $g(u) = a/u$  is not continuous at  $u = 0$ .

If  $\mathbf{y}_i$  are independent and identically distributed,  $\boldsymbol{\mu} = \mathbb{E}(\mathbf{h}(\mathbf{y}))$ , and  $\mathbb{E}\|\mathbf{h}(\mathbf{y})\| < \infty$ , then for  $\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^n \mathbf{h}(\mathbf{y}_i)$ , as  $n \rightarrow \infty$ ,  $\hat{\boldsymbol{\mu}} \xrightarrow{p} \boldsymbol{\mu}$ . Applying the CMT,  $\hat{\boldsymbol{\theta}} = \mathbf{g}(\hat{\boldsymbol{\mu}}) \xrightarrow{p} \mathbf{g}(\boldsymbol{\mu}) = \boldsymbol{\theta}$ .

**Theorem 6.20** If  $\mathbf{y}_i$  are i.i.d.,  $\boldsymbol{\theta} = \mathbf{g}(\mathbb{E}(\mathbf{h}(\mathbf{y})))$ ,  $\mathbb{E}\|\mathbf{h}(\mathbf{y})\| < \infty$ , and  $\mathbf{g}(\mathbf{u})$  is continuous at  $\mathbf{u} = \boldsymbol{\mu}$ , for  $\hat{\boldsymbol{\theta}} = \mathbf{g}(\hat{\boldsymbol{\mu}})$  with  $\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^n \mathbf{h}(\mathbf{y}_i)$ , then  $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$  as  $n \rightarrow \infty$ .

To apply Theorem 6.20 it is necessary to check if the function  $\mathbf{g}$  is continuous at  $\boldsymbol{\mu}$ . In our first example  $g(u) = \exp(u)$  is continuous everywhere. It therefore follows from Theorem 6.6 and Theorem 6.20 that if  $\mathbb{E}|\log(w)| < \infty$  then as  $n \rightarrow \infty$ ,  $\hat{\gamma} \xrightarrow{p} \gamma$ .

In the example of the variance,  $g$  is continuous for all  $\boldsymbol{\mu}$ . Thus if  $\mathbb{E}(w^2) < \infty$  then as  $n \rightarrow \infty$ ,  $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$ .

In our third example  $g$  defined in (6.18) is continuous for all  $\boldsymbol{\mu}$  such that  $\text{var}(w) = \mu_2 - \mu_1^2 > 0$ , which holds unless  $w$  has a degenerate distribution. Thus if  $\mathbb{E}|w|^3 < \infty$  and  $\text{var}(w) > 0$  then as  $n \rightarrow \infty$ ,  $\hat{s}k \xrightarrow{p} sk$ .

## 6.14 Delta Method

In this section we introduce two tools – an extended version of the CMT and the Delta Method – which allow us to calculate the asymptotic distribution of the plug-in estimator  $\hat{\boldsymbol{\theta}}$ .

We first present an extended version of the continuous mapping theorem which allows convergence in distribution.

**Theorem 6.21 Continuous Mapping Theorem**

If  $z_n \xrightarrow{d} z$  as  $n \rightarrow \infty$  and  $\mathbf{g} : \mathbb{R}^m \rightarrow \mathbb{R}^k$  has the set of discontinuity points  $D_g$  such that  $\mathbb{P}(z \in D_g) = 0$ , then  $\mathbf{g}(z_n) \xrightarrow{d} \mathbf{g}(z)$  as  $n \rightarrow \infty$ .

For a proof of Theorem 6.21 see Theorem 2.3 of van der Vaart (1998). It was first proved by Mann and Wald (1943) and is therefore sometimes referred to as the Mann-Wald Theorem.

Theorem 6.21 allows the function  $\mathbf{g}$  to be discontinuous only if the probability at being at a discontinuity point is zero. For example, the function  $g(u) = u^{-1}$  is discontinuous at  $u = 0$ , but if  $z_n \xrightarrow{d} z \sim N(0, 1)$  then  $\mathbb{P}(z = 0) = 0$  so  $z_n^{-1} \xrightarrow{d} z^{-1}$ .

A special case of the Continuous Mapping Theorem is known as Slutsky's Theorem.

**Theorem 6.22 Slutsky's Theorem**

If  $z_n \xrightarrow{d} z$  and  $c_n \xrightarrow{p} c$  as  $n \rightarrow \infty$ , then

1.  $z_n + c_n \xrightarrow{d} z + c$
2.  $z_n c_n \xrightarrow{d} zc$
3.  $\frac{z_n}{c_n} \xrightarrow{d} \frac{z}{c}$  if  $c \neq 0$ .

Even though Slutsky's Theorem is a special case of the CMT, it is a useful statement as it focuses on the most common applications – addition, multiplication, and division.

Despite the fact that the plug-in estimator  $\hat{\boldsymbol{\theta}}$  is a function of  $\hat{\boldsymbol{\mu}}$  for which we have an asymptotic distribution, Theorem 6.21 does not directly give us an asymptotic distribution for  $\hat{\boldsymbol{\theta}}$ . This is because  $\hat{\boldsymbol{\theta}} = \mathbf{g}(\hat{\boldsymbol{\mu}})$  is written as a function of  $\hat{\boldsymbol{\mu}}$ , not of the standardized sequence  $\sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$ . We need an intermediate step – a first order Taylor series expansion. This step is so critical to statistical theory that it has its own name – **The Delta Method**.

**Theorem 6.23 Delta Method**

If  $\sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \xrightarrow{d} \boldsymbol{\xi}$ , where  $\mathbf{g}(\boldsymbol{u})$  is continuously differentiable in a neighborhood of  $\boldsymbol{\mu}$  then as  $n \rightarrow \infty$

$$\sqrt{n}(\mathbf{g}(\hat{\boldsymbol{\mu}}) - \mathbf{g}(\boldsymbol{\mu})) \xrightarrow{d} \mathbf{G}' \boldsymbol{\xi} \quad (6.19)$$

where  $\mathbf{G}(\boldsymbol{u}) = \frac{\partial}{\partial \boldsymbol{u}} \mathbf{g}(\boldsymbol{u})'$  and  $\mathbf{G} = \mathbf{G}(\boldsymbol{\mu})$ . In particular, if  $\boldsymbol{\xi} \sim N(\mathbf{0}, \mathbf{V})$  then as  $n \rightarrow \infty$

$$\sqrt{n}(\mathbf{g}(\hat{\boldsymbol{\mu}}) - \mathbf{g}(\boldsymbol{\mu})) \xrightarrow{d} N(0, \mathbf{G}' \mathbf{V} \mathbf{G}). \quad (6.20)$$

A proof is presented in Section 6.28.

## 6.15 Asymptotic Distribution for Smooth Function Model

The Delta Method allows us to complete our derivation of the asymptotic distribution of the plug-in estimator  $\hat{\boldsymbol{\theta}}$  of  $\boldsymbol{\theta}$  in the smooth function model. By combining Theorems 6.18 and 6.23 we find the following.

**Theorem 6.24** If  $y_i$  are independent and identically distributed,  $\mu = \mathbb{E}(\mathbf{h}(y))$ ,  $\theta = \mathbf{g}(\mu)$ ,  $\mathbb{E}\|\mathbf{h}(y)\|^2 < \infty$ , and  $\mathbf{G}(\mathbf{u}) = \frac{\partial}{\partial \mathbf{u}} \mathbf{g}(\mathbf{u})'$  is continuous in a neighborhood of  $\mu$ , for  $\hat{\theta} = \mathbf{g}(\hat{\mu})$  with  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \mathbf{h}(y_i)$ , then as  $n \rightarrow \infty$

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_\theta)$$

where  $\mathbf{V}_\theta = \mathbf{G}' \mathbf{V} \mathbf{G}$ ,  $\mathbf{V} = \mathbb{E}((\mathbf{h}(y) - \mu)(\mathbf{h}(y) - \mu)')$  and  $\mathbf{G} = \mathbf{G}(\mu)$ .

Theorem 6.20 established the consistency of  $\hat{\theta}$  for  $\theta$ , and Theorem 6.24 established its asymptotic normality. It is instructive to compare the conditions required for these results. Consistency required that  $\mathbf{h}(y)$  have a finite mean, while asymptotic normality requires that this variable have a finite variance. Consistency required that  $\mathbf{g}(\mathbf{u})$  be continuous, while our proof of asymptotic normality used the assumption that  $\mathbf{g}(\mathbf{u})$  is continuously differentiable.

## 6.16 Covariance Matrix Estimation

To use asymptotic distribution in Theorem 6.24 we need an estimator of the asymptotic variance matrix  $\mathbf{V}_\theta = \mathbf{G}' \mathbf{V} \mathbf{G}$ . The natural plug-in estimator is

$$\begin{aligned}\hat{\mathbf{V}}_\theta &= \hat{\mathbf{G}}' \hat{\mathbf{V}} \hat{\mathbf{G}} \\ \hat{\mathbf{G}} &= \mathbf{G}(\hat{\mu}) \\ \hat{\mathbf{V}} &= \frac{1}{n} \sum_{i=1}^n (\mathbf{h}(y_i) - \hat{\mu})(\mathbf{h}(y_i) - \hat{\mu})'.\end{aligned}$$

Under the assumptions of Theorem 6.24, the WLLN implies  $\hat{\mu} \xrightarrow{p} \mu$ , and  $\hat{\mathbf{V}} \xrightarrow{p} \mathbf{V}$ . The CMT implies  $\hat{\mathbf{G}} \xrightarrow{p} \mathbf{G}$  and with a second application,  $\hat{\mathbf{V}}_\theta = \hat{\mathbf{G}}' \hat{\mathbf{V}} \hat{\mathbf{G}} \xrightarrow{p} \mathbf{G}' \mathbf{V} \mathbf{G} = \mathbf{V}_\theta$ . We have established that  $\hat{\mathbf{V}}_\theta$  is consistent for  $\mathbf{V}_\theta$ .

**Theorem 6.25** Under the assumptions of Theorem 6.24,  $\hat{\mathbf{V}}_\theta \xrightarrow{p} \mathbf{V}_\theta$  as  $n \rightarrow \infty$ .

## 6.17 t-ratios

When  $\ell = 1$  we can combine Theorems 6.24 and 6.25 to obtain the asymptotic distribution of the studentized statistic

$$T = \frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\hat{\mathbf{V}}_\theta}} \xrightarrow{d} \frac{N(0, V_\theta)}{\sqrt{V_\theta}} \sim N(0, 1).$$

The final equality is by the property that affine functions of normal random variables are normally distributed (Theorem 5.4).

**Theorem 6.26** Under the assumptions of Theorem 6.24,  $T \xrightarrow{d} N(0, 1)$  as  $n \rightarrow \infty$ .

## 6.18 Stochastic Order Symbols

It is convenient to have simple symbols for random variables and vectors which converge in probability to zero or are stochastically bounded. In this section we introduce some of the most commonly found notation.

It might be useful to review the common notation for non-random convergence and boundedness. Let  $x_n$  and  $a_n$ ,  $n = 1, 2, \dots$ , be non-random sequences. The notation

$$x_n = o(1)$$

(pronounced “small oh-one”) is equivalent to  $x_n \rightarrow 0$  as  $n \rightarrow \infty$ . The notation

$$x_n = o(a_n)$$

is equivalent to  $a_n^{-1}x_n \rightarrow 0$  as  $n \rightarrow \infty$ . The notation

$$x_n = O(1)$$

(pronounced “big oh-one”) means that  $x_n$  is bounded uniformly in  $n$  – there exists an  $M < \infty$  such that  $|x_n| \leq M$  for all  $n$ . The notation

$$x_n = O(a_n)$$

is equivalent to  $a_n^{-1}x_n = O(1)$ .

We now introduce similar concepts for sequences of random variables. Let  $z_n$  and  $a_n$ ,  $n = 1, 2, \dots$  be sequences of random variables. (In most applications,  $a_n$  is non-random.) The notation

$$z_n = o_p(1)$$

(“small oh-P-one”) means that  $z_n \xrightarrow{p} 0$  as  $n \rightarrow \infty$ . For example, for any consistent estimator  $\hat{\boldsymbol{\theta}}$  for  $\boldsymbol{\theta}$  we can write

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + o_p(1).$$

We also write

$$z_n = o_p(a_n)$$

if  $a_n^{-1}z_n = o_p(1)$ .

Similarly, the notation  $z_n = O_p(1)$  (“big oh-P-one”) means that  $z_n$  is bounded in probability. Precisely, for any  $\varepsilon > 0$  there is a constant  $M_\varepsilon < \infty$  such that

$$\limsup_{n \rightarrow \infty} \mathbb{P}(|z_n| > M_\varepsilon) \leq \varepsilon.$$

Furthermore, we write

$$z_n = O_p(a_n)$$

if  $a_n^{-1}z_n = O_p(1)$ .

$O_p(1)$  is weaker than  $o_p(1)$  in the sense that  $z_n = o_p(1)$  implies  $z_n = O_p(1)$  but not the reverse. However, if  $z_n = O_p(a_n)$  then  $z_n = o_p(b_n)$  for any  $b_n$  such that  $a_n/b_n \rightarrow 0$ .

If a random vector converges in distribution  $\mathbf{z}_n \xrightarrow{d} \mathbf{z}$  (for example, if  $\mathbf{z} \sim N(\mathbf{0}, \mathbf{V})$ ) then  $\mathbf{z}_n = O_p(1)$ . It follows that for estimators  $\hat{\boldsymbol{\theta}}$  which satisfy the convergence of Theorem 6.24 then we can write

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + O_p(n^{-1/2}).$$

In words, this statement says that the estimator  $\hat{\boldsymbol{\theta}}$  equals the true coefficient  $\boldsymbol{\theta}$  plus a random component which is bounded when scaled by  $n^{1/2}$ . Equivalently, we can write

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = O_p(1).$$

Another useful observation is that a random sequence with a bounded moment is stochastically bounded.

**Theorem 6.27** If  $\mathbf{z}_n$  is a random vector which satisfies

$$\mathbb{E} \|\mathbf{z}_n\|^\delta = O(a_n)$$

for some sequence  $a_n$  and  $\delta > 0$ , then

$$\mathbf{z}_n = O_p(a_n^{1/\delta}).$$

Similarly,  $\mathbb{E} \|\mathbf{z}_n\|^\delta = o(a_n)$  implies  $\mathbf{z}_n = o_p(a_n^{1/\delta})$ .

This can be shown using Markov's inequality (B.35). The assumptions imply that there is some  $M < \infty$  such that  $\mathbb{E} \|\mathbf{z}_n\|^\delta \leq Ma_n$  for all  $n$ . For any  $\varepsilon$  set  $B = \left(\frac{M}{\varepsilon}\right)^{1/\delta}$ . Then

$$\mathbb{P}\left(a_n^{-1/\delta} \|\mathbf{z}_n\| > B\right) = \mathbb{P}\left(\|\mathbf{z}_n\|^\delta > \frac{Ma_n}{\varepsilon}\right) \leq \frac{\varepsilon}{Ma_n} \mathbb{E} \|\mathbf{z}_n\|^\delta \leq \varepsilon$$

as required.

There are many simple rules for manipulating  $o_p(1)$  and  $O_p(1)$  sequences which can be deduced from the continuous mapping theorem or Slutsky's Theorem. For example,

$$\begin{aligned} o_p(1) + o_p(1) &= o_p(1) \\ o_p(1) + O_p(1) &= O_p(1) \\ O_p(1) + o_p(1) &= O_p(1) \\ o_p(1)o_p(1) &= o_p(1) \\ o_p(1)O_p(1) &= o_p(1) \\ O_p(1)O_p(1) &= O_p(1). \end{aligned}$$

## 6.19 Uniform WLLN\*

The weak law of large numbers (Theorem 6.2) states that for any distribution with a finite mean, the sample mean approaches the population mean in probability as the sample size gets large. One difficulty with this result, however, is that it is not uniform across distributions. Specifically, for any sample size  $n$ , no matter how large, there is a valid data distribution with a high probability that the sample mean is far from the population mean. To see this we use a constructed example. Suppose that the observations  $y_i$  come from the two-point distribution

$$y_i = \begin{cases} 1 & \text{with probability } 1-p \\ \frac{p-1}{p} & \text{with probability } p \end{cases}.$$

This random variable satisfies  $\mathbb{E}(y_i) = 0$ , and  $\mathbb{E}|y_i| = 2(1-p) < \infty$  so the WLLN implies that for any distribution (any  $p$ ),  $\bar{y} \xrightarrow{p} \mathbb{E}(y) = 0$ . However, the probability that the entire sample consists of all 1's (and thus  $\bar{y} = 1$ ) is  $(1-p)^n \simeq \exp(-pn)$ . For any sample size  $n$  there is a value of  $p$  such that this probability can be made arbitrarily close to 1. Thus, with probability arbitrarily close to 100%, the sample mean equals 1, not the population mean of zero.

We say that this is a failure of **uniform convergence**. While  $\bar{y}$  converges pointwise to  $\mathbb{E}(y)$ , it does not converge uniformly across all distributions for which  $\mathbb{E}|y_i| < \infty$ .

The solution is to restrict the set of distributions. It turns out that a sufficient condition is to require slightly more than one finite moment. That is, let  $\mathcal{F}$  denote the set of distributions which satisfy

$$\mathbb{E}|y_i|^r \leq B \tag{6.21}$$

for some given  $r > 1$  and  $B < \infty$ . This excludes, for example, two-point distributions with arbitrarily small probabilities  $p$ . Under this restriction the WLLN holds uniformly across the set of distributions.

**Theorem 6.28** Let  $\mathcal{F}$  denote the set of distributions which satisfy (6.21) for some  $r > 1$  and  $B < \infty$ . Then for all  $\varepsilon > 0$ , as  $n \rightarrow \infty$

$$\sup_{F \in \mathcal{F}} \mathbb{P}(|\bar{y} - \mathbb{E}(y)| > \varepsilon) \longrightarrow 0.$$

To prove this result, observe that by an application of Markov's inequality (B.35), the Bahr-Esseen inequality (B.44), and then (6.21),

$$\begin{aligned} \mathbb{P}(|\bar{y} - \mathbb{E}(y)| > \varepsilon) &= \mathbb{P}(|\bar{y} - \mathbb{E}(y)|^r > \varepsilon^r) \\ &\leq \frac{\mathbb{E}|\sum_{i=1}^n (y_i - \mathbb{E}(y_i))|^r}{\varepsilon^r n^r} \\ &\leq \frac{\sum_{i=1}^n \mathbb{E}|y_i - \mathbb{E}(y_i)|^r}{\varepsilon^r n^r} \\ &\leq \frac{2^r B}{\varepsilon^r n^{r-1}}. \end{aligned}$$

The right-hand-side does not depend on the distribution, only on the bound  $B$ . Thus

$$\sup_{F \in \mathcal{F}} \mathbb{P}(|\bar{y} - \mathbb{E}(y)| > \varepsilon) \leq \frac{2^r B}{\varepsilon^r n^{r-1}} \longrightarrow 0$$

as  $n \rightarrow \infty$ .

## 6.20 Uniform CLT\*

The Lindeberg-Lévy CLT (Theorem 6.11) states that for any distribution with a finite variance, the sample mean is approximately normally distributed for sufficiently large  $n$ . This does not mean, however, that a large sample size implies that the normal distribution is necessarily a good approximation. There is always a finite-variance distribution for which the normal approximation can be made arbitrarily poor.

Again consider the example from the previous section. Recall that  $\bar{y} \leq 1$ . Thus the standardized sample mean is bounded by

$$\frac{\bar{y}}{\sqrt{\text{var}(\bar{y})}} \leq \sqrt{\frac{np}{1-p}}.$$

Suppose  $p = 1/(n+1)$ . Then the distribution is truncated at 1. It follows that the  $N(0, 1)$  approximation is not accurate, in particular in the right tail.

As in the previous section the problem is a failure of uniform convergence. While the standardized sample mean converges to the normal distribution for every sampling distribution, it does not do so uniformly. For every sample size there is a distribution which can cause arbitrary failure in the asymptotic normal approximation.

Similar to the previous section the solution is to restrict the set of distributions. Unlike in the previous section, however, it is not sufficient to impose an upper moment bound. We also need to prevent asymptotically degenerate variances. A sufficient set of conditions is now given.

**Theorem 6.29** Let  $\mathcal{F}$  denote the set of distributions which satisfy (6.21) for some  $r > 2$  and  $B < \infty$  and in addition  $\text{var}(y_i) \geq \delta$  for some  $\delta > 0$ . Then for all  $x$ , as  $n \rightarrow \infty$

$$\sup_{F \in \mathcal{F}} \left| \mathbb{P} \left( \frac{\sqrt{n}(\bar{y} - \mathbb{E}(y))}{\sqrt{\text{var}(y_i)}} \leq x \right) - \Phi(x) \right| \rightarrow 0$$

where  $\Phi(x)$  is the standard normal distribution function.

Theorem 6.29 states that the standardized sample mean converges uniformly over  $\mathcal{F}$  to the normal distribution. This is a much stronger result than the classic CLT (Theorem 6.11) which is pointwise in  $\mathcal{F}$ .

The proof of Theorem 6.29 is refreshingly straightforward. If it were false there would be a sequence of distributions  $F_n \in \mathcal{F}$  and some  $x$  for which

$$\mathbb{P}_n \left( \frac{\sqrt{n}(\bar{y} - \mathbb{E}(y))}{\sqrt{\text{var}(y_i)}} \leq x \right) \rightarrow \Phi(x) \quad (6.22)$$

fails, where  $\mathbb{P}_n$  denotes the probability calculated under the distribution  $F_n$ . However, as discussed after Theorem 6.12, the assumptions of Theorem 6.29 imply that under the sequence  $F_n$  the Lindeberg condition (6.5) holds and hence the Lindeberg CLT holds. Thus (6.22) holds for all  $x$ .

## 6.21 Convergence of Moments\*

Sometimes we are interested in moments (often the mean and variance) of a statistic  $\mathbf{z}_n$ . When  $\mathbf{z}_n$  is a normalized sample mean we have direct expressions for the integer moments of  $\mathbf{z}_n$  (as presented in Section 6.9). But for other statistics, such as nonlinear functions of the sample mean, such expressions are not available.

The statement  $\mathbf{z}_n \xrightarrow{d} \mathbf{z}$  means that we can approximate the distribution of  $\mathbf{z}_n$  with that of  $\mathbf{z}$ . In this case we may approximate the moments of  $\mathbf{z}_n$  with those of  $\mathbf{z}$ . This can be rigorously justified if the moments of  $\mathbf{z}_n$  converge to the corresponding moments of  $\mathbf{z}$ . In this section we explore conditions under which this holds.

We first give a sufficient condition for the existence of the mean of the asymptotic distribution.

**Theorem 6.30** If  $\mathbf{z}_n \xrightarrow{d} \mathbf{z}$  and  $\mathbb{E} \|\mathbf{z}_n\| \leq C$  then  $\mathbb{E} \|\mathbf{z}\| \leq C$ .

A corollary is that  $\mathbb{E} (\|\mathbf{z}_n\|^r) \leq C$  implies  $\mathbb{E} (\|\mathbf{z}\|^r) \leq \infty$ .

To prove Theorem 6.30, let  $F_n(u)$  and  $F(u)$  be the distribution functions of  $\|\mathbf{z}_n\|$  and  $\|\mathbf{z}\|$ . Using Theorem 2.12, Definition 6.7, Fatou's Lemma, again Theorem 2.12, and the bound  $\mathbb{E} \|\mathbf{z}_n\| \leq C$ ,

$$\begin{aligned} \mathbb{E} \|\mathbf{z}\| &= \int_0^\infty (1 - F(x)) dx \\ &= \int_0^\infty \lim_{n \rightarrow \infty} (1 - F_n(x)) dx \\ &\leq \liminf_{n \rightarrow \infty} \int_0^\infty (1 - F_n(x)) dx \\ &= \liminf_{n \rightarrow \infty} \mathbb{E} \|\mathbf{z}_n\| \leq C \end{aligned}$$

as required.

We next consider conditions under which  $\mathbb{E}(z_n)$  converges to  $\mathbb{E}(z)$ . One might guess that the conditions of Theorem 6.30 would be sufficient, but a counter-example demonstrates that this is incorrect. Let  $z_n$  be a random variable which equals  $n$  with probability  $1/n$  and equals 0 with probability  $1 - 1/n$ . Then  $z_n \xrightarrow{d} z$  where  $\mathbb{P}(z = 0) = 1$ . We can also calculate that  $\mathbb{E}(z_n) = 1$ . Thus the assumptions of Theorem 6.30 are satisfied. However,  $\mathbb{E}(z_n) = 1$  does not converge to  $\mathbb{E}(z) = 0$ . Thus the boundedness of moments  $\mathbb{E}|z_n| \leq C < \infty$  is not sufficient to ensure the convergence of moments.

The problem is due to a lack of what is called “tightness” of the sequence of distributions. The culprit is the small probability mass which “escapes to infinity”.

The solution is to strengthen the assumption of boundedness of moments (integrability) to what is called uniform integrability. Recall that a random variable  $z$  is integrable if  $\mathbb{E}|z| = \int_{-\infty}^{\infty} |z| dF < \infty$ , or equivalently if

$$\lim_{M \rightarrow \infty} \mathbb{E}(|z| \mathbf{1}(|z| > M)) = \lim_{M \rightarrow \infty} \int_{-M}^M |z| dF = 0.$$

We say that a sequence of random variables is uniformly integrable if the limit is zero uniformly over the sequence.

**Definition 6.8** The random vector  $z_n$  is **uniformly integrable** as  $n \rightarrow \infty$  if

$$\lim_{M \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{E}(\|z_n\| \mathbf{1}(\|z_n\| > M)) = 0.$$

Uniform integrability is stronger than uniformly bounded moments. Indeed, the condition in the definition holds if for any  $\varepsilon > 0$  there is an  $M$  sufficiently large such that  $\mathbb{E}(\|z_N\| \mathbf{1}(\|z_N\| > M)) \leq \varepsilon$  for all  $N \geq n$ . This means  $\mathbb{E}\|z_N\| \leq M + \varepsilon$  so the moments are uniformly bounded. Uniform integrability is stronger than the uniform bound, as the example given previously does not satisfy uniform integrability. Specifically, take  $a_n$  as given previously. For any  $M < \infty$  set  $n = M + 1$ . Then  $\mathbb{E}(|a_n| \mathbf{1}(|a_n| > M)) = 1$  so does not limit to zero.

We can apply uniform integrability to powers of random variables. In particular we say  $z_n$  is **uniformly square integrable** if  $\|z_n\|^2$  is uniformly integrable, thus if

$$\lim_{M \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{E}(\|z_n\|^2 \mathbf{1}(\|z_n\|^2 > M)) = 0. \quad (6.23)$$

Uniform square integrability is similar (but slightly stronger) to the Lindeberg condition (6.5) when  $\bar{\sigma}_n^2 \geq \delta > 0$ . To see this, assume (6.23) holds for  $z_n = y_{ni}$ . Then for any  $\varepsilon > 0$  there is an  $M$  large enough so that  $\limsup_{n \rightarrow \infty} \mathbb{E}(z_n^2 \mathbf{1}(z_n^2 > M)) \leq \varepsilon \delta$ . Since  $\varepsilon n \bar{\sigma}_n^2 \rightarrow \infty$ , we have

$$\frac{1}{n \bar{\sigma}_n^2} \sum_{i=1}^n \mathbb{E}(y_{ni}^2 \mathbf{1}(y_{ni}^2 \geq \varepsilon n \bar{\sigma}_n^2)) \leq \frac{\varepsilon \delta}{\bar{\sigma}_n^2} \leq \varepsilon$$

which implies (6.5).

Uniform integrability is also implied by a bounded  $1 + \delta$  moment for some  $\delta > 0$ .

**Theorem 6.31** If for some  $\delta > 0$ ,  $\mathbb{E}\|z_n\|^{1+\delta} \leq C < \infty$ , then  $z_n$  is uniformly integrable.

A corollary is that  $\mathbb{E}\|\mathbf{z}_n\|^r \leq C < \infty$  implies  $\|\mathbf{z}_n\|^s$  is uniformly integrable for any  $s < r$ . To prove this theorem, fix  $\varepsilon$  and set  $M \geq (C/\varepsilon)^{1/\delta}$ . Then

$$\begin{aligned}\mathbb{E}(\|\mathbf{z}_n\| \mathbf{1}(\|\mathbf{z}_n\| > M)) &= \mathbb{E}\left(\frac{\|\mathbf{z}_n\|^{1+\delta}}{\|\mathbf{z}_n\|^\delta} \mathbf{1}(\|\mathbf{z}_n\| > M)\right) \\ &\leq \frac{\mathbb{E}(\|\mathbf{z}_n\|^{1+\delta} \mathbf{1}(\|\mathbf{z}_n\| > M))}{M^\delta} \\ &\leq \frac{\mathbb{E}\|\mathbf{z}_n\|^{1+\delta}}{M^\delta} \\ &\leq \frac{C}{M^\delta} \\ &\leq \varepsilon.\end{aligned}$$

Uniform integrability is the key condition which allows us to establish the convergence of moments.

**Theorem 6.32** If  $\mathbf{z}_n \xrightarrow{d} \mathbf{z}$  and  $\mathbf{z}_n$  is uniformly integrable then  $\mathbb{E}(\mathbf{z}_n) \rightarrow \mathbb{E}(\mathbf{z})$ .

Furthermore, if  $\mathbf{z}_n^r$  is uniformly integrable then  $\mathbb{E}(\mathbf{z}_n^r) \rightarrow \mathbb{E}(\mathbf{z}^r)$ .

We now prove Theorem 6.32. Without loss of generality assume  $\mathbf{z}_n$  is scalar and  $\mathbf{z}_n \geq 0$ . Let  $a \wedge b = \min(a, b)$ . Fix  $\varepsilon > 0$ . By Theorem 6.30  $\mathbf{z}$  is integrable, and by assumption  $\mathbf{z}_n$  is uniformly integrable. Thus we can find an  $M < \infty$  such that for all large  $n$ ,

$$\mathbb{E}(\mathbf{z} - \mathbf{z} \wedge M) = \mathbb{E}((\mathbf{z} - M) \mathbf{1}(\mathbf{z} > M)) \leq \mathbb{E}(\mathbf{z} \mathbf{1}(\mathbf{z} > M)) \leq \varepsilon$$

and

$$\mathbb{E}(\mathbf{z}_n - \mathbf{z}_n \wedge M) = \mathbb{E}((\mathbf{z}_n - M) \mathbf{1}(\mathbf{z}_n > M)) \leq \mathbb{E}(\mathbf{z}_n \mathbf{1}(\mathbf{z}_n > M)) \leq \varepsilon.$$

The function  $(\mathbf{z}_n \wedge M)$  is continuous and bounded. Since  $\mathbf{z}_n \xrightarrow{d} \mathbf{z}$ , Theorem 6.8 implies  $\mathbb{E}(\mathbf{z}_n \wedge M) \rightarrow \mathbb{E}(\mathbf{z} \wedge M)$ . Thus for  $n$  sufficiently large

$$|\mathbb{E}((\mathbf{z}_n \wedge M) - (\mathbf{z} \wedge M))| \leq \varepsilon.$$

Applying the triangle inequality (B.1) and the above three inequalities we find

$$|\mathbb{E}(\mathbf{z}_n - \mathbf{z})| \leq |\mathbb{E}(\mathbf{z}_n - (\mathbf{z}_n \wedge M))| + |\mathbb{E}((\mathbf{z}_n \wedge M) - (\mathbf{z} \wedge M))| + |\mathbb{E}(\mathbf{z} - (\mathbf{z} \wedge M))| \leq 3\varepsilon.$$

Since  $\varepsilon$  is arbitrary we conclude  $|\mathbb{E}(\mathbf{z}_n - \mathbf{z})| \rightarrow 0$  as required.

We complete this section by giving conditions under which moments of  $\mathbf{z}_n = \sqrt{n}(\bar{\mathbf{y}} - \mathbb{E}(\bar{\mathbf{y}}))$  converge to those of the normal distribution. In Section 6.9 we presented exact expressions for the integer moments of  $\mathbf{z}_n$ . We now consider non-integer moments as well.

**Theorem 6.33** If  $y_{ni}$  satisfies the conditions of Theorem 6.13, and  $\sup_{n,i} \mathbb{E}|y_{ni}|^r < \infty$  for some  $r > 2$ , then for any  $0 < s < r$ ,  $\mathbb{E}|z_n|^s \rightarrow \mathbb{E}|z|^s$  where  $z \sim N(0, \sigma^2)$ .

We now prove this result. Theorem 6.13 establishes  $z_n \xrightarrow{d} z$ , and the CMT establishes  $z_n^s \xrightarrow{d} z^s$ . We now establish that  $z_n^s$  is uniformly integrable. By Liapunov's inequality (B.34) and Minkowski's inequality (B.33), and  $\sup_{n,i} \mathbb{E} |y_{ni}|^r = B < \infty$

$$(\mathbb{E} |y_{ni} - \mathbb{E}(y_{ni})|^2)^{1/2} \leq (\mathbb{E} |y_{ni} - \mathbb{E}(y_{ni})|^r)^{1/r} \leq 2(\mathbb{E} |y_{ni}|^r)^{1/r} \leq 2B^{1/r}. \quad (6.24)$$

The Rosenthal inequality (B.50) establishes that there is a constant  $R_r < \infty$  such that

$$\begin{aligned} \mathbb{E} |z_n|^r &= \frac{1}{n^{r/2}} \mathbb{E} \left( \left| \sum_{i=1}^n (y_{ni} - \mathbb{E}(y_{ni})) \right|^r \right) \\ &\leq \frac{1}{n^{r/2}} R_r \left\{ \left( \sum_{i=1}^n \mathbb{E} |y_{ni} - \mathbb{E}(y_{ni})|^2 \right)^{r/2} + \sum_{i=1}^n \mathbb{E} |y_{ni} - \mathbb{E}(y_{ni})|^r \right\} \\ &\leq \frac{1}{n^{r/2}} R_r \left\{ (n4B^{2/r})^{r/2} + n2^r B \right\} \\ &\leq 2^{r+1} R_r B. \end{aligned}$$

The second inequality is (6.24). This shows that  $\mathbb{E} |z_n|^r$  is uniformly bounded, so  $|z_n|^s$  is uniformly integrable for any  $s < r$  by Theorem 6.31. Since  $|z_n|^s \xrightarrow{d} |z|^s$  and  $|z_n|^s$  is uniformly integrable, by Theorem 6.32, we conclude that  $\mathbb{E} |z_n|^s \rightarrow \mathbb{E} |z|^s$  as stated.

## 6.22 Edgeworth Expansion for the Sample Mean\*

The central limit theorem shows that normalized estimators are approximately normally distributed if the sample size  $n$  is sufficiently large. In practice, how good is this approximation? One way to measure the discrepancy between the actual distribution and the asymptotic distribution is by **higher-order expansions**. Higher-order expansions of the distribution function are known as **Edgeworth expansions**.

Let  $G_n(x)$  be the distribution function of the normalized mean  $z_n = \sqrt{n}(\bar{y}_n - \mu)/\sigma$  where  $\bar{y}_n = n^{-1} \sum_{i=1}^n y_i$  is the sample mean of i.i.d. random variables. An Edgeworth expansion is a series representation for  $G_n(x)$  expressed as powers of  $n^{-1/2}$ . It equals

$$G_n(x) = \Phi(x) - n^{-1/2} \frac{\kappa_3}{6} H e_2(x) \phi(x) - n^{-1} \left( \frac{\kappa_4}{24} H e_3(x) + \frac{\kappa_3^2}{72} H e_5(x) \right) \phi(x) + o(n^{-1}) \quad (6.25)$$

where  $\Phi(x)$  and  $\phi(x)$  are the standard normal distribution and density functions,  $\kappa_3$  and  $\kappa_4$  are the third and fourth cumulants of  $y_i$ , and

$$H e_j(x) = (-1)^j \frac{\phi^{(j)}(x)}{\phi(x)}$$

is the  $j^{\text{th}}$  Hermite polynomial. In particular,  $H e_2(x) = x^2 - 1$ ,  $H e_3(x) = x^3 - 3x$ , and  $H e_5(x) = x^5 - 10x^3 + 15x$ .

Below we give a justification for (6.25).

The expansion (6.25) may not be convergent. It is interpreted as an asymptotic series, meaning that the remainder is of a smaller order than the last included term. Sufficient regularity conditions for the validity of the expansion (6.25) are  $\mathbb{E}(y_i^4) < \infty$  and that the characteristic function of  $y$  is bounded below one. This latter – known as Cramer's condition – requires  $y$  to have an absolutely continuous distribution.

The expression (6.25) shows that the exact distribution  $G_n(x)$  can be written as the sum of the normal distribution, a  $n^{-1/2}$  correction for the main effect of skewness, and a  $n^{-1}$  correction for the main effect of kurtosis and the secondary effect of skewness. The  $n^{-1/2}$  skewness correction is an even function<sup>2</sup> of  $x$  which means that it changes the distribution function symmetrically about zero. This means that this

---

<sup>2</sup>A function  $f(x)$  is **even** if  $f(-x) = f(x)$ .

term captures skewness in the distribution function  $G_n(x)$ . The  $n^{-1}$  correction is an odd function<sup>3</sup> of  $x$  which means that this term moves probability mass symmetrically either away from, or towards, the center. This term captures kurtosis in the distribution of  $z_n$ .

We now derive (6.25) using the moment generating function. For a more rigorous argument the characteristic function could be used with minimal change in details. Let  $C_n(t) = \mathbb{E} \exp(tz_n)$  be the moment generating function of the normalized mean  $z_n$ . For simplicity assume  $\mu = 0$  and  $\sigma^2 = 1$ . In the proof of the central limit theorem (Theorem 6.11) we showed that

$$C_n(t) = \exp\left(nK\left(\frac{t}{\sqrt{n}}\right)\right)$$

where  $K(t) = \log(\mathbb{E} \exp(ty_i))$  is the cumulant generating function of  $y_i$  (see Section 2.33). By a series expansion about  $t = 0$ , the facts  $K(0) = K^{(1)}(0) = 0$ ,  $K^{(2)}(0) = 1$ ,  $K^{(3)}(0) = \kappa_3$  and  $K^{(4)}(0) = \kappa_4$ , this equals

$$\begin{aligned} C_n(t) &= \exp\left(\frac{t^2}{2} + n^{-1/2} \frac{\kappa_3}{6} t^3 + n^{-1} \frac{\kappa_4}{24} t^4 + o(n^{-1})\right) \\ &= \exp(t^2/2) + n^{-1/2} \exp(t^2/2) \frac{\kappa_3}{6} t^3 + n^{-1} \exp(t^2/2) \left(\frac{\kappa_4}{24} t^4 + \frac{\kappa_3^2}{72} t^6\right) + o(n^{-1}). \end{aligned} \quad (6.26)$$

The second line holds by taking a second-order expansion of the exponential function.

The Hermite polynomials satisfy

$$\frac{d}{dx} (He_j(x)\phi(x)) = -He_{j+1}(x)\phi(x). \quad (6.27)$$

By the formula for the normal MGF, the fact  $He_0(x) = 1$ , and repeated integration by parts applying (6.27), we find

$$\begin{aligned} \exp(t^2/2) &= \int_{-\infty}^{\infty} e^{tx} \phi(x) dx \\ &= \int_{-\infty}^{\infty} e^{tx} He_0(x) \phi(x) dx \\ &= t^{-1} \int_{-\infty}^{\infty} e^{tx} He_1(x) \phi(x) dx \\ &= t^{-2} \int_{-\infty}^{\infty} e^{tx} He_2(x) \phi(x) dx \\ &\vdots \\ &= t^{-j} \int_{-\infty}^{\infty} e^{tx} He_j(x) \phi(x) dx. \end{aligned}$$

This implies that for any  $j \geq 0$ ,

$$\exp(t^2/2) t^j = \int_{-\infty}^{\infty} e^{tx} He_j(x) \phi(x) dx.$$

Substituting into (6.26) we find

$$\begin{aligned} C_n(t) &= \int_{-\infty}^{\infty} e^{tx} \phi(x) dx + n^{-1/2} \frac{\kappa_3}{6} \int_{-\infty}^{\infty} e^{tx} He_3(x) \phi(x) dx \\ &\quad + n^{-1} \left( \frac{\kappa_4}{24} \int_{-\infty}^{\infty} e^{tx} He_4(x) \phi(x) dx + \frac{\kappa_3^2}{72} \int_{-\infty}^{\infty} e^{tx} He_6(x) \phi(x) dx \right) + o(n^{-1}) \\ &= \int_{-\infty}^{\infty} e^{tx} \left( \phi(x) + n^{-1/2} \frac{\kappa_3}{6} He_3(x) \phi(x) + n^{-1} \left( \frac{\kappa_4}{24} He_4(x) \phi(x) + \frac{\kappa_3^2}{72} He_6(x) \phi(x) \right) \right) dx \\ &= \int_{-\infty}^{\infty} e^{tx} d \left( \Phi(x) - n^{-1/2} \frac{\kappa_3}{6} He_2(x) \phi(x) - n^{-1} \left( \frac{\kappa_4}{24} He_3(x) + \frac{\kappa_3^2}{72} He_5(x) \right) \phi(x) \right) \end{aligned}$$

---

<sup>3</sup>A function  $f(x)$  is **odd** if  $f(-x) = -f(x)$ .

where the third equality uses (6.27). The final line shows that this is the MGF of the distribution in brackets which is (6.25). We have shown that the MGF expansion of  $z_n$  equals that of (6.25), so they are identical as claimed.

### Francis Edgeworth

Francis Ysidro Edgeworth (1845-1926) of Ireland, founding editor of the *Economic Journal*, was a profound economic and statistical theorist, developing the theories of indifference curves and asymptotic expansions. He also could be viewed as the first econometrician due to his early use of mathematical statistics in the study of economic data.

## 6.23 Edgeworth Expansion for Smooth Function Model\*

Most applications of Edgeworth expansions concern statistics which are more complicated than sample means. The following result applies to general smooth functions of means, which includes most estimators. This result includes that of the previous section as a special case.

**Theorem 6.34** If  $y_i$  are independent and identically distributed,  $\mu = \mathbb{E}(\mathbf{h}(y))$ ,  $\mathbb{E}\|\mathbf{h}(y)\|^4 < \infty$ ,  $g(u)$  has four continuous derivatives in a neighborhood of  $\mu$ , and  $\mathbb{E}(\exp(t\|\mathbf{h}(y)\|)) \leq B < 1$ , for  $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n \mathbf{h}(y_i)$ ,  $V = \mathbb{E}[(\mathbf{h}(y) - \mu)(\mathbf{h}(y) - \mu)']$  and  $\mathbf{G} = \mathbf{G}(\mu)$ , as  $n \rightarrow \infty$

$$\begin{aligned} \mathbb{P}\left(\frac{\sqrt{n}(g(\hat{\mu}) - g(\mu))}{\sqrt{\mathbf{G}'V\mathbf{G}}} \leq x\right) &= \Phi(x) + n^{-1/2} p_1(x)\phi(x) \\ &\quad + n^{-1} p_2(x)\phi(x) + o(n^{-1}) \end{aligned}$$

uniformly in  $x$ , where  $p_1(x)$  is an even polynomial of order 2, and  $p_2(x)$  is an odd polynomial of degree 5, with coefficients depending on the moments of  $\mathbf{h}(y)$  up to order 4.

For a proof see Theorem 2.2 of Hall (1992).

This Edgeworth expansion is identical in form to (6.25) derived in the previous section for the sample mean. The only difference is in the coefficients of the polynomials.

We are also interested in expansions for studentized statistics such as the t-ratio. Theorem 6.34 applies to such cases as well, so long as the variance estimator can be written as a function of sample means.

**Theorem 6.35** Under the assumptions of Theorem 6.34, if in addition  $\mathbb{E}\|\mathbf{h}(\mathbf{y})\|^8 < \infty$ ,  $g(\mathbf{u})$  has five continuous derivatives in a neighborhood of  $\boldsymbol{\mu}$ ,  $\mathbf{G}'\mathbf{V}\mathbf{G} > 0$ , and  $\mathbb{E}\left(\exp\left(t\|\mathbf{h}(\mathbf{y})\|^2\right)\right) \leq B < 1$ , for  $T$  and  $\widehat{\mathbf{G}}'\widehat{\mathbf{V}}\widehat{\mathbf{G}}$  as defined in Section 6.16, as  $n \rightarrow \infty$

$$\mathbb{P}(T \leq x) = \Phi(x) + n^{-1/2} p_1(x)\phi(x) + n^{-1} p_2(x)\phi(x) + o(n^{-1})$$

uniformly in  $x$ , where  $p_1(x)$  is an even polynomial of order 2, and  $p_2(x)$  is an odd polynomial of degree 5, with coefficients depending on the moments of  $\mathbf{h}(\mathbf{y})$  up to order 8.

Again this Edgeworth expansion is identical in form to the others presented, with the only difference appearing in the coefficients of the polynomials.

To see that Theorem 6.34 implies Theorem 6.35, define

$$\begin{aligned}\bar{\mathbf{h}}(\mathbf{y}_i) &= \begin{pmatrix} \mathbf{h}(\mathbf{y}_i) \\ \text{vec}(\mathbf{h}(\mathbf{y}_i)\mathbf{h}(\mathbf{y}_i)') \end{pmatrix} \\ \bar{\boldsymbol{\mu}} &= \frac{1}{n} \sum_{i=1}^n \bar{\mathbf{h}}(\mathbf{y}_i) = \begin{pmatrix} \widehat{\boldsymbol{\mu}} \\ \frac{1}{n} \sum_{i=1}^n \text{vec}(\mathbf{h}(\mathbf{y}_i)\mathbf{h}(\mathbf{y}_i)') \end{pmatrix}.\end{aligned}$$

Notice  $g(\widehat{\boldsymbol{\mu}})$ ,  $\widehat{\mathbf{G}} = \mathbf{G}(\widehat{\boldsymbol{\mu}})$  and  $\widehat{\mathbf{V}} = \frac{1}{n} \sum_{i=1}^n \mathbf{h}(\mathbf{y}_i)\mathbf{h}(\mathbf{y}_i)' - \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}'$  are all functions of  $\bar{\boldsymbol{\mu}}$ . We apply Theorem 6.34 to  $\sqrt{n}\bar{g}(\bar{\boldsymbol{\mu}})$  where

$$\bar{g}(\bar{\boldsymbol{\mu}}) = \frac{g(\widehat{\boldsymbol{\mu}}) - g(\boldsymbol{\mu})}{\sqrt{\widehat{\mathbf{G}}'\widehat{\mathbf{V}}\widehat{\mathbf{G}}}}.$$

The assumption  $\mathbb{E}\|\mathbf{h}(\mathbf{y})\|^8 < \infty$  implies  $\mathbb{E}\|\bar{\mathbf{h}}(\mathbf{y})\|^4 < \infty$ , and the assumptions that  $g(\mathbf{u})$  has five continuous derivatives and  $\mathbf{G}'\mathbf{V}\mathbf{G} > 0$  imply that  $\bar{g}(\mathbf{u})$  has four continuous derivatives. Thus the conditions of Theorem 6.34 are satisfied.

Theorem 6.35 is an Edgeworth expansion for a standard t-ratio. One implication is that when the normal distribution  $\Phi(x)$  is used as an approximation to the actual distribution  $\mathbb{P}(T \leq x)$ , the error is  $O(n^{-1/2})$

$$\mathbb{P}(T \leq x) - \Phi(x) = n^{-1/2} p_1(x)\phi(x) + O(n^{-1}) = O(n^{-1/2}).$$

Sometimes we are interested in the distribution of the absolute value of the t-ratio  $|T|$ . It has the distribution

$$\mathbb{P}(|T| \leq x) = \mathbb{P}(-x \leq T \leq x) = \mathbb{P}(T \leq x) - \mathbb{P}(T < x).$$

From Theorem 6.35 we find that this equals

$$\begin{aligned}&\Phi(x) + n^{-1/2} p_1(x)\phi(x) + n^{-1} p_2(x)\phi(x) \\ &- (\Phi(-x) + n^{-1/2} p_1(-x)\phi(-x) + n^{-1} p_2(-x)\phi(-x)) + o(n^{-1}) \\ &= 2\Phi(x) - 1 + n^{-1} 2p_2(x)\phi(x) + o(n^{-1})\end{aligned}$$

where the equality holds since  $\Phi(-x) = 1 - \Phi(x)$ ,  $\phi(-x) = \phi(x)$ ,  $p_1(-x) = p_1(x)$  (since  $\phi$  and  $p_1$  are even functions) and  $p_2(-x) = -p_2(x)$  (since  $p_2$  is an odd function). Thus when the normal distribution  $2\Phi(x) - 1$  is used as an approximation to the actual distribution  $\mathbb{P}(|T| \leq x)$ , the error is  $O(n^{-1})$

$$\mathbb{P}(|T| \leq x) - (2\Phi(x) - 1) = n^{-1} 2p_2(x)\phi(x) + o(n^{-1}) = O(n^{-1}).$$

What is occurring is that the  $O(n^{-1/2})$  skewness term affects the two distributional tails equally and offsetting. One tail has extra probability and the other has too little (relative to the normal approximation) so they offset. On the other hand the  $O(n^{-1})$  kurtosis term affects the two tails equally with the

same sign, so the effect doubles (either both tails have too much probability, or both have too little probability, relative to the normal).

There is also a version of the Delta Method for Edgeworth expansions. Essentially, if two random variables differ by  $O_p(a_n)$  then they have the same Edgeworth expansions up to  $O(a_n)$ .

**Theorem 6.36** Suppose the distribution of a random variable  $T$  has the Edgeworth expansion

$$\mathbb{P}(T \leq x) = \Phi(x) + a_n^{-1} p_1(x) \phi(x) + o(a_n^{-1})$$

and a random variable  $X$  satisfies  $X = T + o_p(a_n^{-1})$ . Then  $X$  has the Edgeworth expansion

$$\mathbb{P}(X \leq x) = \Phi(x) + a_n^{-1} p_1(x) \phi(x) + o(a_n^{-1}).$$

To prove this result, the assumption  $X = T + o_p(a_n^{-1})$  means that for any  $\varepsilon > 0$  there is  $n$  sufficiently large such that  $\mathbb{P}(|X - T| > a_n^{-1} \varepsilon) \leq \varepsilon$ . Then

$$\begin{aligned} \mathbb{P}(X \leq x) &\leq \mathbb{P}(X \leq x, |X - T| \leq a_n^{-1} \varepsilon) + \varepsilon \\ &\leq \mathbb{P}(T \leq x + a_n^{-1} \varepsilon) + \varepsilon \\ &= \Phi(x + a_n^{-1} \varepsilon) + a_n^{-1} p_1(x + a_n^{-1} \varepsilon) \phi(x + a_n^{-1} \varepsilon) + \varepsilon + o(a_n^{-1}) \\ &\leq \Phi(x) + a_n^{-1} p_1(x) \phi(x) + \frac{a_n^{-1} \varepsilon}{\sqrt{2\pi}} + \varepsilon + o(a_n^{-1}) \\ &\leq \Phi(x) + a_n^{-1} p_1(x) \phi(x) + o(a_n^{-1}) \end{aligned}$$

the last inequality since  $\varepsilon$  is arbitrary. Similarly,  $\mathbb{P}(X \leq x) \geq \Phi(x) + n^{-1/2} p_1(x) \phi(x) + o(a_n^{-1})$ .

## 6.24 Cornish-Fisher Expansions\*

The previous two sections described expansions for distribution functions. For some purposes it is useful to have similar expansions for the inverse of the distribution function, which are the quantiles of the distribution. Such expansions are known as Cornish-Fisher expansions. Recall, the  $\alpha^{th}$  quantile of a continuous distribution  $F(u)$  is the solution to  $F(q) = \alpha$ . Suppose that a statistic  $T$  has distribution  $G_n(x) = \mathbb{P}(T \leq x)$ . For any  $\alpha \in (0, 1)$  its  $\alpha^{th}$  quantile  $q_\alpha$  is the solution to  $G_n(q_\alpha) = \alpha$ . Let  $z_\alpha$  be the  $\alpha^{th}$  quantile of the standard normal, e.g.  $\Phi(z_\alpha) = \alpha$ .

**Theorem 6.37** Suppose the distribution of a random variable  $T$  has the Edgeworth expansion

$$G_n(x) = \mathbb{P}(T \leq x) = \Phi(x) + n^{-1/2} p_1(x) \phi(x) + n^{-1} p_2(x) \phi(x) + o(n^{-1})$$

uniformly in  $x$ . For any  $\alpha \in (0, 1)$  let  $q_\alpha$  and  $z_\alpha$  be the  $\alpha^{th}$  quantile of  $G_n(u)$  and  $\Phi(u)$ , that is the solutions to  $G_n(q_\alpha) = \alpha$  and  $\Phi(z_\alpha) = \alpha$ . Then

$$q_\alpha = z_\alpha + n^{-1/2} p_{11}(z_\alpha) + n^{-1} p_{21}(z_\alpha) + o(n^{-1}) \quad (6.28)$$

where

$$p_{11}(x) = -p_1(x) \quad (6.29)$$

$$p_{21}(x) = -p_2(x) + p_1(x) p'_1(x) - \frac{1}{2} x p_1(x)^2. \quad (6.30)$$

Under the conditions of Theorem 6.35, the functions  $p_{11}(x)$  and  $p_{21}(x)$  are even and odd functions of  $x$  with coefficients depending on the moments of  $\mathbf{h}(\mathbf{y})$  up to order 4.

Theorem 6.37 can be derived from the Edgeworth expansion using Taylor expansions. Evaluating the Edgeworth expansion at  $q_\alpha$ , substituting in (6.28), we have

$$\begin{aligned}\alpha &= G_n(q_\alpha) \\ &= \Phi(q_\alpha) + n^{-1/2} p_1(q_\alpha)\phi(q_\alpha) + n^{-1} p_2(q_\alpha)\phi(q_\alpha) + o(n^{-1}) \\ &= \Phi(z_\alpha + n^{-1/2} p_{11}(z_\alpha) + n^{-1} p_{21}(z_\alpha)) \\ &\quad + n^{-1/2} p_1(z_\alpha + n^{-1/2} p_{11}(z_\alpha))\phi(z_\alpha + n^{-1/2} p_{11}(z_\alpha)) \\ &\quad + n^{-1} p_{21}(z_\alpha) + o(n^{-1}).\end{aligned}$$

Next, expand  $\Phi(x)$  in a second-order Taylor expansion and  $p_1(x)$  and  $\phi(x)$  in first-order expansions, both about  $z_\alpha$ . We obtain that the above expression equals

$$\begin{aligned}&\Phi(z_\alpha) + n^{-1/2} \phi(z_\alpha) (p_{11}(z_\alpha) + p_1(z_\alpha)) \\ &+ n^{-1} \phi(z_\alpha) \left( p_{21}(z_\alpha) - \frac{z_\alpha p_1(z_\alpha)^2}{2} + p'_1(z_\alpha) p_{11}(z_\alpha) - z_\alpha p_1(z_\alpha) p_{11}(z_\alpha) + p_2(z_\alpha) \right) + o(n^{-1}).\end{aligned}$$

For this to equal  $\alpha$ , we deduce that  $p_{11}(x)$  and  $p_{21}(x)$  must take the values given in (6.29)-(6.30).

## 6.25 Uniform Stochastic Bounds\*

For some applications it can be useful to obtain the stochastic order of the random variable

$$\max_{1 \leq i \leq n} |y_i|.$$

This is the magnitude of the largest observation in the sample  $\{y_1, \dots, y_n\}$ . If the support of the distribution of  $y_i$  is unbounded, then as the sample size  $n$  increases, the largest observation will also tend to increase. It turns out that there is a simple characterization.

**Theorem 6.38** If  $|y_i|^r$  is uniformly integrable, then as  $n \rightarrow \infty$

$$n^{-1/r} \max_{1 \leq i \leq n} |y_i| \xrightarrow{p} 0. \quad (6.31)$$

Furthermore, if  $\exp(ty_i)$  is uniformly integrable for some  $t > 0$ , then for any  $\eta > 0$

$$(\log n)^{-(1+\eta)} \max_{1 \leq i \leq n} |y_i| \xrightarrow{p} 0. \quad (6.32)$$

The proof of Theorem 6.38 is presented in Section 6.28.

Equivalently, (6.31) can be written as

$$\max_{1 \leq i \leq n} |y_i| = o_p(n^{1/r}) \quad (6.33)$$

and (6.32) as

$$\max_{1 \leq i \leq n} |y_i| = o_p(\log n). \quad (6.34)$$

Equation (6.33) says that if  $y$  has  $r$  finite moments, then the largest observation will diverge at a rate slower than  $n^{1/r}$ . As  $r$  increases this rate decreases. Equation (6.34) shows that if we strengthen this

to  $y$  having all finite moments and a finite moment generating function (for example, if  $y$  is normally distributed) then the largest observation will diverge slower than  $\log n$ . Thus the higher the moments, the slower the rate of divergence.

To simplify the notation, we may write (6.33) as  $y_i = o_p(n^{1/r})$  uniformly in  $1 \leq i \leq n$ . It is important to understand when the  $O_p$  or  $o_p$  symbols are applied to subscript  $i$  random variables whether the convergence is pointwise in  $i$ , or is uniform in  $i$  in the sense of (6.33)-(6.34).

Theorem 6.38 applies to random vectors. For example, if  $\mathbb{E} \|y\|^r < \infty$  then

$$\max_{1 \leq i \leq n} \|y_i\| = o_p(n^{1/r}).$$

## 6.26 Marcinkiewicz Weak Law of Large Numbers\*

**Theorem 6.39** If  $y_i$  are independent and uniformly integrable, then for any  $r > 1$ , as  $n \rightarrow \infty$

$$n^{-r} \sum_{i=1}^n |y_i|^r \xrightarrow{p} 0.$$

To see an interesting implication of Theorem 6.39, recall that the sample mean  $\bar{y}$  has variance  $\text{var}(\bar{y}) = \sigma^2/n$  which has the natural estimator  $\widehat{\text{var}}(\bar{y}) = \hat{\sigma}^2/n$  where  $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (y_i - \bar{y})^2$  is the sample variance. Theorem 6.39 with  $r = 2$  implies that if the observations are i.i.d. and  $\mathbb{E}|y| < \infty$  then

$$\widehat{\text{var}}(\bar{y}) = n^{-2} \sum_{i=1}^n (y_i - \bar{y})^2 \leq n^{-2} \sum_{i=1}^n y_i^2 \xrightarrow{p} 0.$$

This is notable because it only requires that the first moment of the distribution is finite ( $\mathbb{E}|y| < \infty$ ). The result holds even if the true variance is infinite ( $\mathbb{E}(y^2) = \infty$ ). Thus the estimator of the variance of  $\bar{y}$  converges in probability to zero, even when the true variance is infinite.

We will not use the Marcinkiewicz weak law for our standard asymptotic theory, but will find it useful in our study of bootstrap asymptotic theory in Chapter 10.

We close this section by providing a proof of Theorem 6.39:

$$n^{-r} \sum_{i=1}^n |y_i|^r \leq \left( n^{-1} \max_{1 \leq i \leq n} |y_i| \right)^{r-1} \frac{1}{n} \sum_{i=1}^n |y_i| = o_p(1) O_p(1) \xrightarrow{p} 0$$

by Theorems 6.2 and 6.38, and  $r > 1$ .

## 6.27 Semiparametric Efficiency\*

In this section we argue that the sample mean  $\hat{\mu}$  and plug-in estimator  $\hat{\theta} = \mathbf{g}(\hat{\mu})$  are efficient estimators of the parameters  $\mu$  and  $\theta$ . Our demonstration is based on the rich but technically challenging theory of semiparametric efficiency bounds. An excellent accessible review has been provided by Newey (1990). We will also appeal to the asymptotic theory of maximum likelihood estimation (see Chapter 5).

We start by examining the sample mean  $\hat{\mu}$ , for the asymptotic efficiency of  $\hat{\theta}$  will follow from that of  $\hat{\mu}$ .

Recall, we know that if  $\mathbb{E}\|y\|^2 < \infty$  then the sample mean has the asymptotic distribution  $\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} N(0, V)$ . We want to know if  $\hat{\mu}$  is the best feasible estimator, or if there is another estimator with a smaller asymptotic variance. While it seems intuitively unlikely that another estimator could have a smaller asymptotic variance, how do we know that this is not the case?

When we ask if  $\hat{\mu}$  is the best estimator, we need to be clear about the class of models – the class of permissible distributions. For estimation of the mean  $\mu$  of the distribution of  $\mathbf{y}$  the broadest conceivable class is  $\mathcal{L}_1 = \{F : \mathbb{E} \|\mathbf{y}\| < \infty\}$ . This class is too broad for our current purposes, as  $\hat{\mu}$  is not asymptotically  $N(0, V)$  for all  $F \in \mathcal{L}_1$ . A more realistic choice is  $\mathcal{L}_2 = \{F : \mathbb{E} \|\mathbf{y}\|^2 < \infty\}$  – the class of finite-variance distributions. When we seek an efficient estimator of the mean  $\mu$  in the class of models  $\mathcal{L}_2$  what we are seeking is the best estimator, given that all we know is that  $F \in \mathcal{L}_2$ .

To show that the answer is not immediately obvious, it might be helpful to review a setting where the sample mean is inefficient. Suppose that  $y \in \mathbb{R}$  has the double exponential density  $f(y | \mu) = 2^{-1/2} \exp(-|y - \mu| \sqrt{2})$ . Since  $\text{var}(y) = 1$  we see that the sample mean satisfies  $\sqrt{n}(\bar{\mu} - \mu) \xrightarrow{d} N(0, 1)$ . In this model the maximum likelihood estimator (MLE)  $\tilde{\mu}$  for  $\mu$  is the sample median. Recall from the theory of maximum likelihood that the MLE satisfies  $\sqrt{n}(\tilde{\mu} - \mu) \xrightarrow{d} N(0, (\mathbb{E}(S^2))^{-1})$  where  $S = \frac{\partial}{\partial \mu} \log f(y | \mu) = -\sqrt{2} \text{sgn}(y - \mu)$  is the score. We can calculate that  $\mathbb{E}(S^2) = 2$  and thus conclude that  $\sqrt{n}(\tilde{\mu} - \mu) \xrightarrow{d} N(0, 1/2)$ . The asymptotic variance of the MLE is one-half that of the sample mean. Thus when the true density is known to be double exponential the sample mean is inefficient.

But the estimator which achieves this improved efficiency – the sample median – is not generically consistent for the population mean. It is inconsistent if the density is asymmetric or skewed. So the improvement comes at a great cost. Another way of looking at this is that the sample median is efficient in the class of densities  $\{f(y | \mu) = 2^{-1/2} \exp(-|y - \mu| \sqrt{2})\}$  but unless it is known that this is the correct distribution class this knowledge is not very useful.

The relevant question is whether or not the sample mean is efficient when the form of the distribution is unknown. We call this setting **semiparametric** as the parameter of interest (the mean) is finite dimensional while the remaining features of the distribution are unspecified. In the semiparametric context an estimator is called **semiparametrically efficient** if it has the smallest asymptotic variance among all semiparametric estimators.

The mathematical trick is to reduce the semiparametric model to a set of parametric “submodels”. The Cramer-Rao variance bound can be found for each parametric submodel. The variance bound for the semiparametric model (the union of the submodels) is then defined as the supremum of the individual variance bounds.

Formally, suppose that the true density of  $\mathbf{y}$  is the unknown function  $f(\mathbf{y})$  with mean  $\mu = \mathbb{E}(\mathbf{y}) = \int \mathbf{y} f(\mathbf{y}) d\mathbf{y}$ . A parametric submodel  $\eta$  for  $f(\mathbf{y})$  is a density  $f_\eta(\mathbf{y} | \boldsymbol{\theta})$  which is a smooth function of a parameter  $\boldsymbol{\theta}$ , and there is a true value  $\boldsymbol{\theta}_0$  such that  $f_\eta(\mathbf{y} | \boldsymbol{\theta}_0) = f(\mathbf{y})$ . The index  $\eta$  indicates the submodels. The equality  $f_\eta(\mathbf{y} | \boldsymbol{\theta}_0) = f(\mathbf{y})$  means that the submodel class passes through the true density, so the submodel is a true model. The class of submodels  $\eta$  and parameter  $\boldsymbol{\theta}_0$  depend on the true density  $f$ . In the submodel  $f_\eta(\mathbf{y} | \boldsymbol{\theta})$ , the mean is  $\mu_\eta(\boldsymbol{\theta}) = \int \mathbf{y} f_\eta(\mathbf{y} | \boldsymbol{\theta}) d\mathbf{y}$  which varies with the parameter  $\boldsymbol{\theta}$ . Let  $\eta \in \aleph$  be the class of all submodels for  $f$ .

Since each submodel  $\eta$  is parametric we can calculate the efficiency bound for estimation of  $\mu$  within this submodel. Specifically, given the density  $f_\eta(\mathbf{y} | \boldsymbol{\theta})$  its likelihood score is

$$S_\eta = \frac{\partial}{\partial \boldsymbol{\theta}} \log f_\eta(\mathbf{y} | \boldsymbol{\theta}_0),$$

so the Cramer-Rao lower bound for estimation of  $\boldsymbol{\theta}$  is  $(\mathbb{E}(S_\eta S'_\eta))^{-1}$ . Defining  $\mathbf{M}_\eta = \frac{\partial}{\partial \boldsymbol{\theta}} \mu_\eta(\boldsymbol{\theta}_0)'$ , by Theorem 5.25 the Cramer-Rao lower bound for estimation of  $\mu$  within the submodel  $\eta$  is  $V_\eta = \mathbf{M}'_\eta (\mathbb{E}(S_\eta S'_\eta))^{-1} \mathbf{M}_\eta$ .

As  $V_\eta$  is the efficiency bound for the submodel class  $f_\eta(\mathbf{y} | \boldsymbol{\theta})$ , no estimator can have an asymptotic variance smaller than  $V_\eta$  for any density  $f_\eta(\mathbf{y} | \boldsymbol{\theta})$  in the submodel class, including the true density  $f$ . This is true for all submodels  $\eta$ . Thus the asymptotic variance of any semiparametric estimator cannot be smaller than  $V_\eta$  for any conceivable submodel. Taking the supremum of the Cramer-Rao bounds

from all conceivable submodels we define<sup>4</sup>

$$\bar{V} = \sup_{\eta \in \mathfrak{X}} V_\eta.$$

The asymptotic variance of any semiparametric estimator cannot be smaller than  $\bar{V}$ , since it cannot be smaller than any individual  $V_\eta$ . We call  $\bar{V}$  the **semiparametric asymptotic variance bound** or **semiparametric efficiency bound** for estimation of  $\mu$ , as it is a lower bound on the asymptotic variance for any semiparametric estimator. If the asymptotic variance of a specific semiparametric estimator equals the bound  $\bar{V}$  we say that the estimator is **semiparametrically efficient**.

For many statistical problems it is quite challenging to calculate the semiparametric variance bound. However, in some cases there is a simple method to find the solution. Suppose that we can find a submodel  $\eta_0$  whose Cramer-Rao lower bound satisfies  $V_{\eta_0} = V_\mu$  where  $V_\mu$  is the asymptotic variance of a known semiparametric estimator. In this case, we can deduce that  $\bar{V} = V_{\eta_0} = V_\mu$ . Otherwise (that is, if  $V_{\eta_0}$  is not the efficiency bound) there would exist another submodel  $\eta_1$  whose Cramer-Rao lower bound satisfies  $V_{\eta_0} < V_{\eta_1}$  (because  $V_{\eta_0}$  is not the supremum). This would imply  $V_\mu < V_{\eta_1}$  which contradicts the Cramer-Rao Theorem (since when submodel  $\eta_1$  is true then no estimator can have a lower variance than  $V_{\eta_1}$ ).

We now find this submodel for the sample mean  $\hat{\mu}$ . Our goal is to find a parametric submodel whose Cramer-Rao bound for  $\mu$  is  $V$ . This can be done by creating a tilted version of the true density. Consider the parametric submodel

$$f_\eta(\mathbf{y} | \boldsymbol{\theta}) = f(\mathbf{y}) (1 + \boldsymbol{\theta}' \mathbf{V}^{-1} (\mathbf{y} - \mu)) \quad (6.35)$$

where  $f(\mathbf{y})$  is the true density and  $\mu = \mathbb{E}(\mathbf{y})$ . Note that

$$\int f_\eta(\mathbf{y} | \boldsymbol{\theta}) d\mathbf{y} = \int f(\mathbf{y}) d\mathbf{y} + \boldsymbol{\theta}' \mathbf{V}^{-1} \int f(\mathbf{y}) (\mathbf{y} - \mu) d\mathbf{y} = 1$$

and for all  $\boldsymbol{\theta}$  close to zero  $f_\eta(\mathbf{y} | \boldsymbol{\theta}) \geq 0$ . Thus  $f_\eta(\mathbf{y} | \boldsymbol{\theta})$  is a valid density function. It is a parametric submodel since  $f_\eta(\mathbf{y} | \boldsymbol{\theta}_0) = f(\mathbf{y})$  when  $\boldsymbol{\theta}_0 = 0$ . This parametric submodel has the mean

$$\begin{aligned} \mu(\boldsymbol{\theta}) &= \int \mathbf{y} f_\eta(\mathbf{y} | \boldsymbol{\theta}) d\mathbf{y} \\ &= \int \mathbf{y} f(\mathbf{y}) d\mathbf{y} + \int f(\mathbf{y}) \mathbf{y} (\mathbf{y} - \mu)' \mathbf{V}^{-1} \boldsymbol{\theta} d\mathbf{y} \\ &= \mu + \boldsymbol{\theta} \end{aligned}$$

which is a smooth function of  $\boldsymbol{\theta}$ .

Since

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log f_\eta(\mathbf{y} | \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log (1 + \boldsymbol{\theta}' \mathbf{V}^{-1} (\mathbf{y} - \mu)) = \frac{\mathbf{V}^{-1} (\mathbf{y} - \mu)}{1 + \boldsymbol{\theta}' \mathbf{V}^{-1} (\mathbf{y} - \mu)}$$

it follows that the score function for  $\boldsymbol{\theta}$  is

$$\mathbf{S}_\eta = \frac{\partial}{\partial \boldsymbol{\theta}} \log f_\eta(\mathbf{y} | \boldsymbol{\theta}_0) = \mathbf{V}^{-1} (\mathbf{y} - \mu).$$

By Theorem 5.25 the Cramer-Rao lower bound for  $\boldsymbol{\theta}$  is

$$(\mathbb{E}(\mathbf{S}_\eta \mathbf{S}_\eta'))^{-1} = (\mathbf{V}^{-1} \mathbb{E}((\mathbf{y} - \mu)(\mathbf{y} - \mu)') \mathbf{V}^{-1})^{-1} = \mathbf{V}.$$

The Cramer-Rao lower bound for  $\mu(\boldsymbol{\theta}) = \mu + \boldsymbol{\theta}$  is also  $\mathbf{V}$ , and this equals the asymptotic variance of the moment estimator  $\hat{\mu}$ . This was what we set out to show.

In summary, we have shown that in the submodel (6.35) the Cramer-Rao lower bound for estimation of  $\mu$  is  $\mathbf{V}$  which equals the asymptotic variance of the sample mean. This establishes the following result.

---

<sup>4</sup>It is not obvious that this supremum exists, as  $V_\eta$  is a matrix so there is not a unique ordering of matrices. However, in many cases (including the ones we study) the supremum exists and is unique.

**Proposition 6.1** In the class of distributions  $F \in \mathcal{L}_2$ , the semiparametric variance bound for estimation of  $\boldsymbol{\mu}$  is  $V = \text{var}(y_i)$ , and the sample mean  $\hat{\boldsymbol{\mu}}$  is a semiparametrically efficient estimator of the population mean  $\boldsymbol{\mu}$ .

We call this result a proposition rather than a theorem as we have not attended to the regularity conditions.

It is a simple matter to extend this result to the plug-in estimator  $\hat{\boldsymbol{\theta}} = \mathbf{g}(\hat{\boldsymbol{\mu}})$ . We know from Theorem 6.24 that if  $\mathbb{E}(\|y\|^2) < \infty$  and  $\mathbf{g}(\mathbf{u})$  is continuously differentiable at  $\mathbf{u} = \boldsymbol{\mu}$  then the plug-in estimator has the asymptotic distribution  $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(0, \mathbf{G}'V\mathbf{G})$ . We therefore consider the class of distributions

$$\mathcal{L}_2(\mathbf{g}) = \left\{ F : \mathbb{E}(\|y\|^2) < \infty, \mathbf{g}(\mathbf{u}) \text{ is continuously differentiable at } \mathbf{u} = \mathbb{E}(y) \right\}.$$

For example, if  $\theta = \mu_1/\mu_2$  where  $\mu_1 = \mathbb{E}(y_1)$  and  $\mu_2 = \mathbb{E}(y_2)$  then

$$\mathcal{L}_2(g) = \left\{ F : \mathbb{E}(y_1^2) < \infty, \mathbb{E}(y_2^2) < \infty, \text{ and } \mathbb{E}(y_2) \neq 0 \right\}.$$

For any submodel  $\eta$  the Cramer-Rao lower bound for estimation of  $\boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\mu})$  is  $\mathbf{G}'V_\eta\mathbf{G}$ . For the submodel (6.35) this bound is  $\mathbf{G}'V\mathbf{G}$  which equals the asymptotic variance of  $\hat{\boldsymbol{\theta}}$  from Theorem 6.24. Thus  $\hat{\boldsymbol{\theta}}$  is semiparametrically efficient.

**Proposition 6.2** In the class of distributions  $F \in \mathcal{L}_2(\mathbf{g})$  the semiparametric variance bound for estimation of  $\boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\mu})$  is  $\mathbf{G}'V\mathbf{G}$ , and the plug-in estimator  $\hat{\boldsymbol{\theta}} = \mathbf{g}(\hat{\boldsymbol{\mu}})$  is a semiparametrically efficient estimator of  $\boldsymbol{\theta}$ .

The result in Proposition 6.2 is quite general. Smooth functions of sample moments are efficient estimators for their population counterparts. This is a very powerful result, as most econometric estimators can be written (or approximated) as smooth functions of sample means.

## 6.28 Technical Proofs\*

In this section we provide proofs of some of the more technical points in the chapter. These proofs may only be of interest to more mathematically inclined students.

**Proof of Theorem 6.2:** Without loss of generality, we can assume  $\mathbb{E}(y_i) = 0$  by recentering  $y_i$  on its expectation.

We need to show that for all  $\delta > 0$  and  $\eta > 0$  there is some  $N < \infty$  so that for all  $n \geq N$ ,  $\mathbb{P}(|\bar{y}| > \delta) \leq \eta$ . Fix  $\delta$  and  $\eta$ . Set  $\varepsilon = \delta\eta/3$ . Pick  $C < \infty$  large enough so that

$$\mathbb{E}(|y_i| \mathbf{1}(|y_i| > C)) \leq \varepsilon \quad (6.36)$$

which is possible since  $y_i$  is uniformly integrable (or if  $y_i$  is i.i.d. and  $\mathbb{E}|y_i| < \infty$ ). Define the random variables

$$\begin{aligned} w_i &= y_i \mathbf{1}(|y_i| \leq C) - \mathbb{E}(y_i \mathbf{1}(|y_i| \leq C)) \\ z_i &= y_i \mathbf{1}(|y_i| > C) - \mathbb{E}(y_i \mathbf{1}(|y_i| > C)) \end{aligned}$$

so that

$$\bar{y} = \bar{w} + \bar{z}$$

and

$$\mathbb{E}|\bar{y}| \leq \mathbb{E}|\bar{w}| + \mathbb{E}|\bar{z}|. \quad (6.37)$$

We now show that sum of the expectations on the right-hand-side can be bounded below  $3\epsilon$ .

First, by the triangle inequality (B.1) and the expectation inequality (B.29),

$$\begin{aligned} \mathbb{E}|z_i| &= \mathbb{E}|y_i \mathbf{1}(|y_i| > C) - \mathbb{E}(y_i \mathbf{1}(|y_i| > C))| \\ &\leq \mathbb{E}|y_i \mathbf{1}(|y_i| > C)| + |\mathbb{E}(y_i \mathbf{1}(|y_i| > C))| \\ &\leq 2\mathbb{E}|y_i \mathbf{1}(|y_i| > C)| \\ &\leq 2\epsilon, \end{aligned} \quad (6.38)$$

and thus by the triangle inequality (B.1) and (6.38)

$$\mathbb{E}|\bar{z}| = \mathbb{E}\left|\frac{1}{n} \sum_{i=1}^n z_i\right| \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}|z_i| \leq 2\epsilon. \quad (6.39)$$

Second, by a similar argument

$$\begin{aligned} |w_i| &= |y_i \mathbf{1}(|y_i| \leq C) - \mathbb{E}(y_i \mathbf{1}(|y_i| \leq C))| \\ &\leq |y_i \mathbf{1}(|y_i| \leq C)| + |\mathbb{E}(y_i \mathbf{1}(|y_i| \leq C))| \\ &\leq 2|y_i \mathbf{1}(|y_i| \leq C)| \\ &\leq 2C \end{aligned} \quad (6.40)$$

where the final inequality is (6.36). Then by Jensen's inequality (B.26), the fact that the  $w_i$  are iid and mean zero, and (6.40),

$$(\mathbb{E}|\bar{w}|)^2 \leq \mathbb{E}(|\bar{w}|^2) = \frac{\mathbb{E}(w_i^2)}{n} \leq \frac{4C^2}{n} \leq \epsilon^2 \quad (6.41)$$

the final inequality holding for  $n \geq 4C^2/\epsilon^2 = 36C^2/\delta^2\eta^2$ . Equations (6.37), (6.39) and (6.41) together show that

$$\mathbb{E}|\bar{y}| \leq 3\epsilon \quad (6.42)$$

as desired.

Finally, by Markov's inequality (B.35) and (6.42),

$$\mathbb{P}(|\bar{y}| > \delta) \leq \frac{\mathbb{E}|\bar{y}|}{\delta} \leq \frac{3\epsilon}{\delta} = \eta,$$

the final equality by the definition of  $\epsilon$ . We have shown that for any  $\delta > 0$  and  $\eta > 0$  then for all  $n \geq 36C^2/\delta^2\eta^2$ ,  $\mathbb{P}(|\bar{y}| > \delta) \leq \eta$ , as needed. ■

**Proof of Theorem 6.5:** Assume  $\mathbb{E}|y_j| \leq C < \infty$  for  $j = 1, \dots, m$ . Applying the triangle inequality (B.9)

$$\mathbb{E}\|\mathbf{y}\| \leq \sum_{j=1}^m \mathbb{E}|y_j| \leq mC < \infty.$$

For the reverse inequality, the Euclidean norm of a vector is larger than the length of any individual component, so for any  $j$ ,  $|y_j| \leq \|\mathbf{y}\|$ . Thus, if  $\mathbb{E}\|\mathbf{y}\| < \infty$ , then  $\mathbb{E}|y_j| < \infty$  for  $j = 1, \dots, m$ . ■

**Proof of Theorem 6.7:** We present a case for the one-dimensional case. Fix  $\epsilon > 0$  and set  $J = 1/\epsilon$ . By the continuity of  $F(u)$  we can find points  $u_0 < u_1 < \dots < u_J$  with  $F(u_j) = j/J = j\epsilon$ . Since  $\mathbf{z}_n \xrightarrow{d} \mathbf{z}$  and  $J$  is fixed there is an  $n$  sufficiently large such that

$$\max_{j \leq J} |F_n(u_j) - F(u_j)| \leq \epsilon. \quad (6.43)$$

Since both  $F_n(u)$  and  $F(u)$  are monotonically increasing, for any  $u$  satisfying  $u_{j-1} \leq u \leq u_j$

$$F_n(u) - F(u) \leq F_n(u_j) - F(u_{j-1}) = F_n(u_j) - F(u_j) + \varepsilon \leq 2\varepsilon$$

where the final inequality is (6.43). Similarly,

$$F_n(u) - F(u) \geq F_n(u_{j-1}) - F(u_j) = F_n(u_{j-1}) - F(u_{j-1}) - \varepsilon \geq -2\varepsilon.$$

We have shown that for any  $u$ ,  $|F_n(u) - F(u)| \leq 2\varepsilon$ . Since  $\varepsilon$  is arbitrary this shows  $\sup_u |F_n(u) - F(u)| \rightarrow 0$  as  $n \rightarrow \infty$ . ■

**Proof of Theorem 6.10:** By Lévy's Continuity Theorem (Theorem 6.9),  $z_n \xrightarrow{d} z$  if and only if  $\mathbb{E}(\exp(is'z_n)) \rightarrow \mathbb{E}(\exp(is'z))$  for every  $s \in \mathbb{R}^k$ . We can write  $s = t\lambda$  where  $t \in \mathbb{R}$  and  $\lambda \in \mathbb{R}^k$  with  $\lambda'\lambda = 1$ . Thus the convergence holds if and only if  $\mathbb{E}(\exp(it\lambda'z_n)) \rightarrow \mathbb{E}(\exp(it\lambda'z))$  for every  $t \in \mathbb{R}$  and  $\lambda \in \mathbb{R}^k$  with  $\lambda'\lambda = 1$ . Again by Lévy's Continuity Theorem, this holds if and only if  $\lambda'z_n \xrightarrow{d} \lambda'z$  for every  $\lambda \in \mathbb{R}^k$  and with  $\lambda'\lambda = 1$ . ■

**Proof of Theorem 6.11:** The moment bound  $\mathbb{E}(y_i^2) < \infty$  is sufficient to guarantee that  $\mu$  and  $\sigma^2$  are well defined and finite. Without loss of generality, it is sufficient to consider the case  $\mu = 0$ .

Our proof method is to calculate the moment generating function of  $\sqrt{n}\bar{y}_n$  and show that it converges pointwise to  $\exp(t^2\sigma^2/2)$ , the MGF function of  $N(0, \sigma^2)$ . With a slight increase of notation this extends to the characteristic function. By Lévy's Continuity Theorem (Theorem 6.9) this implies  $\sqrt{n}\bar{y}_n \xrightarrow{d} N(0, \sigma^2)$ .

Let  $K(t) = \log(\mathbb{E}\exp(ty_i))$  denote the cumulant generating function of  $y_i$  (see Section 2.33). From the results in Section 2.33 we know  $K(0) = 0$ ,  $K^{(1)}(0) = \mu = 0$  and  $K^{(2)}(0) = \sigma^2$ . Since the second moment of  $y_i$  is finite,  $K^{(2)}(t)$  is continuous at  $t = 0$ . Thus we can apply a second order Taylor series expansion about 0 to find that for  $t$  sufficiently small

$$\begin{aligned} K(t) &= K(0) + K^{(1)}(0)t + \frac{1}{2}K^{(2)}(t^*)t^2 \\ &= \frac{1}{2}K^{(2)}(t^*)t^2 \end{aligned} \tag{6.44}$$

where  $t^*$  lies on the line segment joining 0 and  $t$ , and  $K^{(2)}(t^*)$  approaches  $\sigma^2$  as  $t \rightarrow 0$ .

We now compute  $C_n(t) = \mathbb{E}\exp(t\sqrt{n}\bar{y}_n)$ , the MGF of  $\sqrt{n}\bar{y}_n$ . By the properties of the exponential function, the independence of the  $y_i$ , and the definition of  $K(t)$

$$\begin{aligned} \log C_n(t) &= \log \mathbb{E}\left(\exp\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n ty_i\right)\right) \\ &= \log \mathbb{E}\left(\prod_{i=1}^n \exp\left(\frac{1}{\sqrt{n}} ty_i\right)\right) \\ &= \log \prod_{i=1}^n \mathbb{E}\left(\exp\left(\frac{1}{\sqrt{n}} ty_i\right)\right) \\ &= \sum_{i=1}^n \log \mathbb{E}\left(\exp\left(\frac{1}{\sqrt{n}} ty_i\right)\right) \\ &= nK\left(\frac{t}{\sqrt{n}}\right) \\ &= \frac{1}{2}K^{(2)}(t_n)t^2 \end{aligned}$$

where the final equality is (6.44) with  $t_n \rightarrow 0$  as  $n \rightarrow \infty$ . Since  $t_n$  is bounded we deduce that  $K^{(2)}(t_n) \rightarrow K^{(2)}(0) = \sigma^2$ . Hence, as  $n \rightarrow \infty$ ,

$$\log C_n(t) \rightarrow \frac{1}{2}\sigma^2 t^2$$

and

$$C_n(t) \rightarrow \exp\left(\frac{1}{2}\sigma^2 t^2\right)$$

which is the MGF of the  $N(0, \sigma^2)$  distribution, as shown in Exercise 5.5. This completes the proof. ■

**Proof of Theorem 6.13:** Suppose that  $\sigma^2 = 0$ . Then  $\text{var}(\sqrt{n}(\bar{y} - \mathbb{E}(\bar{y}))) = \bar{\sigma}_n^2 \rightarrow \sigma^2 = 0$  so  $\sqrt{n}(\bar{y} - \mathbb{E}(\bar{y})) \xrightarrow{p} 0$  and hence  $\sqrt{n}(\bar{y} - \mathbb{E}(\bar{y})) \xrightarrow{d} 0$ . The random variable  $N(0, \sigma^2) = N(0, 0)$  is 0 with probability 1, so this is  $\sqrt{n}(\bar{y} - \mathbb{E}(\bar{y})) \xrightarrow{d} N(0, \sigma^2)$  as stated.

Now suppose that  $\sigma^2 > 0$ . This implies (6.8). Together with (6.7) this implies Lyapunov's condition, and hence Lindeberg's condition, and hence Theorem 6.12, which states

$$\frac{\sqrt{n}(\bar{y} - \mathbb{E}(\bar{y}))}{\bar{\sigma}_n^{1/2}} \xrightarrow{d} N(0, 1).$$

Combined with (6.9) we deduce  $\sqrt{n}(\bar{y} - \mathbb{E}(\bar{y})) \xrightarrow{d} N(0, \sigma^2)$  as stated. ■

**Proof of Theorem 6.14:** Set  $\lambda \in \mathbb{R}^k$  with  $\lambda' \lambda = 1$  and define  $u_i = \lambda' (\mathbf{y}_i - \boldsymbol{\mu})$ . The  $u_i$  are i.i.d. with  $\mathbb{E}(u_i^2) = \lambda' V \lambda < \infty$ . By Theorem 6.11,

$$\lambda' \sqrt{n}(\bar{y} - \boldsymbol{\mu}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n u_i \xrightarrow{d} N(0, \lambda' V \lambda)$$

Notice that if  $\mathbf{z} \sim N(\mathbf{0}, V)$  then  $\lambda' \mathbf{z} \sim N(\mathbf{0}, \lambda' V \lambda)$ . Thus

$$\lambda' \sqrt{n}(\bar{y} - \boldsymbol{\mu}) \xrightarrow{d} \lambda' \mathbf{z}.$$

Since this holds for all  $\lambda$ , the conditions of Theorem 6.10 are satisfied and we deduce that

$$\sqrt{n}(\bar{y} - \boldsymbol{\mu}) \xrightarrow{d} \mathbf{z} \sim N(\mathbf{0}, V)$$

as stated. ■

**Proof of Theorem 6.15:** Set  $\lambda \in \mathbb{R}^k$  with  $\lambda' \lambda = 1$  and define  $u_{ni} = \lambda' \bar{V}_n^{-1/2} \mathbf{y}_{ni}$ . Notice that  $u_{ni}$  are independent and has variance  $\sigma_{ni}^2 = \lambda' \bar{V}_n^{-1/2} V_{ni} \bar{V}_n^{-1/2} \lambda$  and  $\bar{\sigma}_n^2 = \sum_{i=1}^{r_n} \sigma_{ni}^2 = 1$ . It is sufficient to verify (6.5). By the Schwarz inequality (B.12),

$$\begin{aligned} u_{ni}^2 &= (\lambda' \bar{V}_n^{-1/2} \mathbf{y}_{ni})^2 \\ &\leq \lambda' \bar{V}_n^{-1} \lambda \|\mathbf{y}_{ni}\|^2 \\ &\leq \frac{\|\mathbf{y}_{ni}\|^2}{\lambda_{\min}(\bar{V}_n)} \\ &= \frac{\|\mathbf{y}_{ni}\|^2}{v_n^2}. \end{aligned}$$

Then

$$\begin{aligned} \frac{1}{\bar{\sigma}_n^2} \sum_{i=1}^{r_n} \mathbb{E}(u_{ni}^2 \mathbf{1}(u_{ni}^2 \geq \varepsilon \bar{\sigma}_n^2)) &= \sum_{i=1}^{r_n} \mathbb{E}(u_{ni}^2 \mathbf{1}(u_{ni}^2 \geq \varepsilon)) \\ &\leq \frac{1}{v_n^2} \sum_{i=1}^{r_n} \mathbb{E}(\|\mathbf{y}_{ni}\|^2 \mathbf{1}(\|\mathbf{y}_{ni}\|^2 \geq \varepsilon v_n^2)) \\ &\rightarrow 0 \end{aligned}$$

by (6.14). This establishes (6.5). We deduce from Theorem 6.12 that

$$\sum_{i=1}^{r_n} u_{ni} = \boldsymbol{\lambda}' \bar{V}_n^{-1/2} \sum_{i=1}^{r_n} \mathbf{y}_{ni} \xrightarrow{d} N(0, 1) = \boldsymbol{\lambda}' \mathbf{z}$$

where  $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}_k)$ . Since this holds for all  $\boldsymbol{\lambda}$ , the conditions of Theorem 6.10 are satisfied and we deduce that

$$\bar{V}_n^{-1/2} \sum_{i=1}^{r_n} \mathbf{y}_{ni} \xrightarrow{d} N(0, \mathbf{I}_k)$$

as stated. ■

**Proof of Theorem 6.16:** Set  $\boldsymbol{\lambda} \in \mathbb{R}^k$  with  $\boldsymbol{\lambda}' \boldsymbol{\lambda} = 1$  and define  $u_{ni} = \boldsymbol{\lambda}' \mathbf{y}_{ni}$ . Using the Schwarz inequality (B.12) and (6.15) we obtain

$$\sup_{n,i} \mathbb{E}(|u_{ni}|^{2+\delta}) = \sup_{n,i} \mathbb{E}(|\boldsymbol{\lambda}' \mathbf{y}_{ni}|^{2+\delta}) \leq \|\boldsymbol{\lambda}\|^{2+\delta} \sup_{n,i} \mathbb{E}(\|\mathbf{y}_{ni}\|^{2+\delta}) = \sup_{n,i} \mathbb{E}(\|\mathbf{y}_{ni}\|^{2+\delta}) < \infty$$

which is (6.7). Notice that

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}(u_{ni}^2) = \boldsymbol{\lambda}' \frac{1}{n} \sum_{i=1}^n V_{ni} \boldsymbol{\lambda} = \boldsymbol{\lambda}' \bar{V}_n \boldsymbol{\lambda} \rightarrow \boldsymbol{\lambda}' \mathbf{V} \boldsymbol{\lambda}$$

which is (6.9). Since the  $u_{ni}$  are independent, by Theorem 6.14,

$$\boldsymbol{\lambda}' \sqrt{n} \bar{\mathbf{y}} = \frac{1}{\sqrt{n}} \sum_{i=1}^n u_{ni} \xrightarrow{d} N(0, \boldsymbol{\lambda}' \mathbf{V} \boldsymbol{\lambda}) = \boldsymbol{\lambda}' \mathbf{z}$$

where  $\mathbf{z} \sim N(\mathbf{0}, \mathbf{V})$ . Since this holds for all  $\boldsymbol{\lambda}$ , the conditions of Theorem 6.10 are satisfied and we deduce that

$$\sqrt{n} \bar{\mathbf{y}} \xrightarrow{d} N(\mathbf{0}, \mathbf{V})$$

as stated. ■

**Proof of Theorem 6.19:** Fix  $\varepsilon > 0$ . Continuity of  $\mathbf{g}(\mathbf{u})$  at  $\mathbf{c}$  means that there exists  $\delta > 0$  such that  $\|\mathbf{u} - \mathbf{c}\| \leq \delta$  implies  $\|\mathbf{g}(\mathbf{u}) - \mathbf{g}(\mathbf{c})\| \leq \varepsilon$ . Evaluated at  $\mathbf{u} = \mathbf{z}_n$  we find

$$\mathbb{P}(\|\mathbf{g}(\mathbf{z}_n) - \mathbf{g}(\mathbf{c})\| \leq \varepsilon) \geq \mathbb{P}(\|\mathbf{z}_n - \mathbf{c}\| \leq \delta) \rightarrow 1$$

where the final convergence holds as  $n \rightarrow \infty$  by the assumption that  $\mathbf{z}_n \xrightarrow{p} \mathbf{c}$ . This implies  $\mathbf{g}(\mathbf{z}_n) \xrightarrow{p} \mathbf{g}(\mathbf{c})$ .

■

**Proof of Theorem 6.23:** By a vector Taylor series expansion, for each element of  $\mathbf{g}$ ,

$$g_j(\boldsymbol{\theta}_n) = g_j(\boldsymbol{\theta}) + g_{j\boldsymbol{\theta}}(\boldsymbol{\theta}_{jn}^*) (\boldsymbol{\theta}_n - \boldsymbol{\theta})$$

where  $\boldsymbol{\theta}_{jn}^*$  lies on the line segment between  $\boldsymbol{\theta}_n$  and  $\boldsymbol{\theta}$  and therefore converges in probability to  $\boldsymbol{\theta}$ . It follows that  $a_{jn} = g_{j\boldsymbol{\theta}}(\boldsymbol{\theta}_{jn}^*) - g_{j\boldsymbol{\theta}} \xrightarrow{p} 0$ . Stacking across elements of  $\mathbf{g}$ , we find

$$\sqrt{n}(\mathbf{g}(\boldsymbol{\theta}_n) - \mathbf{g}(\boldsymbol{\theta})) = (\mathbf{G} + a_n)' \sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathbf{G}' \boldsymbol{\xi}. \quad (6.45)$$

The convergence is by Theorem 6.21, as  $\mathbf{G} + a_n \xrightarrow{d} \mathbf{G}$ ,  $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{\xi}$ , and their product is continuous. This establishes (6.19)

When  $\boldsymbol{\xi} \sim N(\mathbf{0}, \mathbf{V})$ , the right-hand-side of (6.45) equals

$$\mathbf{G}' \boldsymbol{\xi} = \mathbf{G}' N(\mathbf{0}, \mathbf{V}) = N(\mathbf{0}, \mathbf{G}' \mathbf{V} \mathbf{G})$$

establishing (6.20). ■

**Proof of Theorem 6.38:** First consider (6.31). Take any  $\delta > 0$ . The event  $\{\max_{1 \leq i \leq n} |y_i| > \delta n^{1/r}\}$  means that at least one of the  $|y_i|$  exceeds  $\delta n^{1/r}$ , which is the same as the event  $\bigcup_{i=1}^n \{|y_i| > \delta n^{1/r}\}$  or equivalently  $\bigcup_{i=1}^n \{|y_i|^r > \delta^r n\}$ . Since the probability of the union of events is smaller than the sum of the probabilities,

$$\begin{aligned} \mathbb{P}\left(n^{-1/r} \max_{1 \leq i \leq n} |y_i| > \delta\right) &= \mathbb{P}\left(\bigcup_{i=1}^n \{|y_i|^r > \delta^r n\}\right) \\ &\leq \sum_{i=1}^n \mathbb{P}(|y_i|^r > n\delta^r) \\ &\leq \frac{1}{n\delta^r} \sum_{i=1}^n \mathbb{E}(|y_i|^r \mathbf{1}(|y_i|^r > n\delta^r)) \\ &= \frac{1}{\delta^r} \max_{i \leq n} \mathbb{E}(|y_i|^r \mathbf{1}(|y_i|^r > n\delta^r)) \end{aligned}$$

where the second inequality is the strong form of Markov's inequality (Theorem B.36) and the final equality is since the  $y_i$  have identical distributions. Since  $|y_i|^r$  is uniformly integrable, this converges to zero as  $n\delta^r \rightarrow \infty$ . This establishes (6.31).

Now consider (6.32). Take any  $\delta > 0$  and pick  $n$  large enough so that  $(\log n)^\eta t\delta \geq 1$ . By a similar calculation

$$\begin{aligned} \mathbb{P}\left((\log n)^{-(1+\eta)} \max_{1 \leq i \leq n} |y_i| > \delta\right) &= \mathbb{P}\left(\bigcup_{i=1}^n \{\exp|ty_i| > \exp((\log n)^{1+\eta} t\delta)\}\right) \\ &\leq \sum_{i=1}^n \mathbb{P}(\exp|ty_i| > n) \\ &\leq \max_{i \leq n} \mathbb{E}(\exp|ty_i| \mathbf{1}(\exp|ty_i| > n)) \end{aligned}$$

where the second line uses  $\exp((\log n)^{1+\eta} t\delta) \geq \exp(\log n) = n$ . Since  $\exp|ty_i|$  is uniformly integrable, this converges to zero as  $n \rightarrow \infty$ . This establishes (6.32). ■

## Exercises

**Exercise 6.1** For the following sequences, show  $a_n \rightarrow 0$  as  $n \rightarrow \infty$

(a)  $a_n = 1/n$

(b)  $a_n = \frac{1}{n} \sin\left(\frac{\pi}{2} n\right)$

**Exercise 6.2** Does the sequence  $a_n = \sin\left(\frac{\pi}{2} n\right)$  converge? Find the liminf and limsup as  $n \rightarrow \infty$ .

**Exercise 6.3** A weighted sample mean takes the form  $\bar{y}^* = \frac{1}{n} \sum_{i=1}^n w_i y_i$  for some non-negative constants  $w_i$  satisfying  $\frac{1}{n} \sum_{i=1}^n w_i = 1$ . Assume  $y_i$  is iid.

(a) Show that  $\bar{y}^*$  is unbiased for  $\mu = \mathbb{E}(y_i)$ .

(b) Calculate  $\text{var}(\bar{y}^*)$ .

(c) Show that a sufficient condition for  $\bar{y}^* \xrightarrow{p} \mu$  is that  $\frac{1}{n^2} \sum_{i=1}^n w_i^2 \rightarrow 0$ .

(d) Show that a sufficient condition for the condition in part 3 is  $\max_{i \leq n} w_i = o(n)$ .

**Exercise 6.4** Consider a random variable  $X_n$  with the probability distribution

$$X_n = \begin{cases} -n & \text{with probability } 1/n \\ 0 & \text{with probability } 1 - 2/n \\ n & \text{with probability } 1/n \end{cases}$$

(a) Does  $X_n \rightarrow_p 0$  as  $n \rightarrow \infty$ ?

(b) Calculate  $\mathbb{E}(X_n)$

(c) Calculate  $\text{var}(X_n)$

(d) Now suppose the distribution is

$$X_n = \begin{cases} 0 & \text{with probability } 1 - n \\ n & \text{with probability } 1/n \end{cases}$$

Calculate  $\mathbb{E}(X_n)$

(e) Conclude that  $X_n \rightarrow_p 0$  as  $n \rightarrow \infty$  and  $\mathbb{E}(X_n) \rightarrow 0$  are unrelated.

**Exercise 6.5** A weighted sample mean takes the form  $\bar{y}^* = \frac{1}{n} \sum_{i=1}^n w_i y_i$  for some non-negative constants  $w_i$  satisfying  $\frac{1}{n} \sum_{i=1}^n w_i = 1$ . Assume  $y_i$  is iid.

(a) Show that  $\bar{y}^*$  is unbiased for  $\mu = \mathbb{E}(y_i)$ .

(b) Calculate  $\text{var}(\bar{y}^*)$ .

(c) Show that a sufficient condition for  $\bar{y}^* \xrightarrow{p} \mu$  is that  $\frac{1}{n^2} \sum_{i=1}^n w_i^2 \rightarrow 0$ .

(d) Show that a sufficient condition for the condition in part c is  $\max_{i \leq n} w_i / n \rightarrow 0$ .

**Exercise 6.6** Take a random sample  $\{y_1, \dots, y_n\}$ . Which statistics converge in probability by the weak law of large numbers and continuous mapping theorem, assuming the moment exists?

(a)  $\frac{1}{n} \sum_{i=1}^n y_i^2$ .

- (b)  $\frac{1}{n} \sum_{i=1}^n y_i^3$ .
- (c)  $\max_{i \leq n} y_i$ .
- (d)  $\frac{1}{n} \sum_{i=1}^n y_i^2 - \left( \frac{1}{n} \sum_{i=1}^n y_i \right)^2$ .
- (e)  $\frac{\sum_{i=1}^n y_i^2}{\sum_{i=1}^n y_i}$  assuming  $\mathbb{E}(y_i) > 0$ .
- (f)  $\mathbf{1}\left(\frac{1}{n} \sum_{i=1}^n y_i > 0\right)$  where  $\mathbf{1}(a)$  is the indicator function.

**Exercise 6.7** Take a random sample  $\{X_1, \dots, X_n\}$  where  $X > 0$ . Consider the sample geometric mean

$$\hat{\mu} = \left( \prod_{i=1}^n X_i \right)^{1/n}$$

and population geometric mean

$$\mu = \exp(\mathbb{E}(\log X))$$

Assuming  $\mu$  is finite, show that  $\hat{\mu} \xrightarrow{p} \mu$  as  $n \rightarrow \infty$ .

**Exercise 6.8** Take a random variable  $Z$  such that  $\mathbb{E}(Z) = 0$  and  $\text{var}(Z) = 1$ . Use Chebyshev's inequality to find a  $\delta$  such that  $\mathbb{P}(|Z| > \delta) \leq 0.05$ . Contrast this with the exact  $\delta$  which solves  $\mathbb{P}(|Z| > \delta) = 0.05$  when  $Z \sim N(0, 1)$ . Comment on the difference.

**Exercise 6.9** Find the moment estimator  $\hat{\mu}_3$  of  $\mu_3 = \mathbb{E}(y_i^3)$  and show that  $\sqrt{n}(\hat{\mu}_3 - \mu_3) \xrightarrow{d} N(0, v^2)$  for some  $v^2$ . Write  $v^2$  as a function of the moments of  $y_i$ .

**Exercise 6.10** Suppose  $z_n \xrightarrow{p} c$  as  $n \rightarrow \infty$ . Show that  $z_n^2 \xrightarrow{p} c^2$  as  $n \rightarrow \infty$  using the definition of convergence in probability, but not appealing to the CMT.

**Exercise 6.11** Let  $\mu_k = \mathbb{E}(y^k)$  for some integer  $k \geq 1$ .

- (a) Write down the natural moment estimator  $\hat{\mu}_k$  of  $\mu_k$ .
- (b) Find the asymptotic distribution of  $\sqrt{n}(\hat{\mu}_k - \mu_k)$  as  $n \rightarrow \infty$ . (Assume  $\mathbb{E}(X^{2k}) < \infty$ .)

**Exercise 6.12** Let  $m_k = (\mathbb{E}(y^k))^{1/k}$  for some integer  $k \geq 1$ .

- (a) Write down an estimator  $\hat{m}_k$  of  $m_k$ .
- (b) Find the asymptotic distribution of  $\sqrt{n}(\hat{m}_k - m_k)$  as  $n \rightarrow \infty$ .

**Exercise 6.13** Suppose  $\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} N(0, v^2)$  and set  $\beta = \mu^2$  and  $\hat{\beta} = \hat{\mu}^2$ .

- (a) Use the Delta Method to obtain an asymptotic distribution for  $\sqrt{n}(\hat{\beta} - \beta)$ .
- (b) Now suppose  $\mu = 0$ . Describe what happens to the asymptotic distribution from the previous part.
- (c) Improve on the previous answer. Under the assumption  $\mu = 0$ , find the asymptotic distribution for  $n\hat{\beta} = n\hat{\mu}^2$ .
- (d) Comment on the differences between the answers in parts 1 and 3.

**Exercise 6.14** Let  $y$  be distributed Bernoulli  $P(y = 1) = p$  and  $P(y = 0) = 1 - p$  for some unknown  $0 < p < 1$ .

- (a) Show that  $p = \mathbb{E}(y)$ .
- (b) Write down the natural moment estimator  $\hat{p}$  of  $p$ .
- (c) Find  $\text{var}(\hat{p})$ .
- (d) Find the asymptotic distribution of  $\sqrt{n}(\hat{p} - p)$  as  $n \rightarrow \infty$ .

# Chapter 7

## Asymptotic Theory for Least Squares

### 7.1 Introduction

It turns out that the asymptotic theory of least-squares estimation applies equally to the projection model and the linear CEF model, and therefore the results in this chapter will be stated for the broader projection model described in Section 2.18. Recall that the model is

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + e_i$$

for  $i = 1, \dots, n$ , where the linear projection coefficient  $\boldsymbol{\beta}$  is

$$\boldsymbol{\beta} = (\mathbb{E}(\mathbf{x}_i \mathbf{x}'_i))^{-1} \mathbb{E}(\mathbf{x}_i y_i).$$

Maintained assumptions in this chapter will be random sampling (Assumption 1.2) and finite second moments (Assumption 2.1). We restate these conditions here for clarity.

#### Assumption 7.1

1. The observations  $(y_i, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ , are independent and identically distributed.
2.  $\mathbb{E}(y^2) < \infty$ .
3.  $\mathbb{E}(\|\mathbf{x}\|^2) < \infty$ .
4.  $\mathbf{Q}_{\mathbf{x}\mathbf{x}} = \mathbb{E}(\mathbf{x}\mathbf{x}')$  is positive definite.

The distributional results will require a strengthening of these assumptions to finite fourth moments. We discuss the specific conditions in Section 7.3.

### 7.2 Consistency of Least-Squares Estimator

In this section we use the weak law of large numbers (WLLN, Theorem 6.2 and Theorem 6.6) and continuous mapping theorem (CMT, Theorem 6.19) to show that the least-squares estimator  $\hat{\boldsymbol{\beta}}$  is consistent for the projection coefficient  $\boldsymbol{\beta}$ .

This derivation is based on three key components. First, the OLS estimator can be written as a continuous function of a set of sample moments. Second, the WLLN shows that sample moments converge in probability to population moments. And third, the CMT states that continuous functions preserve convergence in probability. We now explain each step in brief and then in greater detail.

First, observe that the OLS estimator

$$\hat{\beta} = \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i y_i \right) = \hat{\mathbf{Q}}_{xx}^{-1} \hat{\mathbf{Q}}_{xy}$$

is a function of the sample moments  $\hat{\mathbf{Q}}_{xx} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i$  and  $\hat{\mathbf{Q}}_{xy} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i y_i$ .

Second, by an application of the WLLN these sample moments converge in probability to the population moments. Specifically, the fact that  $(y_i, \mathbf{x}_i)$  are mutually independent and identically distributed implies that any function of  $(y_i, \mathbf{x}_i)$  is iid, including  $\mathbf{x}_i \mathbf{x}'_i$  and  $\mathbf{x}_i y_i$ . These variables also have finite expectations under Assumption 7.1. Under these conditions, the WLLN (Theorem 6.6) implies that as  $n \rightarrow \infty$ ,

$$\hat{\mathbf{Q}}_{xx} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \xrightarrow{p} \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i) = \mathbf{Q}_{xx} \quad (7.1)$$

and

$$\hat{\mathbf{Q}}_{xy} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i y_i \xrightarrow{p} \mathbb{E}(\mathbf{x}_i y_i) = \mathbf{Q}_{xy}.$$

Third, the CMT (Theorem 6.19) allows us to combine these equations to show that  $\hat{\beta}$  converges in probability to  $\beta$ . Specifically, as  $n \rightarrow \infty$ ,

$$\begin{aligned} \hat{\beta} &= \hat{\mathbf{Q}}_{xx}^{-1} \hat{\mathbf{Q}}_{xy} \\ &\xrightarrow{p} \mathbf{Q}_{xx}^{-1} \mathbf{Q}_{xy} \\ &= \beta. \end{aligned} \quad (7.2)$$

We have shown that  $\hat{\beta} \xrightarrow{p} \beta$ , as  $n \rightarrow \infty$ . In words, the OLS estimator converges in probability to the projection coefficient vector  $\beta$  as the sample size  $n$  gets large.

To fully understand the application of the CMT we walk through it in detail. We can write

$$\hat{\beta} = \mathbf{g}(\hat{\mathbf{Q}}_{xx}, \hat{\mathbf{Q}}_{xy})$$

where  $\mathbf{g}(\mathbf{A}, \mathbf{b}) = \mathbf{A}^{-1} \mathbf{b}$  is a function of  $\mathbf{A}$  and  $\mathbf{b}$ . The function  $\mathbf{g}(\mathbf{A}, \mathbf{b})$  is a continuous function of  $\mathbf{A}$  and  $\mathbf{b}$  at all values of the arguments such that  $\mathbf{A}^{-1}$  exists. Assumption 7.1 specifies that  $\mathbf{Q}_{xx}^{-1}$  exists and thus  $\mathbf{g}(\mathbf{A}, \mathbf{b})$  is continuous at  $\mathbf{A} = \mathbf{Q}_{xx}$ . This justifies the application of the CMT in (7.2).

For a slightly different demonstration of (7.2), recall that (4.6) implies that

$$\hat{\beta} - \beta = \hat{\mathbf{Q}}_{xx}^{-1} \hat{\mathbf{Q}}_{xe} \quad (7.3)$$

where

$$\hat{\mathbf{Q}}_{xe} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i e_i.$$

The WLLN and (2.23) imply

$$\hat{\mathbf{Q}}_{xe} \xrightarrow{p} \mathbb{E}(\mathbf{x}_i e_i) = \mathbf{0}.$$

Therefore

$$\begin{aligned} \hat{\beta} - \beta &= \hat{\mathbf{Q}}_{xx}^{-1} \hat{\mathbf{Q}}_{xe} \\ &\xrightarrow{p} \mathbf{Q}_{xx}^{-1} \mathbf{0} \\ &= \mathbf{0} \end{aligned}$$

which is the same as  $\hat{\beta} \xrightarrow{p} \beta$ .

**Theorem 7.1 Consistency of Least-Squares**

Under Assumption 7.1,  $\hat{\mathbf{Q}}_{xx} \xrightarrow{p} \mathbf{Q}_{xx}$ ,  $\hat{\mathbf{Q}}_{xy} \xrightarrow{p} \mathbf{Q}_{xy}$ ,  $\hat{\mathbf{Q}}_{xx}^{-1} \xrightarrow{p} \mathbf{Q}_{xx}^{-1}$ ,  $\hat{\mathbf{Q}}_{xe} \xrightarrow{p} \mathbf{0}$ , and  $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$  as  $n \rightarrow \infty$ .

Theorem 7.1 states that the OLS estimator  $\hat{\boldsymbol{\beta}}$  converges in probability to  $\boldsymbol{\beta}$  as  $n$  increases, and thus  $\hat{\boldsymbol{\beta}}$  is consistent for  $\boldsymbol{\beta}$ . In the stochastic order notation, Theorem 7.1 can be equivalently written as

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + o_p(1). \quad (7.4)$$

To illustrate the effect of sample size on the least-squares estimator consider the least-squares regression

$$\ln(Wage_i) = \beta_1 Education_i + \beta_2 Experience_i + \beta_3 Experience_i^2 + \beta_4 + e_i.$$

We use the sample of 24,344 white men from the March 2009 CPS. Randomly sorting the observations, and sequentially estimating the model by least-squares, starting with the first 5 observations, and continuing until the full sample is used, the sequence of estimates are displayed in Figure 7.1. You can see how the least-squares estimate changes with the sample size, but as the number of observations increases it settles down to the full-sample estimate  $\hat{\beta}_1 = 0.114$ .

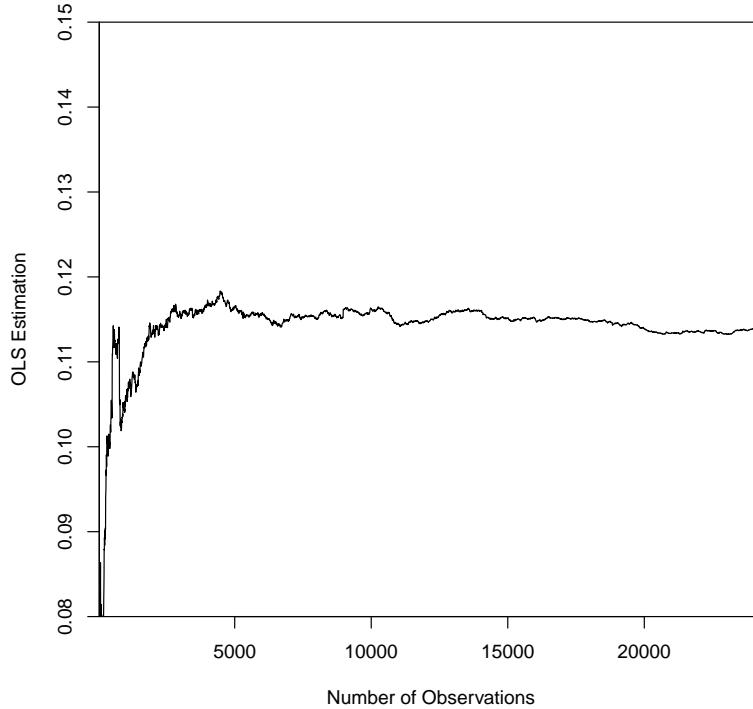


Figure 7.1: The Least-Squares Estimator  $\hat{\boldsymbol{\beta}}_1$  as a Function of Sample Size  $n$

### 7.3 Asymptotic Normality

We started this chapter discussing the need for an approximation to the distribution of the OLS estimator  $\hat{\boldsymbol{\beta}}$ . In Section 7.2 we showed that  $\hat{\boldsymbol{\beta}}$  converges in probability to  $\boldsymbol{\beta}$ . Consistency is a good first step,

but in itself does not describe the distribution of the estimator. In this section we derive an approximation typically called the **asymptotic distribution**.

The derivation starts by writing the estimator as a function of sample moments. One of the moments must be written as a sum of zero-mean random vectors and normalized so that the central limit theorem can be applied. The steps are as follows.

Take equation (7.3) and multiply it by  $\sqrt{n}$ . This yields the expression

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_i e_i \right). \quad (7.5)$$

This shows that the normalized and centered estimator  $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$  is a function of the sample average  $\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i$  and the normalized sample average  $\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_i e_i$ . Furthermore, the latter has mean zero so the central limit theorem (CLT, Theorem 6.11) applies.

The product  $\mathbf{x}_i e_i$  is i.i.d. (since the observations  $(y_i, \mathbf{x}_i)$  are i.i.d.) and mean zero (since  $\mathbb{E}(\mathbf{x}_i e_i) = \mathbf{0}$ ). Define the  $k \times k$  covariance matrix

$$\boldsymbol{\Omega} = \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i e_i^2).$$

The CLT requires the elements of  $\boldsymbol{\Omega}$  to be finite, written  $\boldsymbol{\Omega} < \infty$ . This requires a strengthening of Assumption 7.1. We state the required conditions here.

### Assumption 7.2

1. The observations  $(y_i, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ , are independent and identically distributed.
2.  $\mathbb{E}(y^4) < \infty$ .
3.  $\mathbb{E}(\|\mathbf{x}\|^4) < \infty$ .
4.  $\mathbf{Q}_{xx} = \mathbb{E}(\mathbf{x}\mathbf{x}')$  is positive definite.

Assumption 7.2 implies that  $\boldsymbol{\Omega} < \infty$ . To see this, take the  $j\ell^{th}$  element of  $\boldsymbol{\Omega}$ ,  $\mathbb{E}(x_{ji} x_{\ell i} e_i^2)$ . By the expectation inequality (B.29), the  $j\ell^{th}$  element of  $\boldsymbol{\Omega}$  is bounded by

$$|\mathbb{E}(x_{ji} x_{\ell i} e_i^2)| \leq \mathbb{E}|x_{ji} x_{\ell i} e_i^2| = \mathbb{E}(|x_{ji}| |x_{\ell i}| e_i^2).$$

By two applications of the Cauchy-Schwarz inequality (B.31), this is smaller than

$$(\mathbb{E}(x_{ji}^2 x_{\ell i}^2))^{1/2} (\mathbb{E}(e_i^4))^{1/2} \leq (\mathbb{E}(x_{ji}^4))^{1/4} (\mathbb{E}(x_{\ell i}^4))^{1/4} (\mathbb{E}(e_i^4))^{1/2} < \infty$$

where the finiteness holds under Assumption 7.2.2 and 7.2.3. Thus  $\boldsymbol{\Omega} < \infty$ .

An alternative way to show that the elements of  $\boldsymbol{\Omega}$  are finite is by using a matrix norm  $\|\cdot\|$  (See Appendix A.23). Then by the expectation inequality, the Cauchy-Schwarz inequality, and Assumption 7.2

$$\|\boldsymbol{\Omega}\| \leq \mathbb{E}\|\mathbf{x}_i \mathbf{x}'_i e_i^2\| = \mathbb{E}(\|\mathbf{x}_i\|^2 e_i^2) \leq (\mathbb{E}\|\mathbf{x}_i\|^4)^{1/2} (\mathbb{E}(e_i^4))^{1/2} < \infty.$$

This is a more compact argument (often described as more *elegant*) but such manipulations should not be done without understanding the notation and the applicability of each step of the argument.

Regardless, the finiteness of the covariance matrix means that we can then apply the multivariate CLT (Theorem 6.14).

**Theorem 7.2** Under Assumption 7.2,

$$\Omega < \infty \quad (7.6)$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_i e_i \xrightarrow{d} N(\mathbf{0}, \Omega) \quad (7.7)$$

as  $n \rightarrow \infty$ .

Putting together (7.1), (7.5), and (7.7),

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta) &\xrightarrow{d} Q_{xx}^{-1} N(\mathbf{0}, \Omega) \\ &= N(\mathbf{0}, Q_{xx}^{-1} \Omega Q_{xx}^{-1}) \end{aligned}$$

as  $n \rightarrow \infty$ , where the final equality follows from the property that linear combinations of normal vectors are also normal (Theorem 5.4).

We have derived the asymptotic normal approximation to the distribution of the least-squares estimator.

**Theorem 7.3 Asymptotic Normality of Least-Squares Estimator**

Under Assumption 7.2, as  $n \rightarrow \infty$

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(\mathbf{0}, V_\beta)$$

where

$$V_\beta = Q_{xx}^{-1} \Omega Q_{xx}^{-1}, \quad (7.8)$$

$$Q_{xx} = \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i), \text{ and } \Omega = \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i e_i^2).$$

In the stochastic order notation, Theorem 7.3 implies that

$$\hat{\beta} = \beta + O_p(n^{-1/2})$$

which is stronger than (7.4).

The matrix  $V_\beta = Q_{xx}^{-1} \Omega Q_{xx}^{-1}$  is the variance of the asymptotic distribution of  $\sqrt{n}(\hat{\beta} - \beta)$ . Consequently,  $V_\beta$  is often referred to as the **asymptotic covariance matrix** of  $\hat{\beta}$ . The expression  $V_\beta = Q_{xx}^{-1} \Omega Q_{xx}^{-1}$  is called a **sandwich form**, as the matrix  $\Omega$  is sandwiched between two copies of  $Q_{xx}^{-1}$ .

It is useful to compare the variance of the asymptotic distribution given in (7.8) and the finite-sample conditional variance in the CEF model as given in (4.10):

$$V_{\hat{\beta}} = \text{var}(\hat{\beta} | X) = (X'X)^{-1} (X'DX) (X'X)^{-1}. \quad (7.9)$$

Notice that  $V_{\hat{\beta}}$  is the exact conditional variance of  $\hat{\beta}$  and  $V_\beta$  is the asymptotic variance of  $\sqrt{n}(\hat{\beta} - \beta)$ . Thus  $V_\beta$  should be (roughly)  $n$  times as large as  $V_{\hat{\beta}}$ , or  $V_\beta \approx nV_{\hat{\beta}}$ . Indeed, multiplying (7.9) by  $n$  and distributing, we find

$$nV_{\hat{\beta}} = \left( \frac{1}{n} X'X \right)^{-1} \left( \frac{1}{n} X'DX \right) \left( \frac{1}{n} X'X \right)^{-1}$$

which looks like an estimator of  $V_\beta$ . Indeed, as  $n \rightarrow \infty$

$$nV_{\hat{\beta}} \xrightarrow{p} V_\beta.$$

The expression  $V_{\hat{\beta}}$  is useful for practical inference (such as computation of standard errors and tests) since it is the variance of the estimator  $\hat{\beta}$ , while  $V_{\beta}$  is useful for asymptotic theory as it is well defined in the limit as  $n$  goes to infinity. We will make use of both symbols and it will be advisable to adhere to this convention.

There is a special case where  $\Omega$  and  $V_{\beta}$  simplify. Suppose that

$$\text{cov}(\mathbf{x}_i \mathbf{x}'_i, e_i^2) = \mathbf{0}. \quad (7.10)$$

Condition (7.10) holds in the homoskedastic linear regression model, but is somewhat broader. Under (7.10) the asymptotic variance formulae simplify as

$$\begin{aligned}\Omega &= \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i) \mathbb{E}(e_i^2) = \mathbf{Q}_{xx} \sigma^2 \\ V_{\beta} &= \mathbf{Q}_{xx}^{-1} \Omega \mathbf{Q}_{xx}^{-1} = \mathbf{Q}_{xx}^{-1} \sigma^2 \equiv V_{\beta}^0.\end{aligned} \quad (7.11)$$

In (7.11) we define  $V_{\beta}^0 = \mathbf{Q}_{xx}^{-1} \sigma^2$  whether (7.10) is true or false. When (7.10) is true then  $V_{\beta} = V_{\beta}^0$ , otherwise  $V_{\beta} \neq V_{\beta}^0$ . We call  $V_{\beta}^0$  the **homoskedastic asymptotic covariance matrix**.

Theorem 7.3 states that the sampling distribution of the least-squares estimator, after rescaling, is approximately normal when the sample size  $n$  is sufficiently large. This holds true for all joint distributions of  $(y_i, \mathbf{x}_i)$  which satisfy the conditions of Assumption 7.2, and is therefore broadly applicable. Consequently, asymptotic normality is routinely used to approximate the finite sample distribution of  $\sqrt{n}(\hat{\beta} - \beta)$ .

A difficulty is that for any fixed  $n$  the sampling distribution of  $\hat{\beta}$  can be arbitrarily far from the normal distribution. In Figure 6.1 we have already seen a simple example where the least-squares estimate is quite asymmetric and non-normal even for reasonably large sample sizes. The normal approximation improves as  $n$  increases, but how large should  $n$  be in order for the approximation to be useful? Unfortunately, there is no simple answer to this reasonable question. The trouble is that no matter how large is the sample size, the normal approximation is arbitrarily poor for some data distribution satisfying the assumptions. We illustrate this problem using a simulation. Let  $y_i = \beta_1 x_i + \beta_2 + e_i$  where  $x_i$  is  $N(0, 1)$ , and  $e_i$  is independent of  $x_i$  with the Double Pareto density  $f(e) = \frac{\alpha}{2} |e|^{-\alpha-1}$ ,  $|e| \geq 1$ . If  $\alpha > 2$  the error  $e_i$  has zero mean and variance  $\alpha/(\alpha-2)$ . As  $\alpha$  approaches 2, however, its variance diverges to infinity. In this context the normalized least-squares slope estimator  $\sqrt{n} \frac{\alpha-2}{\alpha} (\hat{\beta}_1 - \beta_1)$  has the  $N(0, 1)$  asymptotic distribution for any  $\alpha > 2$ . In Figure 7.2 we display the finite sample densities of the normalized estimator  $\sqrt{n} \frac{\alpha-2}{\alpha} (\hat{\beta}_1 - \beta_1)$ , setting  $n = 100$  and varying the parameter  $\alpha$ . For  $\alpha = 3.0$  the density is very close to the  $N(0, 1)$  density. As  $\alpha$  diminishes the density changes significantly, concentrating most of the probability mass around zero.

Another example is shown in Figure 7.3. Here the model is  $y_i = \beta + e_i$  where

$$e_i = \frac{u_i^r - \mathbb{E}(u_i^r)}{\left(\mathbb{E}(u_i^{2r}) - (\mathbb{E}(u_i^r))^2\right)^{1/2}} \quad (7.12)$$

and  $u_i \sim N(0, 1)$  and some integer  $r \geq 1$ . We show the sampling distribution of  $\sqrt{n}(\hat{\beta} - \beta)$  setting  $n = 100$ , for  $r = 1, 4, 6$  and  $8$ . As  $r$  increases, the sampling distribution becomes highly skewed and non-normal. The lesson from Figures 7.2 and 7.3 is that the  $N(0, 1)$  asymptotic approximation is never guaranteed to be accurate.

## 7.4 Joint Distribution

Theorem 7.3 gives the joint asymptotic distribution of the coefficient estimators. We can use the result to study the covariance between the coefficient estimators. For simplicity, suppose  $k = 2$  with no

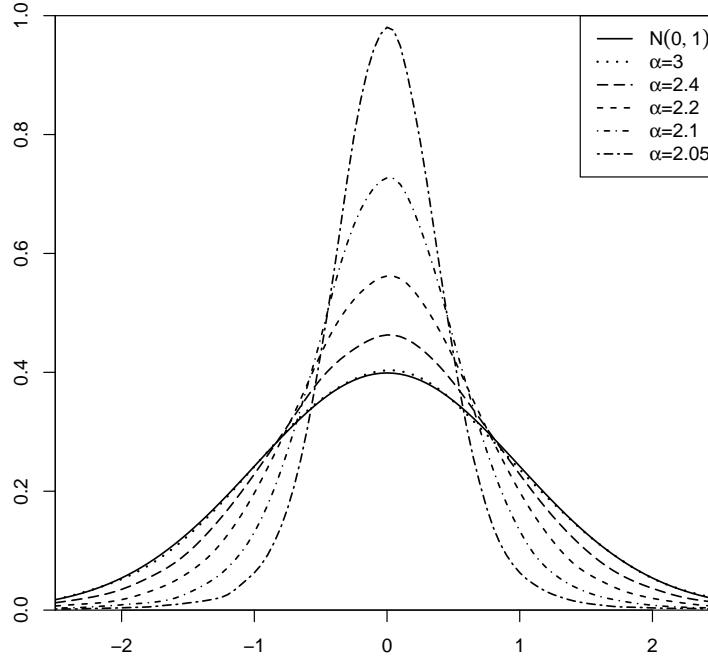


Figure 7.2: Density of Normalized OLS Estimator with Double Pareto Error

intercept, both regressors are mean zero and the error is homoskedastic. Let  $\sigma_1^2$  and  $\sigma_2^2$  be the variances of  $x_{1i}$  and  $x_{2i}$ , and  $\rho$  be their correlation. Then using the formula for inversion of a  $2 \times 2$  matrix,

$$\mathbf{V}_{\beta}^0 = \sigma^2 \mathbf{Q}_{xx}^{-1} = \frac{\sigma^2}{\sigma_1^2 \sigma_2^2 (1 - \rho^2)} \begin{bmatrix} \sigma_2^2 & -\rho \sigma_1 \sigma_2 \\ -\rho \sigma_1 \sigma_2 & \sigma_1^2 \end{bmatrix}.$$

Thus if  $x_{1i}$  and  $x_{2i}$  are positively correlated ( $\rho > 0$ ) then  $\hat{\beta}_1$  and  $\hat{\beta}_2$  are negatively correlated (and vice-versa).

For illustration, Figure 7.4 displays the probability contours of the joint asymptotic distribution of  $\hat{\beta}_1 - \beta_1$  and  $\hat{\beta}_2 - \beta_2$  when  $\beta_1 = \beta_2 = 0$ ,  $\sigma_1^2 = \sigma_2^2 = \sigma^2 = 1$ , and  $\rho = 0.5$ . The coefficient estimators are negatively correlated since the regressors are positively correlated. This means that if  $\hat{\beta}_1$  is unusually negative, it is likely that  $\hat{\beta}_2$  is unusually positive, or conversely. It is also unlikely that we will observe both  $\hat{\beta}_1$  and  $\hat{\beta}_2$  unusually large and of the same sign.

This finding that the correlation of the regressors is of opposite sign of the correlation of the coefficient estimates is sensitive to the assumption of homoskedasticity. If the errors are heteroskedastic then this relationship is not guaranteed.

This can be seen through a simple constructed example. Suppose that  $x_{1i}$  and  $x_{2i}$  only take the values  $\{-1, +1\}$ , symmetrically, with  $\mathbb{P}(x_{1i} = x_{2i} = 1) = \mathbb{P}(x_{1i} = x_{2i} = -1) = 3/8$ , and  $\mathbb{P}(x_{1i} = 1, x_{2i} = -1) = \mathbb{P}(x_{1i} = -1, x_{2i} = 1) = 1/8$ . You can check that the regressors are mean zero, unit variance and correlation 0.5, which is identical with the setting displayed in Figure 7.4.

Now suppose that the error is heteroskedastic. Specifically, suppose that  $\mathbb{E}(e_i^2 | x_{1i} = x_{2i}) = \frac{5}{4}$  and

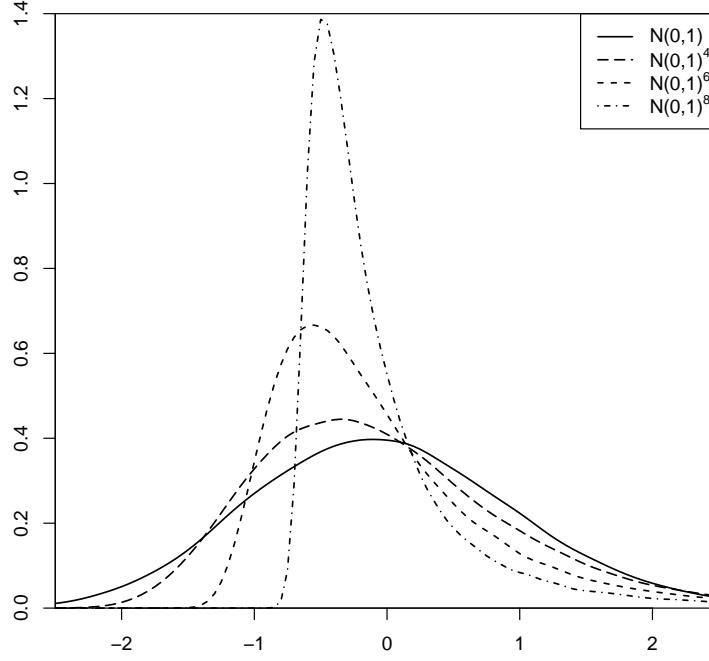


Figure 7.3: Density of Normalized OLS Estimator with Error Process (7.12)

$\mathbb{E}(e_i^2 | x_{1i} \neq x_{2i}) = \frac{1}{4}$ . You can check that  $\mathbb{E}(e_i^2) = 1$ ,  $\mathbb{E}(x_{1i}^2 e_i^2) = \mathbb{E}(x_{2i}^2 e_i^2) = 1$  and  $\mathbb{E}(x_{1i} x_{2i} e_i^2) = \frac{7}{8}$ . Therefore

$$\begin{aligned} V_{\beta} &= Q_{xx}^{-1} \Omega Q_{xx}^{-1} \\ &= \frac{9}{16} \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 1 & \frac{7}{8} \\ \frac{7}{8} & 1 \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \\ &= \frac{4}{3} \begin{bmatrix} 1 & \frac{1}{4} \\ \frac{1}{4} & 1 \end{bmatrix}. \end{aligned}$$

Thus the coefficient estimators  $\hat{\beta}_1$  and  $\hat{\beta}_2$  are positively correlated (their correlation is  $1/4$ .) The joint probability contours of their asymptotic distribution is displayed in Figure 7.5. We can see how the two estimators are positively associated.

What we found through this example is that in the presence of heteroskedasticity there is no simple relationship between the correlation of the regressors and the correlation of the parameter estimators.

We can extend the above analysis to study the covariance between coefficient sub-vectors. For example, partitioning  $x'_i = (x'_{1i}, x'_{2i})$  and  $\beta' = (\beta'_1, \beta'_2)$ , we can write the general model as

$$y_i = x'_{1i} \beta_1 + x'_{2i} \beta_2 + e_i$$

and the coefficient estimates as  $\hat{\beta}' = (\hat{\beta}'_1, \hat{\beta}'_2)$ . Make the partitions

$$Q_{xx} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}, \quad \Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}.$$

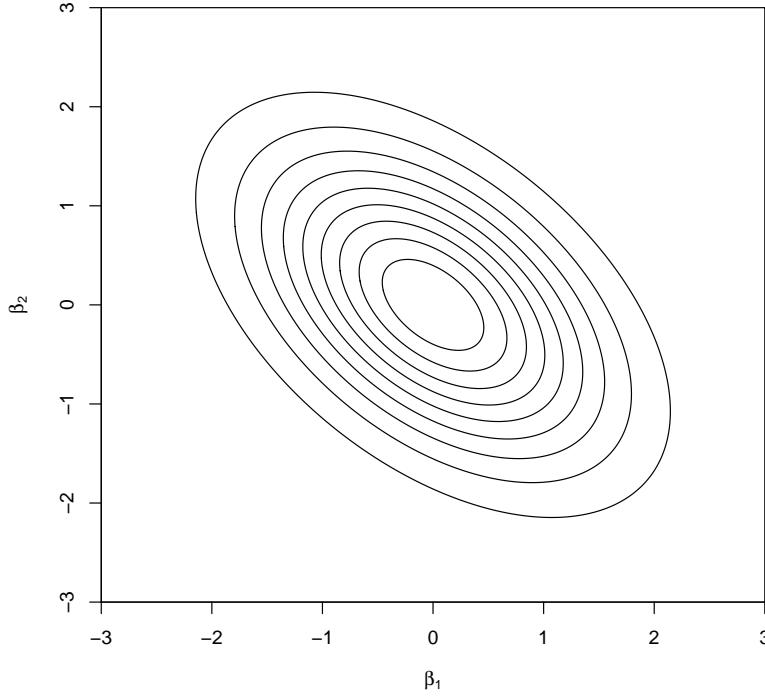


Figure 7.4: Contours of Joint Distribution of  $(\hat{\beta}_1, \hat{\beta}_2)$ , Homoskedastic Case

From (2.41)

$$\mathbf{Q}_{xx}^{-1} = \begin{bmatrix} \mathbf{Q}_{11 \cdot 2}^{-1} & -\mathbf{Q}_{11 \cdot 2}^{-1} \mathbf{Q}_{12} \mathbf{Q}_{22}^{-1} \\ -\mathbf{Q}_{22 \cdot 1}^{-1} \mathbf{Q}_{21} \mathbf{Q}_{11}^{-1} & \mathbf{Q}_{22 \cdot 1}^{-1} \end{bmatrix}$$

where  $\mathbf{Q}_{11 \cdot 2} = \mathbf{Q}_{11} - \mathbf{Q}_{12} \mathbf{Q}_{22}^{-1} \mathbf{Q}_{21}$  and  $\mathbf{Q}_{22 \cdot 1} = \mathbf{Q}_{22} - \mathbf{Q}_{21} \mathbf{Q}_{11}^{-1} \mathbf{Q}_{12}$ . Thus when the error is homoskedastic,

$$\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = -\sigma^2 \mathbf{Q}_{11 \cdot 2}^{-1} \mathbf{Q}_{12} \mathbf{Q}_{22}^{-1}$$

which is a matrix generalization of the two-regressor case.

In the general case, you can show that (Exercise 7.5)

$$\mathbf{V}_{\beta} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix} \quad (7.13)$$

where

$$\mathbf{V}_{11} = \mathbf{Q}_{11 \cdot 2}^{-1} (\Omega_{11} - \mathbf{Q}_{12} \mathbf{Q}_{22}^{-1} \Omega_{21} - \Omega_{12} \mathbf{Q}_{22}^{-1} \mathbf{Q}_{21} + \mathbf{Q}_{12} \mathbf{Q}_{22}^{-1} \Omega_{22} \mathbf{Q}_{22}^{-1} \Omega_{21}) \mathbf{Q}_{11 \cdot 2}^{-1} \quad (7.14)$$

$$\mathbf{V}_{21} = \mathbf{Q}_{22 \cdot 1}^{-1} (\Omega_{21} - \mathbf{Q}_{21} \mathbf{Q}_{11}^{-1} \Omega_{11} - \Omega_{22} \mathbf{Q}_{11}^{-1} \mathbf{Q}_{21} + \mathbf{Q}_{21} \mathbf{Q}_{11}^{-1} \Omega_{12} \mathbf{Q}_{11}^{-1} \Omega_{21}) \mathbf{Q}_{11 \cdot 2}^{-1} \quad (7.15)$$

$$\mathbf{V}_{22} = \mathbf{Q}_{22 \cdot 1}^{-1} (\Omega_{22} - \mathbf{Q}_{21} \mathbf{Q}_{11}^{-1} \Omega_{12} - \Omega_{21} \mathbf{Q}_{11}^{-1} \mathbf{Q}_{12} + \mathbf{Q}_{21} \mathbf{Q}_{11}^{-1} \Omega_{11} \mathbf{Q}_{11}^{-1} \Omega_{12}) \mathbf{Q}_{22 \cdot 1}^{-1} \quad (7.16)$$

Unfortunately, these expressions are not easily interpretable.

## 7.5 Consistency of Error Variance Estimators

Using the methods of Section 7.2 we can show that the estimators  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{e}_i^2$  and  $s^2 = \frac{1}{n-k} \sum_{i=1}^n \hat{e}_i^2$  are consistent for  $\sigma^2$ .

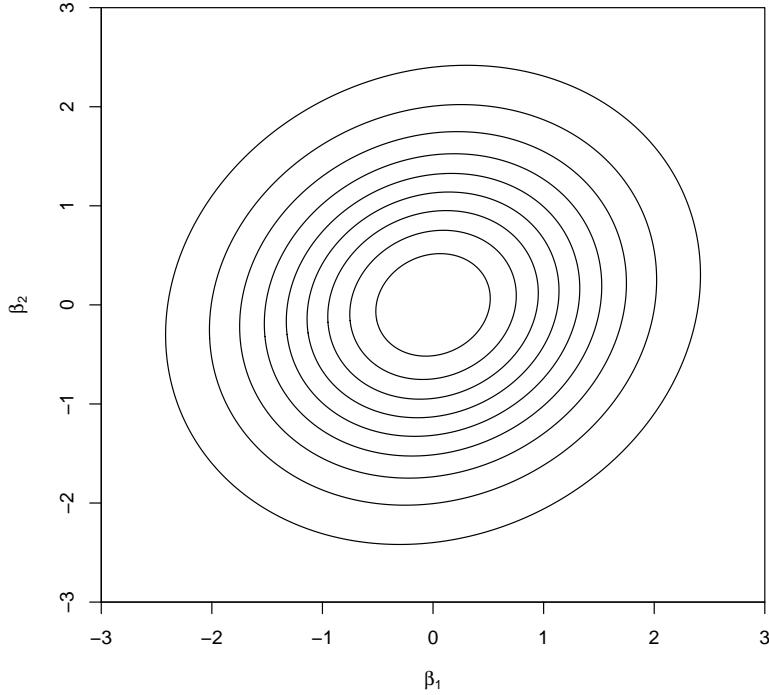


Figure 7.5: Contours of Joint Distribution of  $\hat{\beta}_1$  and  $\hat{\beta}_2$ , Heteroskedastic Case

The trick is to write the residual  $\hat{e}_i$  as equal to the error  $e_i$  plus a deviation term

$$\begin{aligned}\hat{e}_i &= y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}} \\ &= e_i + \mathbf{x}'_i \boldsymbol{\beta} - \mathbf{x}'_i \hat{\boldsymbol{\beta}} \\ &= e_i - \mathbf{x}'_i (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).\end{aligned}$$

Thus the squared residual equals the squared error plus a deviation

$$\hat{e}_i^2 = e_i^2 - 2e_i \mathbf{x}'_i (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{x}_i \mathbf{x}'_i (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \quad (7.17)$$

So when we take the average of the squared residuals we obtain the average of the squared errors, plus two terms which are (hopefully) asymptotically negligible.

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n e_i^2 - 2 \left( \frac{1}{n} \sum_{i=1}^n e_i \mathbf{x}'_i \right) (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\quad + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \right) (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).\end{aligned} \quad (7.18)$$

Indeed, the WLLN shows that

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n e_i^2 &\xrightarrow{p} \sigma^2 \\ \frac{1}{n} \sum_{i=1}^n e_i \mathbf{x}'_i &\xrightarrow{p} \mathbb{E}(e_i \mathbf{x}'_i) = 0 \\ \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i &\xrightarrow{p} \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i) = \mathbf{Q}_{xx}\end{aligned}$$

and Theorem 7.1 shows that  $\hat{\beta} \xrightarrow{p} \beta$ . Hence (7.18) converges in probability to  $\sigma^2$ , as desired.

Finally, since  $n/(n-k) \rightarrow 1$  as  $n \rightarrow \infty$ , it follows that

$$s^2 = \left( \frac{n}{n-k} \right) \hat{\sigma}^2 \xrightarrow{p} \sigma^2.$$

Thus both estimators are consistent.

**Theorem 7.4** Under Assumption 7.1,  $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$  and  $s^2 \xrightarrow{p} \sigma^2$  as  $n \rightarrow \infty$ .

## 7.6 Homoskedastic Covariance Matrix Estimation

Theorem 7.3 shows that  $\sqrt{n}(\hat{\beta} - \beta)$  is asymptotically normal with asymptotic covariance matrix  $V_\beta$ . For asymptotic inference (confidence intervals and tests) we need a consistent estimator of  $V_\beta$ . Under homoskedasticity,  $V_\beta$  simplifies to  $V_\beta^0 = Q_{xx}^{-1}\sigma^2$ , and in this section we consider the simplified problem of estimating  $V_\beta^0$ .

The standard moment estimator of  $Q_{xx}$  is  $\hat{Q}_{xx}$  defined in (7.1), and thus an estimator for  $Q_{xx}^{-1}$  is  $\hat{Q}_{xx}^{-1}$ . Also, the standard estimator of  $\sigma^2$  is the unbiased estimator  $s^2$  defined in (4.26). Thus a natural plug-in estimator for  $V_\beta^0 = Q_{xx}^{-1}\sigma^2$  is  $\hat{V}_\beta^0 = \hat{Q}_{xx}^{-1}s^2$ .

Consistency of  $\hat{V}_\beta^0$  for  $V_\beta^0$  follows from consistency of the moment estimators  $\hat{Q}_{xx}$  and  $s^2$ , and an application of the continuous mapping theorem. Specifically, Theorem 7.1 established that  $\hat{Q}_{xx} \xrightarrow{p} Q_{xx}$ , and Theorem 7.4 established  $s^2 \xrightarrow{p} \sigma^2$ . The function  $V_\beta^0 = Q_{xx}^{-1}\sigma^2$  is a continuous function of  $Q_{xx}$  and  $\sigma^2$  so long as  $Q_{xx} > 0$ , which holds true under Assumption 7.1.4. It follows by the CMT that

$$\hat{V}_\beta^0 = \hat{Q}_{xx}^{-1}s^2 \xrightarrow{p} Q_{xx}^{-1}\sigma^2 = V_\beta^0$$

so that  $\hat{V}_\beta^0$  is consistent for  $V_\beta^0$ , as desired.

**Theorem 7.5** Under Assumption 7.1,  $\hat{V}_\beta^0 \xrightarrow{p} V_\beta^0$  as  $n \rightarrow \infty$ .

It is instructive to notice that Theorem 7.5 does not require the assumption of homoskedasticity. That is,  $\hat{V}_\beta^0$  is consistent for  $V_\beta^0$  regardless if the regression is homoskedastic or heteroskedastic. However,  $V_\beta^0 = V_\beta = \text{avar}(\hat{\beta})$  only under homoskedasticity. Thus in the general case,  $\hat{V}_\beta^0$  is consistent for a well-defined but non-useful object.

## 7.7 Heteroskedastic Covariance Matrix Estimation

Theorems 7.3 established that the asymptotic covariance matrix of  $\sqrt{n}(\hat{\beta} - \beta)$  is  $V_\beta = Q_{xx}^{-1}\Omega Q_{xx}^{-1}$ . We now consider estimation of this covariance matrix without imposing homoskedasticity. The standard approach is to use a plug-in estimator which replaces the unknowns with sample moments.

As described in the previous section, a natural estimator for  $Q_{xx}^{-1}$  is  $\hat{Q}_{xx}^{-1}$ , where  $\hat{Q}_{xx}$  defined in (7.1).

The moment estimator for  $\Omega$  is

$$\hat{\Omega} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \hat{\epsilon}_i^2,$$

leading to the plug-in covariance matrix estimator

$$\widehat{V}_{\beta}^{\text{HC0}} = \widehat{\mathbf{Q}}_{xx}^{-1} \widehat{\Omega} \widehat{\mathbf{Q}}_{xx}^{-1}. \quad (7.19)$$

You can check that  $\widehat{V}_{\beta}^{\text{HC0}} = n \widehat{V}_{\widehat{\beta}}^{\text{HC0}}$  where  $\widehat{V}_{\widehat{\beta}}^{\text{HC0}}$  is the HC0 covariance matrix estimator introduced in (4.31).

As shown in Theorem 7.1,  $\widehat{\mathbf{Q}}_{xx}^{-1} \xrightarrow{p} \mathbf{Q}_{xx}^{-1}$ , so we just need to verify the consistency of  $\widehat{\Omega}$ . The key is to replace the squared residual  $\widehat{e}_i^2$  with the squared error  $e_i^2$ , and then show that the difference is asymptotically negligible.

Specifically, observe that

$$\begin{aligned} \widehat{\Omega} &= \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \widehat{e}_i^2 \\ &= \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i e_i^2 + \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i (\widehat{e}_i^2 - e_i^2). \end{aligned}$$

The first term is an average of the i.i.d. random variables  $\mathbf{x}_i \mathbf{x}'_i e_i^2$ , and therefore by the WLLN converges in probability to its expectation, namely,

$$\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i e_i^2 \xrightarrow{p} \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i e_i^2) = \Omega.$$

Technically, this requires that  $\Omega$  has finite elements, which was shown in (7.6).

So to establish that  $\widehat{\Omega}$  is consistent for  $\Omega$  it remains to show that

$$\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i (\widehat{e}_i^2 - e_i^2) \xrightarrow{p} 0. \quad (7.20)$$

There are multiple ways to do this. A reasonable straightforward yet slightly tedious derivation is to start by applying the triangle inequality (B.16) using a matrix norm:

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i (\widehat{e}_i^2 - e_i^2) \right\| &\leq \frac{1}{n} \sum_{i=1}^n \| \mathbf{x}_i \mathbf{x}'_i (\widehat{e}_i^2 - e_i^2) \| \\ &= \frac{1}{n} \sum_{i=1}^n \| \mathbf{x}_i \|^2 | \widehat{e}_i^2 - e_i^2 |. \end{aligned} \quad (7.21)$$

Then recalling the expression for the squared residual (7.17), apply the triangle inequality (B.1) and then the Schwarz inequality (B.12) twice

$$\begin{aligned} | \widehat{e}_i^2 - e_i^2 | &\leq 2 | e_i \mathbf{x}'_i (\widehat{\beta} - \beta) | + (\widehat{\beta} - \beta)' \mathbf{x}_i \mathbf{x}'_i (\widehat{\beta} - \beta) \\ &= 2 | e_i | | \mathbf{x}'_i (\widehat{\beta} - \beta) | + | (\widehat{\beta} - \beta)' \mathbf{x}_i |^2 \\ &\leq 2 | e_i | \| \mathbf{x}_i \| \| \widehat{\beta} - \beta \| + \| \mathbf{x}_i \|^2 \| \widehat{\beta} - \beta \|^2. \end{aligned} \quad (7.22)$$

Combining (7.21) and (7.22), we find

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i (\widehat{e}_i^2 - e_i^2) \right\| &\leq 2 \left( \frac{1}{n} \sum_{i=1}^n \| \mathbf{x}_i \|^3 | e_i | \right) \| \widehat{\beta} - \beta \| \\ &\quad + \left( \frac{1}{n} \sum_{i=1}^n \| \mathbf{x}_i \|^4 \right) \| \widehat{\beta} - \beta \|^2 \\ &= o_p(1). \end{aligned} \quad (7.23)$$

The expression is  $o_p(1)$  because  $\|\hat{\beta} - \beta\| \xrightarrow{p} 0$  and both averages in parenthesis are averages of random variables with finite mean under Assumption 7.2 (and are thus  $O_p(1)$ ). Indeed, by Hölder's inequality (B.30)

$$\begin{aligned}\mathbb{E}(\|\mathbf{x}_i\|^3 | e_i) &\leq \left(\mathbb{E}(\|\mathbf{x}_i\|^3)^{4/3}\right)^{3/4} (\mathbb{E}(e_i^4))^{1/4} \\ &= (\mathbb{E}(\|\mathbf{x}_i\|^4))^{3/4} (\mathbb{E}(e_i^4))^{1/4} < \infty.\end{aligned}$$

We have established (7.20), as desired.

**Theorem 7.6** Under Assumption 7.2, as  $n \rightarrow \infty$ ,  $\hat{\Omega} \xrightarrow{p} \Omega$  and  $\hat{V}_{\beta}^{\text{HC0}} \xrightarrow{p} V_{\beta}$ .

For an alternative proof of this result, see Section 7.22.

## 7.8 Summary of Covariance Matrix Notation

The notation we have introduced may be somewhat confusing so it is helpful to write it down in one place. The exact variance of  $\hat{\beta}$  (under the assumptions of the linear regression model) and the asymptotic variance of  $\sqrt{n}(\hat{\beta} - \beta)$  (under the more general assumptions of the linear projection model) are

$$\begin{aligned}V_{\hat{\beta}} &= \text{var}(\hat{\beta} | \mathbf{X}) = (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{D} \mathbf{X}) (\mathbf{X}' \mathbf{X})^{-1} \\ V_{\beta} &= \text{avar}(\sqrt{n}(\hat{\beta} - \beta)) = \mathbf{Q}_{xx}^{-1} \Omega \mathbf{Q}_{xx}^{-1}.\end{aligned}$$

The HC0 estimators of these two covariance matrices are

$$\begin{aligned}\hat{V}_{\hat{\beta}}^{\text{HC0}} &= (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \hat{e}_i^2 \right) (\mathbf{X}' \mathbf{X})^{-1} \\ \hat{V}_{\beta}^{\text{HC0}} &= \hat{\mathbf{Q}}_{xx}^{-1} \hat{\Omega} \hat{\mathbf{Q}}_{xx}^{-1}\end{aligned}$$

and satisfy the simple relationship

$$\hat{V}_{\beta}^{\text{HC0}} = n \hat{V}_{\hat{\beta}}^{\text{HC0}}.$$

Similarly, under the assumption of homoskedasticity the exact and asymptotic variances simplify to

$$\begin{aligned}V_{\hat{\beta}}^0 &= (\mathbf{X}' \mathbf{X})^{-1} \sigma^2 \\ V_{\beta}^0 &= \mathbf{Q}_{xx}^{-1} \sigma^2\end{aligned}$$

and their standard estimators are

$$\begin{aligned}\hat{V}_{\hat{\beta}}^0 &= (\mathbf{X}' \mathbf{X})^{-1} s^2 \\ \hat{V}_{\beta}^0 &= \hat{\mathbf{Q}}_{xx}^{-1} s^2\end{aligned}$$

which also satisfy the relationship

$$\hat{V}_{\beta}^0 = n \hat{V}_{\hat{\beta}}^0.$$

The exact formula and estimates are useful when constructing test statistics and standard errors. However, for theoretical purposes the asymptotic formula (variances and their estimates) are more useful, as these retain non-generate limits as the sample sizes diverge. That is why both sets of notation are useful.

## 7.9 Alternative Covariance Matrix Estimators\*

In Section 7.7 we introduced  $\widehat{V}_{\beta}^{\text{HC}0}$  as an estimator of  $V_{\beta}$ .  $\widehat{V}_{\beta}^{\text{HC}0}$  is a scaled version of  $\widehat{V}_{\widehat{\beta}}^{\text{HC}0}$  from Section 4.14, where we also introduced the alternative HC1, HC2 and HC3 heteroskedasticity-robust covariance matrix estimators. We now discuss the consistency properties of these estimators.

To do so we introduce their scaled versions, e.g.  $\widehat{V}_{\beta}^{\text{HC}1} = n\widehat{V}_{\widehat{\beta}}^{\text{HC}1}$ ,  $\widehat{V}_{\beta}^{\text{HC}2} = n\widehat{V}_{\widehat{\beta}}^{\text{HC}2}$ , and  $\widehat{V}_{\beta}^{\text{HC}3} = n\widehat{V}_{\widehat{\beta}}^{\text{HC}3}$ . These are (alternative) estimators of the asymptotic covariance matrix  $V_{\beta}$ .

First, consider  $\widehat{V}_{\beta}^{\text{HC}1}$ . Notice that  $\widehat{V}_{\beta}^{\text{HC}1} = n\widehat{V}_{\widehat{\beta}}^{\text{HC}1} = \frac{n}{n-k}\widehat{V}_{\beta}^{\text{HC}0}$  where  $\widehat{V}_{\beta}^{\text{HC}0}$  was defined in (7.19) and shown consistent for  $V_{\beta}$  in Theorem 7.6. If  $k$  is fixed as  $n \rightarrow \infty$ , then  $\frac{n}{n-k} \rightarrow 1$  and thus

$$\widehat{V}_{\beta}^{\text{HC}1} = (1 + o(1))\widehat{V}_{\beta}^{\text{HC}0} \xrightarrow{p} V_{\beta}.$$

Thus  $\widehat{V}_{\beta}^{\text{HC}1}$  is consistent for  $V_{\beta}$ .

The alternative estimators  $\widehat{V}_{\beta}^{\text{HC}2}$  and  $\widehat{V}_{\beta}^{\text{HC}3}$  take the form (7.19) but with  $\widehat{\Omega}$  replaced by

$$\tilde{\Omega} = \frac{1}{n} \sum_{i=1}^n (1 - h_{ii})^{-2} \mathbf{x}_i \mathbf{x}'_i \hat{e}_i^2$$

and

$$\bar{\Omega} = \frac{1}{n} \sum_{i=1}^n (1 - h_{ii})^{-1} \mathbf{x}_i \mathbf{x}'_i \hat{e}_i^2,$$

respectively. To show that these estimators also consistent for  $V_{\beta}$ , given  $\tilde{\Omega} \xrightarrow{p} \Omega$ , it is sufficient to show that the differences  $\tilde{\Omega} - \widehat{\Omega}$  and  $\bar{\Omega} - \widehat{\Omega}$  converge in probability to zero as  $n \rightarrow \infty$ .

The trick is to use the fact that the leverage values are asymptotically negligible:

$$h_n^* = \max_{1 \leq i \leq n} h_{ii} = o_p(1). \quad (7.24)$$

(See Theorem 7.18 in Section 7.23.) Then using the triangle inequality (B.16)

$$\begin{aligned} \|\bar{\Omega} - \widehat{\Omega}\| &\leq \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i \mathbf{x}'_i\| \hat{e}_i^2 |(1 - h_{ii})^{-1} - 1| \\ &\leq \left( \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i\|^2 \hat{e}_i^2 \right) |(1 - h_n^*)^{-1} - 1|. \end{aligned}$$

The sum in parenthesis can be shown to be  $O_p(1)$  under Assumption 7.2 by the same argument as in the proof of Theorem 7.6. (In fact, it can be shown to converge in probability to  $\mathbb{E}(\|\mathbf{x}_i\|^2 \hat{e}_i^2)$ .) The term in absolute values is  $o_p(1)$  by (7.24). Thus the product is  $o_p(1)$ , which means that  $\bar{\Omega} = \widehat{\Omega} + o_p(1) \xrightarrow{p} \Omega$ .

Similarly,

$$\begin{aligned} \|\tilde{\Omega} - \widehat{\Omega}\| &\leq \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i \mathbf{x}'_i\| \hat{e}_i^2 |(1 - h_{ii})^{-2} - 1| \\ &\leq \left( \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i\|^2 \hat{e}_i^2 \right) |(1 - h_n^*)^{-2} - 1| \\ &= o_p(1). \end{aligned}$$

**Theorem 7.7** Under Assumption 7.2, as  $n \rightarrow \infty$ ,  $\tilde{\Omega} \xrightarrow{p} \Omega$ ,  $\bar{\Omega} \xrightarrow{p} \Omega$ ,  $\widehat{V}_{\beta}^{\text{HC}1} \xrightarrow{p} V_{\beta}$ ,  $\widehat{V}_{\beta}^{\text{HC}2} \xrightarrow{p} V_{\beta}$ , and  $\widehat{V}_{\beta}^{\text{HC}3} \xrightarrow{p} V_{\beta}$ .

Theorem 7.7 shows that the alternative covariance matrix estimators are also consistent for the asymptotic covariance matrix.

To simplify notation, for the remainder of the chapter we will use the notation  $\widehat{V}_\beta$  and  $\widehat{V}_{\widehat{\beta}}$  to refer to any of the heteroskedasticity-consistent covariance matrix estimators HC0, HC1, HC2 and HC3, since they all have the same asymptotic limits.

## 7.10 Functions of Parameters

In most serious applications the researcher is actually interested in a specific transformation of the coefficient vector  $\beta = (\beta_1, \dots, \beta_k)$ . For example, he or she may be interested in a single coefficient  $\beta_j$  or a ratio  $\beta_j/\beta_l$ . More generally, interest may focus on a quantity such as consumer surplus which could be a complicated function of the coefficients. In any of these cases we can write the parameter of interest  $\theta$  as a function of the coefficients, e.g.  $\theta = r(\beta)$  for some function  $r : \mathbb{R}^k \rightarrow \mathbb{R}^q$ . The estimate of  $\theta$  is

$$\widehat{\theta} = r(\widehat{\beta}).$$

By the continuous mapping theorem (Theorem 6.19) and the fact  $\widehat{\beta} \xrightarrow{p} \beta$  we can deduce that  $\widehat{\theta}$  is consistent for  $\theta$  (if the function  $r(\cdot)$  is continuous).

**Theorem 7.8** Under Assumption 7.1, if  $r(\beta)$  is continuous at the true value of  $\beta$ , then as  $n \rightarrow \infty$ ,  $\widehat{\theta} \xrightarrow{p} \theta$ .

Furthermore, if the transformation is sufficiently smooth, by the Delta Method (Theorem 6.23) we can show that  $\widehat{\theta}$  is asymptotically normal.

**Assumption 7.3**  $r(\beta) : \mathbb{R}^k \rightarrow \mathbb{R}^q$  is continuously differentiable at the true value of  $\beta$  and  $R = \frac{\partial}{\partial \beta} r(\beta)'$  has rank  $q$ .

**Theorem 7.9 Asymptotic Distribution of Functions of Parameters**

Under Assumptions 7.2 and 7.3, as  $n \rightarrow \infty$ ,

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(\mathbf{0}, V_\theta) \quad (7.25)$$

where

$$V_\theta = R' V_\beta R.$$

In many cases, the function  $r(\beta)$  is linear:

$$r(\beta) = R' \beta$$

for some  $k \times q$  matrix  $R$ . In particular, if  $R$  is a “selector matrix”

$$R = \begin{pmatrix} I \\ \mathbf{0} \end{pmatrix}$$

then we can partition  $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2)'$  so that  $\mathbf{R}'\boldsymbol{\beta} = \boldsymbol{\beta}_1$  for  $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2)'$ . Then

$$\mathbf{V}_{\boldsymbol{\theta}} = (\begin{array}{cc} \mathbf{I} & \mathbf{0} \end{array}) \mathbf{V}_{\boldsymbol{\beta}} \begin{pmatrix} \mathbf{I} \\ \mathbf{0} \end{pmatrix} = \mathbf{V}_{11},$$

the upper-left sub-matrix of  $\mathbf{V}_{11}$  given in (7.14). In this case (7.25) states that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}_{11}).$$

That is, subsets of  $\hat{\boldsymbol{\beta}}$  are approximately normal with variances given by the conformable subcomponents of  $\mathbf{V}$ .

To illustrate the case of a nonlinear transformation, take the example  $\theta = \beta_j/\beta_l$  for  $j \neq l$ . Then

$$\mathbf{R} = \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{r}(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial}{\partial \beta_1} (\beta_j/\beta_l) \\ \vdots \\ \frac{\partial}{\partial \beta_j} (\beta_j/\beta_l) \\ \vdots \\ \frac{\partial}{\partial \beta_\ell} (\beta_j/\beta_l) \\ \vdots \\ \frac{\partial}{\partial \beta_k} (\beta_j/\beta_l) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 1/\beta_l \\ \vdots \\ -\beta_j/\beta_l^2 \\ \vdots \\ 0 \end{pmatrix} \quad (7.26)$$

so

$$\mathbf{V}_{\boldsymbol{\theta}} = \mathbf{V}_{jj}/\beta_l^2 + \mathbf{V}_{ll}\beta_j^2/\beta_l^4 - 2\mathbf{V}_{jl}\beta_j/\beta_l^3$$

where  $\mathbf{V}_{ab}$  denotes the  $ab^{th}$  element of  $\mathbf{V}_{\boldsymbol{\beta}}$ .

For inference we need an estimator of the asymptotic variance matrix  $\mathbf{V}_{\boldsymbol{\theta}} = \mathbf{R}'\mathbf{V}_{\boldsymbol{\beta}}\mathbf{R}$ , and for this it is typical to use a plug-in estimator. The natural estimator of  $\mathbf{R}$  is the derivative evaluated at the point estimator

$$\hat{\mathbf{R}} = \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{r}(\hat{\boldsymbol{\beta}})'. \quad (7.27)$$

The derivative in (7.27) may be calculated analytically or numerically. By analytically, we mean working out for the formula for the derivative and replacing the unknowns by point estimates. For example, if  $\theta = \beta_j/\beta_l$ , then  $\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{r}(\boldsymbol{\beta})$  is (7.26). However in some cases the function  $\mathbf{r}(\boldsymbol{\beta})$  may be extremely complicated and a formula for the analytic derivative may not be easily available. In this case calculation by numerical differentiation may be preferable. Let  $\delta_l = (0 \cdots 1 \cdots 0)'$  be the unit vector with the “1” in the  $l^{th}$  place. Then the  $jl^{th}$  element of a numerical derivative  $\hat{\mathbf{R}}$  is

$$\hat{\mathbf{R}}_{jl} = \frac{\mathbf{r}_j(\hat{\boldsymbol{\beta}} + \delta_l \varepsilon) - \mathbf{r}_j(\hat{\boldsymbol{\beta}})}{\varepsilon}$$

for some small  $\varepsilon$ .

The estimator of  $\mathbf{V}_{\boldsymbol{\theta}}$  is

$$\hat{\mathbf{V}}_{\boldsymbol{\theta}} = \hat{\mathbf{R}}' \hat{\mathbf{V}}_{\boldsymbol{\beta}} \hat{\mathbf{R}}. \quad (7.28)$$

Alternatively, the homoskedastic covariance matrix estimator could be used, leading to a homoskedastic covariance matrix estimator for  $\boldsymbol{\theta}$ .

$$\hat{\mathbf{V}}_{\boldsymbol{\theta}}^0 = \hat{\mathbf{R}}' \hat{\mathbf{V}}_{\boldsymbol{\beta}}^0 \hat{\mathbf{R}} = \hat{\mathbf{R}}' \hat{\mathbf{Q}}_{xx}^{-1} \hat{\mathbf{R}} s^2. \quad (7.29)$$

Given (7.27), (7.28) and (7.29) are simple to calculate using matrix operations.

As the primary justification for  $\hat{\mathbf{V}}_{\boldsymbol{\theta}}$  is the asymptotic approximation (7.25),  $\hat{\mathbf{V}}_{\boldsymbol{\theta}}$  is often called an **asymptotic covariance matrix estimator**.

The estimator  $\hat{\mathbf{V}}_{\boldsymbol{\theta}}$  is consistent for  $\mathbf{V}_{\boldsymbol{\theta}}$  under the conditions of Theorem 7.9 since  $\hat{\mathbf{V}}_{\boldsymbol{\beta}} \xrightarrow{p} \mathbf{V}_{\boldsymbol{\beta}}$  by Theorem 7.6, and

$$\hat{\mathbf{R}} = \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{r}(\hat{\boldsymbol{\beta}})' \xrightarrow{p} \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{r}(\boldsymbol{\beta}') = \mathbf{R}$$

since  $\hat{\beta} \xrightarrow{P} \beta$  and the function  $\frac{\partial}{\partial \beta} r(\beta)'$  is continuous in  $\beta$ .

**Theorem 7.10** Under Assumptions 7.2 and 7.3, as  $n \rightarrow \infty$ ,

$$\hat{V}_\theta \xrightarrow{P} V_\theta.$$

Theorem 7.10 shows that  $\hat{V}_\theta$  is consistent for  $V_\theta$  and thus may be used for asymptotic inference. In practice, we may set

$$\hat{V}_{\hat{\beta}} = \hat{R}' \hat{V}_{\hat{\beta}} \hat{R} = n^{-1} \hat{R}' \hat{V}_\beta \hat{R} \quad (7.30)$$

as an estimator of the variance of  $\hat{\theta}$ .

## 7.11 Asymptotic Standard Errors

As described in Section 4.15, a standard error is an estimator of the standard deviation of the distribution of an estimator. Thus if  $\hat{V}_{\hat{\beta}}$  is an estimator of the covariance matrix of  $\hat{\beta}$ , then standard errors are the square roots of the diagonal elements of this matrix. These take the form

$$s(\hat{\beta}_j) = \sqrt{\hat{V}_{\hat{\beta}_j}} = \sqrt{[\hat{V}_{\hat{\beta}}]_{jj}}.$$

Standard errors for  $\hat{\theta}$  are constructed similarly. Supposing that  $\theta = h(\beta)$  is real-valued then the standard error for  $\hat{\theta}$  is the square root of (7.30)

$$s(\hat{\theta}) = \sqrt{\hat{R}' \hat{V}_{\hat{\beta}} \hat{R}} = \sqrt{n^{-1} \hat{R}' \hat{V}_\beta \hat{R}}.$$

When the justification is based on asymptotic theory we call  $s(\hat{\beta}_j)$  or  $s(\hat{\theta})$  an **asymptotic standard error** for  $\hat{\beta}_j$  or  $\hat{\theta}$ . When reporting your results, it is good practice to report standard errors for each reported estimate, and this includes functions and transformations of your parameter estimates. This helps users of the work (including yourself) assess the estimation precision.

We illustrate using the log wage regression

$$\log(Wage) = \beta_1 education + \beta_2 experience + \beta_3 experience^2/100 + \beta_4 + e.$$

Consider the following three parameters of interest.

1. Percentage return to education:

$$\theta_1 = 100\beta_1$$

(100 times the partial derivative of the conditional expectation of log wages with respect to *education*.)

2. Percentage return to experience for individuals with 10 years of experience:

$$\theta_2 = 100\beta_2 + 20\beta_3$$

(100 times the partial derivative of the conditional expectation of log wages with respect to *experience*, evaluated at *experience* = 10.)

3. Experience level which maximizes expected log wages:

$$\theta_3 = -50\beta_2/\beta_3$$

(The level of *experience* at which the partial derivative of the conditional expectation of log wages with respect to *experience* equals 0.)

The  $4 \times 1$  vector  $\mathbf{R}$  for these three parameters is

$$\mathbf{R} = \begin{pmatrix} 100 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 100 \\ 20 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ -50/\beta_3 \\ 50\beta_2/\beta_3^2 \\ 0 \end{pmatrix},$$

respectively.

We use the subsample of married black women (all experience levels), which has 982 observations. The point estimates and standard errors are

$$\widehat{\log(Wage)} = \begin{array}{lllll} 0.118 & education & + & 0.016 & experience - 0.022 & experience^2/100 + 0.947 \\ (0.008) & & & (0.006) & & (0.012) \\ & & & & & (0.157) \end{array}. \quad (7.31)$$

The standard errors are the square roots of the Horn-Horn-Duncan covariance matrix estimate

$$\overline{\mathbf{V}}_{\hat{\beta}} = \begin{pmatrix} 0.632 & 0.131 & -0.143 & -11.1 \\ 0.131 & 0.390 & -0.731 & -6.25 \\ -0.143 & -0.731 & 1.48 & 9.43 \\ -11.1 & -6.25 & 9.43 & 246 \end{pmatrix} \times 10^{-4}. \quad (7.32)$$

We calculate that

$$\begin{aligned} \hat{\theta}_1 &= 100\hat{\beta}_1 \\ &= 100 \times 0.118 \\ &= 11.8 \end{aligned}$$

$$\begin{aligned} s(\hat{\theta}_1) &= \sqrt{100^2 \times 0.632 \times 10^{-4}} \\ &= 0.8 \end{aligned}$$

$$\begin{aligned} \hat{\theta}_2 &= 100\hat{\beta}_2 + 20\hat{\beta}_3 \\ &= 100 \times 0.016 - 20 \times 0.022 \\ &= 1.16 \end{aligned}$$

$$\begin{aligned} s(\hat{\theta}_2) &= \sqrt{(\begin{pmatrix} 100 & 20 \end{pmatrix} (\begin{pmatrix} 0.390 & -0.731 \\ -0.731 & 1.48 \end{pmatrix} (\begin{pmatrix} 100 \\ 20 \end{pmatrix}) \times 10^{-4}))} \\ &= 0.55 \end{aligned}$$

$$\begin{aligned} \hat{\theta}_3 &= -50\hat{\beta}_2/\hat{\beta}_3 \\ &= 50 \times 0.016/0.022 \\ &= 35.2 \end{aligned}$$

$$s(\hat{\theta}_3) = \sqrt{(-50/\hat{\beta}_3 \quad 50\hat{\beta}_2/\hat{\beta}_3^2) \begin{pmatrix} 0.390 & -0.731 \\ -0.731 & 1.48 \end{pmatrix} \begin{pmatrix} -50/\hat{\beta}_3 \\ 50\hat{\beta}_2/\hat{\beta}_3^2 \end{pmatrix}} \times 10^{-4}$$

$$= 7.0.$$

The calculations show that the estimate of the percentage return to education (for married black women) is about 12% per year, with a standard error of 0.8. The estimate of the percentage return to experience for those with 10 years of experience is 1.2% per year, with a standard error of 0.6. And the estimate of the experience level which maximizes expected log wages is 35 years, with a standard error of 7.

In Stata, the `nlcom` command can be used after estimation to perform the same calculations. To illustrate, after estimation of (7.31) using the same covariance matrix option, use the commands given below. In each case, Stata reports the coefficient estimate, asymptotic standard error and 95% confidence intervals.

#### Stata Commands

```
nlcom 100*_b[education]
nlcom 100*_b[experience]+20*_b[exp2]
nlcom -50*_b[experience]/_b[exp2]
```

## 7.12 t-statistic

Let  $\theta = r(\boldsymbol{\beta}) : \mathbb{R}^k \rightarrow \mathbb{R}$  be a parameter of interest,  $\hat{\theta}$  its estimator and  $s(\hat{\theta})$  its asymptotic standard error. Consider the statistic

$$T(\theta) = \frac{\hat{\theta} - \theta}{s(\hat{\theta})}. \quad (7.33)$$

Different writers have called (7.33) a **t-statistic**, a **t-ratio**, a **z-statistic** or a **studentized statistic**, sometimes using the different labels to distinguish between finite-sample and asymptotic inference. As the statistics themselves are always (7.33) we won't make this distinction, and will simply refer to  $T(\theta)$  as a t-statistic or a t-ratio. We also often suppress the parameter dependence, writing it as  $T$ . The t-statistic is a simple function of the estimate, its standard error, and the parameter.

By Theorems 7.9 and 7.10,  $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V_\theta)$  and  $\hat{V}_\theta \xrightarrow{p} V_\theta$ . Thus

$$\begin{aligned} T(\theta) &= \frac{\hat{\theta} - \theta}{s(\hat{\theta})} \\ &= \frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\hat{V}_\theta}} \\ &\xrightarrow{d} \frac{N(0, V_\theta)}{\sqrt{V_\theta}} \\ &= Z \sim N(0, 1). \end{aligned}$$

The last equality is by the property that affine functions of normal distributions are normal (Theorem 5.4).

This calculation also requires that  $V_\theta > 0$ , otherwise the continuous mapping theorem cannot be employed. This seems like an innocuous requirement, as it only excludes degenerate sampling distributions. Formally we add the following assumption.

**Assumption 7.4**  $V_{\theta} = \mathbf{R}' V_{\beta} \mathbf{R} > 0$ .

Assumption 7.4 states that  $V_{\theta}$  is positive definite. Since  $\mathbf{R}$  is full rank under Assumption 7.3, a sufficient condition is that  $V_{\beta} > 0$ , and since  $\mathbf{Q}_{xx} > 0$  a sufficient condition is  $\Omega > 0$ . Thus Assumption 7.4 could be replaced by the assumption  $\Omega > 0$ . Assumption 7.4 is weaker so this is what we use.

Thus the asymptotic distribution of the t-ratio  $T(\theta)$  is the standard normal. Since this distribution does not depend on the parameters, we say that  $T(\theta)$  is **asymptotically pivotal**. In finite samples  $T(\theta)$  is not necessarily pivotal (as in the normal regression model) but the property means that the dependence on unknowns diminishes as  $n$  increases.

As we will see in the next section, it is also useful to consider the distribution of the **absolute t-ratio**  $|T(\theta)|$ . Since  $T(\theta) \xrightarrow{d} Z$ , the continuous mapping theorem yields  $|T(\theta)| \xrightarrow{d} |Z|$ . Letting  $\Phi(u) = \mathbb{P}(Z \leq u)$  denote the standard normal distribution function, we can calculate that the distribution function of  $|Z|$  is

$$\begin{aligned}\mathbb{P}(|Z| \leq u) &= \mathbb{P}(-u \leq Z \leq u) \\ &= \mathbb{P}(Z \leq u) - \mathbb{P}(Z < -u) \\ &= \Phi(u) - \Phi(-u) \\ &= 2\Phi(u) - 1.\end{aligned}\tag{7.34}$$

**Theorem 7.11** Under Assumptions 7.2, 7.3 and 7.4,  $T(\theta) \xrightarrow{d} Z \sim N(0, 1)$  and  $|t_n(\theta)| \xrightarrow{d} |Z|$ .

The asymptotic normality of Theorem 7.11 is used to justify confidence intervals and tests for the parameters.

## 7.13 Confidence Intervals

The estimator  $\hat{\theta}$  is a **point estimator** for  $\theta$ , meaning that  $\hat{\theta}$  is a single value in  $\mathbb{R}^q$ . A broader concept is a **set estimator**  $\hat{C}$  which is a collection of values in  $\mathbb{R}^q$ . When the parameter  $\theta$  is real-valued then it is common to focus on sets of the form  $\hat{C} = [\hat{L}, \hat{U}]$  which is called an **interval estimator** for  $\theta$ .

An interval estimate  $\hat{C}$  is a function of the data and hence is random. The **coverage probability** of the interval  $\hat{C} = [\hat{L}, \hat{U}]$  is  $\mathbb{P}(\theta \in \hat{C})$ . The randomness comes from  $\hat{C}$  as the parameter  $\theta$  is treated as fixed. In Section 5.13 we introduced confidence intervals for the normal regression model, which used the finite sample distribution of the t-statistic to construct exact confidence intervals for the regression coefficients. When we are outside the normal regression model we cannot rely on the exact normal distribution theory, but instead use asymptotic approximations. A benefit is that we can construct confidence intervals for general parameters of interest  $\theta$ , not just regression coefficients.

An interval estimator  $\hat{C}$  is called a **confidence interval** when the goal is to set the coverage probability to equal a pre-specified target such as 90% or 95%.  $\hat{C}$  is called a  $1 - \alpha$  confidence interval if  $\inf_{\theta} \mathbb{P}_{\theta}(\theta \in \hat{C}) = 1 - \alpha$ .

When  $\hat{\theta}$  is asymptotically normal with standard error  $s(\hat{\theta})$ , the conventional confidence interval for  $\theta$  takes the form

$$\hat{C} = [\hat{\theta} - c \cdot s(\hat{\theta}), \quad \hat{\theta} + c \cdot s(\hat{\theta})]\tag{7.35}$$

where  $c$  equals the  $1 - \alpha$  quantile of the distribution of  $|Z|$ . Using (7.34) we calculate that  $c$  is equivalently the  $1 - \alpha/2$  quantile of the standard normal distribution. Thus,  $c$  solves

$$2\Phi(c) - 1 = 1 - \alpha.$$

This can be computed by, for example, `norminv(1-alpha/2)` in MATLAB. The confidence interval (7.35) is symmetric about the point estimator  $\hat{\theta}$ , and its length is proportional to the standard error  $s(\hat{\theta})$ .

Equivalently, (7.35) is the set of parameter values for  $\theta$  such that the t-statistic  $T(\theta)$  is smaller (in absolute value) than  $c$ , that is

$$\hat{C} = \{\theta : |T(\theta)| \leq c\} = \left\{ \theta : -c \leq \frac{\hat{\theta} - \theta}{s(\hat{\theta})} \leq c \right\}.$$

The coverage probability of this confidence interval is

$$\mathbb{P}(\theta \in \hat{C}) = \mathbb{P}(|T(\theta)| \leq c) \rightarrow \mathbb{P}(|Z| \leq c) = 1 - \alpha$$

where the limit is taken as  $n \rightarrow \infty$ , and holds since  $T(\theta)$  is asymptotically  $|Z|$  by Theorem 7.11. We call the limit the **asymptotic coverage probability**, and call  $\hat{C}$  an asymptotic  $1 - \alpha\%$  confidence interval for  $\theta$ . Since the t-ratio is asymptotically pivotal, the asymptotic coverage probability is independent of the parameter  $\theta$ .

It is useful to contrast the confidence interval (7.35) with (5.11) for the normal regression model. They are similar, but there are differences. The normal regression interval (5.11) only applies to regression coefficients  $\beta$ , not to functions  $\theta$  of the coefficients. The normal interval (5.11) also is constructed with the homoskedastic standard error, while (7.35) can be constructed with a heteroskedastic-robust standard error. Furthermore, the constants  $c$  in (5.11) are calculated using the student  $t$  distribution, while  $c$  in (7.35) are calculated using the normal distribution. The difference between the student  $t$  and normal values are typically small in practice (since sample sizes are large in typical economic applications). However, since the student  $t$  values are larger, it results in slightly larger confidence intervals, which is probably reasonable. (A practical rule of thumb is that if the sample sizes are sufficiently small that it makes a difference, then probably neither (5.11) nor (7.35) should be trusted.) Despite these differences, the coincidence of the intervals means that inference on regression coefficients is generally robust to using either the exact normal sampling assumption or the asymptotic large sample approximation, at least in large samples.

In Stata, by default the program reports 95% confidence intervals for each coefficient where the critical values  $c$  are calculated using the  $t_{n-k}$  distribution. This is done for all standard error methods even though it is only justified for homoskedastic standard errors and under normality.

The standard coverage probability for confidence intervals is 95%, leading to the choice  $c = 1.96$  for the constant in (7.35). Rounding 1.96 to 2, we obtain what might be the most commonly used confidence interval in applied econometric practice

$$\hat{C} = [\hat{\theta} - 2s(\hat{\theta}), \quad \hat{\theta} + 2s(\hat{\theta})].$$

This is a useful rule-of-thumb. This asymptotic 95% confidence interval  $\hat{C}$  is simple to compute and can be roughly calculated from tables of coefficient estimates and standard errors. (Technically, it is an asymptotic 95.4% interval, due to the substitution of 2.0 for 1.96, but this distinction is overly precise.)

**Theorem 7.12** Under Assumptions 7.2, 7.3 and 7.4, for  $\hat{C}$  defined in (7.35), with  $c = \Phi^{-1}(1 - \alpha/2)$ ,  $\mathbb{P}(\theta \in \hat{C}) \rightarrow 1 - \alpha$ . For  $c = 1.96$ ,  $\mathbb{P}(\theta \in \hat{C}) \rightarrow 0.95$ .

Confidence intervals are a simple yet effective tool to assess estimation uncertainty. When reading a set of empirical results, look at the estimated coefficient estimates and the standard errors. For a parameter of interest, compute the confidence interval  $C_n$  and consider the meaning of the spread of the suggested values. If the range of values in the confidence interval are too wide to learn about  $\theta$ , then do not jump to a conclusion about  $\theta$  based on the point estimate alone.

For illustration, consider the three examples presented in Section 7.11 based on the log wage regression for married black women.

Percentage return to education. A 95% asymptotic confidence interval is  $11.8 \pm 1.96 \times 0.8 = [10.2, 13.3]$ .

Percentage return to experience for individuals with 10 years experience. A 90% asymptotic confidence interval is  $1.1 \pm 1.645 \times 0.4 = [0.5, 1.8]$ .

Experience level which maximizes expected log wages. An 80% asymptotic confidence interval is  $35 \pm 1.28 \times 7 = [26, 44]$ .

## 7.14 Regression Intervals

In the linear regression model the conditional mean of  $y_i$  given  $\mathbf{x}_i = \mathbf{x}$  is

$$m(\mathbf{x}) = \mathbb{E}(y_i | \mathbf{x}_i = \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}.$$

In some cases, we want to estimate  $m(\mathbf{x})$  at a particular point  $\mathbf{x}$ . Notice that this is a linear function of  $\boldsymbol{\beta}$ . Letting  $r(\boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}$  and  $\theta = r(\boldsymbol{\beta})$ , we see that  $\hat{m}(\mathbf{x}) = \hat{\theta} = \mathbf{x}'\hat{\boldsymbol{\beta}}$  and  $\mathbf{R} = \mathbf{x}$ , so  $s(\hat{\theta}) = \sqrt{\mathbf{x}'\hat{V}_{\hat{\boldsymbol{\beta}}}\mathbf{x}}$ . Thus an asymptotic 95% confidence interval for  $m(\mathbf{x})$  is

$$\left[ \mathbf{x}'\hat{\boldsymbol{\beta}} \pm 1.96\sqrt{\mathbf{x}'\hat{V}_{\hat{\boldsymbol{\beta}}}\mathbf{x}} \right].$$

It is interesting to observe that if this is viewed as a function of  $\mathbf{x}$ , the width of the confidence interval is dependent on  $\mathbf{x}$ .

To illustrate, we return to the log wage regression (3.13) of Section 3.7. The estimated regression equation is

$$\widehat{\log(Wage)} = \mathbf{x}'\hat{\boldsymbol{\beta}} = 0.155x + 0.698$$

where  $x = \text{education}$ . The covariance matrix estimate from (4.38) is

$$\hat{V}_{\hat{\boldsymbol{\beta}}} = \begin{pmatrix} 0.001 & -0.015 \\ -0.015 & 0.243 \end{pmatrix}.$$

Thus the 95% confidence interval for the regression is

$$0.155x + 0.698 \pm 1.96\sqrt{0.001x^2 - 0.030x + 0.243}.$$

The estimated regression and 95% intervals are shown in Figure 7.6. Notice that the confidence bands take a hyperbolic shape. This means that the regression line is less precisely estimated for very large and very small values of *education*.

Plots of the estimated regression line and confidence intervals are especially useful when the regression includes nonlinear terms. To illustrate, consider the log wage regression (7.31) which includes experience and its square, with covariance matrix (7.32). We are interested in plotting the regression estimate and regression intervals as a function of *experience*. Since the regression also includes *education*, to plot the estimates in a simple graph we need to fix *education* at a specific value. We select *education*=12. This only affects the level of the estimated regression, since *education* enters without an interaction. Define the points of evaluation

$$\mathbf{z}(x) = \begin{pmatrix} 12 \\ x \\ x^2/100 \\ 1 \end{pmatrix}$$

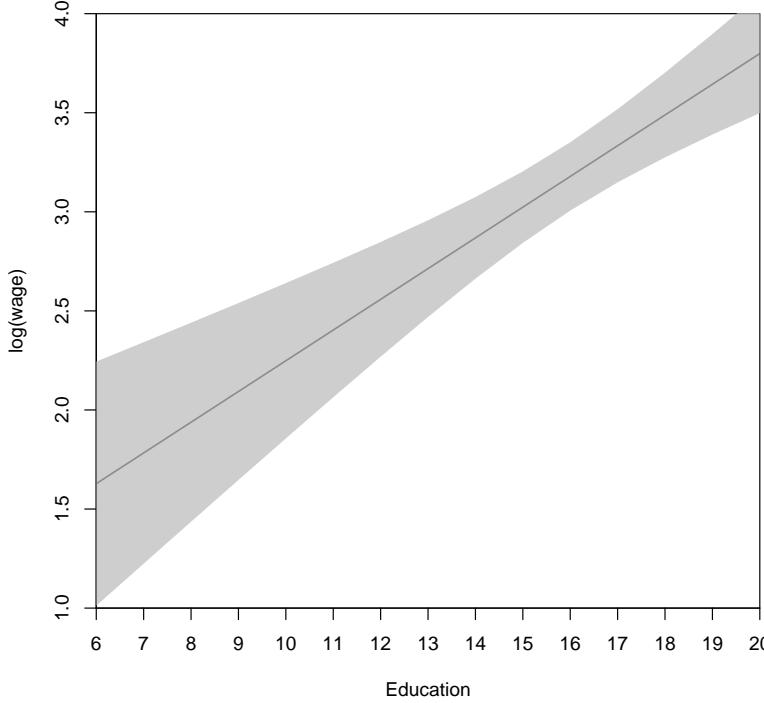


Figure 7.6: Wage on Education Regression Intervals

where  $x = \text{experience}$ .

Thus the 95% regression interval for  $\text{education}=12$ , as a function of  $x = \text{experience}$  is

$$\begin{aligned}
 & 0.118 \times 12 + 0.016 x - 0.022 x^2 / 100 + 0.947 \\
 & \pm 1.96 \sqrt{z(x)' \begin{pmatrix} 0.632 & 0.131 & -0.143 & -11.1 \\ 0.131 & 0.390 & -0.731 & -6.25 \\ -0.143 & -0.731 & 1.48 & 9.43 \\ -11.1 & -6.25 & 9.43 & 246 \end{pmatrix} z(x) \times 10^{-4}} \\
 & = 0.016 x - .00022 x^2 + 2.36 \\
 & \pm 0.0196 \sqrt{70.608 - 9.356 x + 0.54428 x^2 - 0.01462 x^3 + 0.000148 x^4}.
 \end{aligned}$$

The estimated regression and 95% intervals are shown in Figure 7.7. The regression interval widens greatly for small and large values of experience, indicating considerable uncertainty about the effect of experience on mean wages for this population. The confidence bands take a more complicated shape than in Figure 7.6 due to the nonlinear specification.

## 7.15 Forecast Intervals

Suppose we are given a value of the regressor vector  $\mathbf{x}_{n+1}$  for an individual outside the sample, and we want to forecast (guess)  $y_{n+1}$  for this individual. This is equivalent to forecasting  $y_{n+1}$  given  $\mathbf{x}_{n+1} = \mathbf{x}$ , which will generally be a function of  $\mathbf{x}$ . A reasonable forecasting rule is the conditional mean  $m(\mathbf{x})$  as it is the mean-square-minimizing forecast. A point forecast is the estimated conditional mean  $\hat{m}(\mathbf{x}) = \mathbf{x}'\hat{\beta}$ . We would also like a measure of uncertainty for the forecast.

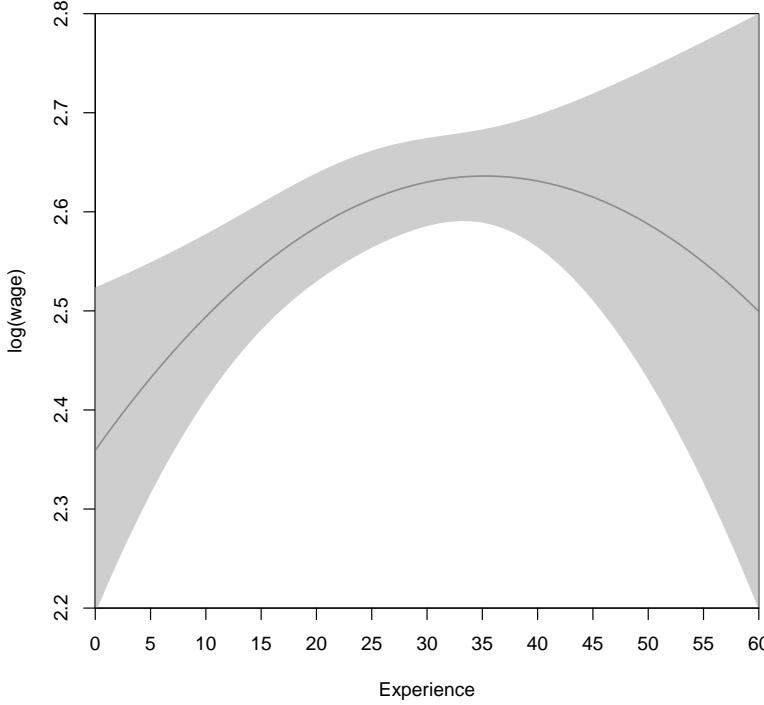


Figure 7.7: Wage on Experience Regression Intervals

The forecast error is  $\hat{e}_{n+1} = y_{n+1} - \hat{m}(\mathbf{x}) = e_{n+1} - \mathbf{x}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ . As the out-of-sample error  $e_{n+1}$  is independent of the in-sample estimate  $\hat{\boldsymbol{\beta}}$ , this has conditional variance

$$\begin{aligned}\mathbb{E}(\hat{e}_{n+1}^2 | \mathbf{x}_{n+1} = \mathbf{x}) &= \mathbb{E}(e_{n+1}^2 - 2\mathbf{x}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})e_{n+1} + \mathbf{x}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{x} | \mathbf{x}_{n+1} = \mathbf{x}) \\ &= \mathbb{E}(e_{n+1}^2 | \mathbf{x}_{n+1} = \mathbf{x}) + \mathbf{x}'\mathbb{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{x} \\ &= \sigma^2(\mathbf{x}) + \mathbf{x}'V_{\hat{\boldsymbol{\beta}}}\mathbf{x}.\end{aligned}\quad (7.36)$$

Under homoskedasticity  $\mathbb{E}(e_{n+1}^2 | \mathbf{x}_{n+1}) = \sigma^2$ . In this case a simple estimator of (7.36) is  $\hat{\sigma}^2 + \mathbf{x}'\hat{V}_{\hat{\boldsymbol{\beta}}}\mathbf{x}$ , so a standard error for the forecast is  $\hat{s}(\mathbf{x}) = \sqrt{\hat{\sigma}^2 + \mathbf{x}'\hat{V}_{\hat{\boldsymbol{\beta}}}\mathbf{x}}$ . Notice that this is different from the standard error for the conditional mean.

The conventional 95% forecast interval for  $y_{n+1}$  uses a normal approximation and sets

$$[\mathbf{x}'\hat{\boldsymbol{\beta}} \pm 2\hat{s}(\mathbf{x})].$$

It is difficult, however, to fully justify this choice. It would be correct if we have a normal approximation to the ratio

$$\frac{e_{n+1} - \mathbf{x}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\hat{s}(\mathbf{x})}.$$

The difficulty is that the equation error  $e_{n+1}$  is generally non-normal, and asymptotic theory cannot be applied to a single observation. The only special exception is the case where  $e_{n+1}$  has the exact distribution  $N(0, \sigma^2)$ , which is generally invalid.

To get an accurate forecast interval, we need to estimate the conditional distribution of  $e_{n+1}$  given  $\mathbf{x}_{n+1} = \mathbf{x}$ , which is a much more difficult task. Perhaps due to this difficulty, many applied forecasters use the simple approximate interval  $[\mathbf{x}'\hat{\boldsymbol{\beta}} \pm 2\hat{s}(\mathbf{x})]$  despite the lack of a convincing justification.

## 7.16 Wald Statistic

Let  $\boldsymbol{\theta} = \mathbf{r}(\boldsymbol{\beta}) : \mathbb{R}^k \rightarrow \mathbb{R}^q$  be any parameter vector of interest,  $\hat{\boldsymbol{\theta}}$  its estimator and  $\hat{V}_{\hat{\boldsymbol{\theta}}}$  its covariance matrix estimator. Consider the quadratic form

$$W(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})' \hat{V}_{\hat{\boldsymbol{\theta}}}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = n (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})' \hat{V}_{\boldsymbol{\theta}}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}). \quad (7.37)$$

where  $\hat{V}_{\boldsymbol{\theta}} = n \hat{V}_{\hat{\boldsymbol{\theta}}}$ . When  $q = 1$ , then  $W(\boldsymbol{\theta}) = T(\boldsymbol{\theta})^2$  is the square of the t-ratio. When  $q > 1$ ,  $W(\boldsymbol{\theta})$  is typically called a **Wald statistic** as it was proposed by Wald (1943). We are interested in its sampling distribution.

The asymptotic distribution of  $W(\boldsymbol{\theta})$  is simple to derive given Theorem 7.9 and Theorem 7.10, which show that

$$\sqrt{n} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} Z \sim N(\mathbf{0}, V_{\boldsymbol{\theta}})$$

and

$$\hat{V}_{\boldsymbol{\theta}} \xrightarrow{p} V_{\boldsymbol{\theta}}.$$

It follows that

$$W(\boldsymbol{\theta}) = \sqrt{n} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})' \hat{V}_{\boldsymbol{\theta}}^{-1} \sqrt{n} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} Z' V_{\boldsymbol{\theta}}^{-1} Z$$

a quadratic in the normal random vector  $Z$ . As shown in Theorem 5.12, the distribution of this quadratic form is  $\chi_q^2$ , a chi-square random variable with  $q$  degrees of freedom.

**Theorem 7.13** Under Assumptions 7.2, 7.3 and 7.4, as  $n \rightarrow \infty$ ,

$$W(\boldsymbol{\theta}) \xrightarrow{d} \chi_q^2.$$

Theorem 7.13 is used to justify multivariate confidence regions and multivariate hypothesis tests.

## 7.17 Homoskedastic Wald Statistic

Under the conditional homoskedasticity assumption  $\mathbb{E}(e_i^2 | \mathbf{x}_i) = \sigma^2$  we can construct the Wald statistic using the homoskedastic covariance matrix estimator  $\hat{V}_{\boldsymbol{\theta}}^0$  defined in (7.29). This yields a homoskedastic Wald statistic

$$W^0(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})' (\hat{V}_{\boldsymbol{\theta}}^0)^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = n (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})' (\hat{V}_{\boldsymbol{\theta}}^0)^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}). \quad (7.38)$$

Under the additional assumption of conditional homoskedasticity, it has the same asymptotic distribution as  $W(\boldsymbol{\theta})$ .

**Theorem 7.14** Under Assumptions 7.2, 7.3, and  $\mathbb{E}(e_i^2 | \mathbf{x}_i) = \sigma^2 > 0$ , as  $n \rightarrow \infty$ ,

$$W^0(\boldsymbol{\theta}) \xrightarrow{d} \chi_q^2.$$

## 7.18 Confidence Regions

A confidence region  $\widehat{C}$  is a set estimator for  $\boldsymbol{\theta} \in \mathbb{R}^q$  when  $q > 1$ . A confidence region  $\widehat{C}$  is a set in  $\mathbb{R}^q$  intended to cover the true parameter value with a pre-selected probability  $1 - \alpha$ . Thus an ideal confidence region has the coverage probability  $\mathbb{P}(\boldsymbol{\theta} \in \widehat{C}) = 1 - \alpha$ . In practice it is typically not possible to construct a region with exact coverage, but we can calculate its asymptotic coverage.

When the parameter estimator satisfies the conditions of Theorem 7.13, a good choice for a confidence region is the ellipse

$$\widehat{C} = \{\boldsymbol{\theta} : W(\boldsymbol{\theta}) \leq c_{1-\alpha}\}$$

with  $c_{1-\alpha}$  the  $1 - \alpha$  quantile of the  $\chi_q^2$  distribution. (Thus  $F_q(c_{1-\alpha}) = 1 - \alpha$ .) It can be computed by, for example, `chi2inv(1-alpha, q)` in MATLAB.

Theorem 7.13 implies

$$\mathbb{P}(\boldsymbol{\theta} \in \widehat{C}) \rightarrow \mathbb{P}\left(\chi_q^2 \leq c_{1-\alpha}\right) = 1 - \alpha$$

which shows that  $\widehat{C}$  has asymptotic coverage  $1 - \alpha$ .

To illustrate the construction of a confidence region, consider the estimated regression (7.31) of the model

$$\widehat{\log(Wage)} = \beta_1 \text{education} + \beta_2 \text{experience} + \beta_3 \text{experience}^2/100 + \beta_4.$$

Suppose that the two parameters of interest are the percentage return to education  $\theta_1 = 100\beta_1$  and the percentage return to experience for individuals with 10 years experience  $\theta_2 = 100\beta_2 + 20\beta_3$ . These two parameters are a linear transformation of the regression parameters with point estimates

$$\widehat{\boldsymbol{\theta}} = \begin{pmatrix} 100 & 0 & 0 & 0 \\ 0 & 100 & 20 & 0 \end{pmatrix} \widehat{\boldsymbol{\beta}} = \begin{pmatrix} 11.8 \\ 1.2 \end{pmatrix},$$

and have the covariance matrix estimate

$$\begin{aligned} \widehat{\mathbf{V}}_{\widehat{\boldsymbol{\theta}}} &= \begin{pmatrix} 0 & 100 & 0 & 0 \\ 0 & 0 & 100 & 20 \end{pmatrix} \widehat{\mathbf{V}}_{\widehat{\boldsymbol{\beta}}} \begin{pmatrix} 0 & 0 \\ 100 & 0 \\ 0 & 100 \\ 0 & 20 \end{pmatrix} \\ &= \begin{pmatrix} 0.632 & 0.103 \\ 0.103 & 0.157 \end{pmatrix} \end{aligned}$$

with inverse

$$\widehat{\mathbf{V}}_{\widehat{\boldsymbol{\theta}}}^{-1} = \begin{pmatrix} 1.77 & -1.16 \\ -1.16 & 7.13 \end{pmatrix}.$$

Thus the Wald statistic is

$$\begin{aligned} W(\boldsymbol{\theta}) &= (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})' \widehat{\mathbf{V}}_{\widehat{\boldsymbol{\theta}}}^{-1} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\ &= \begin{pmatrix} 11.8 - \theta_1 \\ 1.2 - \theta_2 \end{pmatrix}' \begin{pmatrix} 1.77 & -1.16 \\ -1.16 & 7.13 \end{pmatrix} \begin{pmatrix} 11.8 - \theta_1 \\ 1.2 - \theta_2 \end{pmatrix} \\ &= 1.77(11.8 - \theta_1)^2 - 2.32(11.8 - \theta_1)(1.2 - \theta_2) + 7.13(1.2 - \theta_2)^2. \end{aligned}$$

The 90% quantile of the  $\chi_2^2$  distribution is 4.605 (we use the  $\chi_2^2$  distribution as the dimension of  $\boldsymbol{\theta}$  is two), so an asymptotic 90% confidence region for the two parameters is the interior of the ellipse  $W(\boldsymbol{\theta}) = 4.605$  which is displayed in Figure 7.8. Since the estimated correlation of the two coefficient estimates is modest (about 0.3) the region is modestly elliptical.

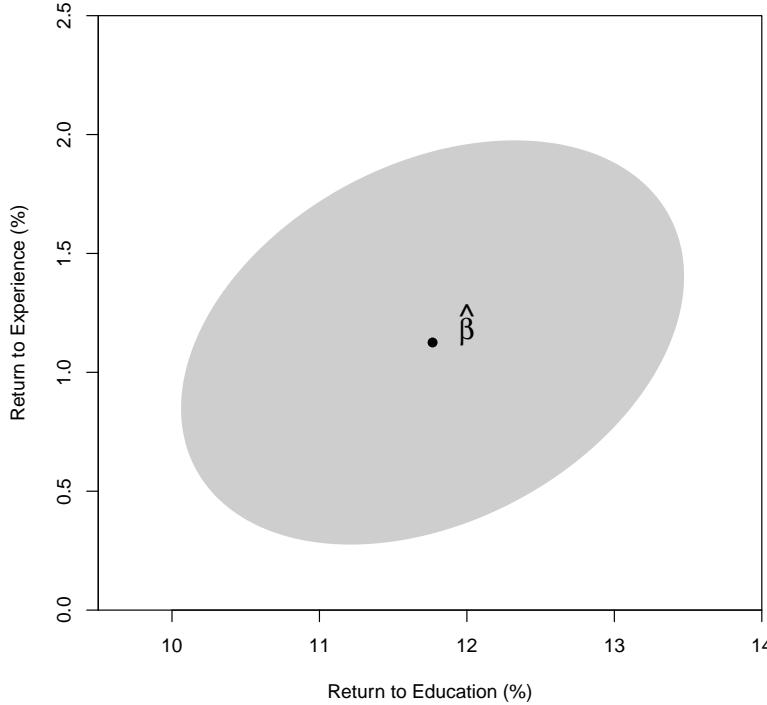


Figure 7.8: Confidence Region for Return to Experience and Return to Education

## 7.19 Edgeworth Expansion\*

Theorem 7.11 showed that the t-ratio  $T(\theta)$  is asymptotically normal. In practice this means that we use the normal distribution to approximate the finite sample distribution of  $T$ . How good is this approximation? Some insight into the accuracy of the normal approximation can be obtained by an Edgeworth expansion, which is a higher-order approximation to the distribution of  $T$ . The following result is an application of Theorem 6.34.

**Theorem 7.15** Under Assumptions 7.2, 7.3 and  $\Omega > 0$ ,  $\mathbb{E} \|e\|^{16} < \infty$ ,  $\mathbb{E} \|x\|^{16} < \infty$ ,  $g(u)$  has five continuous derivatives in a neighborhood of  $\beta$ , and  $\mathbb{E}(\exp(t(\|e\|^4 + \|x\|^4))) \leq B < 1$ , as  $n \rightarrow \infty$

$$\mathbb{P}(T(\theta) \leq x) = \Phi(x) + n^{-1/2} p_1(x) \phi(x) + n^{-1} p_2(x) \phi(x) + o(n^{-1})$$

uniformly in  $x$ , where  $p_1(x)$  is an even polynomial of order 2, and  $p_2(x)$  is an odd polynomial of degree 5, with coefficients depending on the moments of  $e$  and  $x$  up to order 16.

Theorem 7.15 shows that the finite sample distribution of the t-ratio can be approximated up to  $o(n^{-1})$  by the sum of three terms, the first being the standard normal distribution, the second a  $O(n^{-1/2})$  adjustment and the third a  $O(n^{-1})$  adjustment.

Consider a one-sided confidence interval  $C = [\hat{\theta} - z_{1-\alpha} s(\hat{\theta}), \infty)$  where  $z_{1-\alpha}$  is the  $1 - \alpha^{th}$  quantile of

$Z \sim N(0, 1)$ , thus  $\Phi(z_{1-\alpha}) - 1 - \alpha$ . Then

$$\begin{aligned}\mathbb{P}(\theta \in C) &= \mathbb{P}(T(\theta) \leq z_{1-\alpha}) \\ &= \Phi(z_{1-\alpha}) + n^{-1/2} p_1(z_{1-\alpha})\phi(z_{1-\alpha}) + O(n^{-1}) \\ &= 1 - \alpha + O(n^{-1/2}).\end{aligned}$$

This means that the actual coverage is within  $O(n^{-1/2})$  of the desired  $1 - \alpha$  level.

Now consider a two-sided interval  $C = [\hat{\theta} - z_{1-\alpha/2}s(\hat{\theta}), \hat{\theta} + z_{1-\alpha/2}s(\hat{\theta})]$ . It has coverage

$$\begin{aligned}\mathbb{P}(\theta \in C) &= \mathbb{P}(|T(\theta)| \leq z_{1-\alpha/2}) \\ &= 2\Phi(z_{1-\alpha/2}) - 1 + n^{-1/2} p_2(z_{1-\alpha/2})\phi(z_{1-\alpha/2}) + o(n^{-1}) \\ &= 1 - \alpha + O(n^{-1}).\end{aligned}$$

This means that the actual coverage is within  $O(n^{-1})$  of the desired  $1 - \alpha$  level. The accuracy is better than the one-sided interval because the  $O(n^{-1/2})$  term in the Edgeworth expansion has offsetting effects in the two tails of the distribution.

## 7.20 Semiparametric Efficiency in the Projection Model\*

In Section 4.8 we presented the Gauss-Markov theorem, which stated that in the homoskedastic CEF model, in the class of linear unbiased estimators the one with the smallest variance is least-squares. As we noted in that section, the restriction to linear unbiased estimators is unsatisfactory as it leaves open the possibility that an alternative (non-linear) estimator could have a smaller asymptotic variance. In addition, the restriction to the homoskedastic CEF model is also unsatisfactory as the projection model is more relevant for empirical application. The question remains: what is the most efficient estimator of the projection coefficient  $\beta$  (or functions  $\theta = h(\beta)$ ) in the projection model?

It turns out that it is straightforward to show that the projection model falls in the estimator class considered in Proposition 6.2. It follows that the least-squares estimator is semiparametrically efficient in the sense that it has the smallest asymptotic variance in the class of semiparametric estimators of  $\beta$ . This is a more powerful and interesting result than the Gauss-Markov theorem.

To see this, it is worth rephrasing Proposition 6.2 with amended notation. Suppose that a parameter of interest is  $\theta = g(\mu)$  where  $\mu = \mathbb{E}(z_i)$ , for which the moment estimators are  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n z_i$  and  $\hat{\theta} = g(\hat{\mu})$ . Let

$$\mathcal{L}_2(g) = \{F : \mathbb{E}\|z\|^2 < \infty, g(u) \text{ is continuously differentiable at } u = \mathbb{E}(z)\}$$

be the set of distributions for which  $\hat{\theta}$  satisfies the central limit theorem.

**Proposition 7.1** In the class of distributions  $F \in \mathcal{L}_2(g)$ ,  $\hat{\theta}$  is semiparametrically efficient for  $\theta$  in the sense that its asymptotic variance equals the semiparametric efficiency bound.

Proposition 7.1 says that under the minimal conditions in which  $\hat{\theta}$  is asymptotically normal, then no semiparametric estimator can have a smaller asymptotic variance than  $\hat{\theta}$ .

To show that an estimator is semiparametrically efficient it is sufficient to show that it falls in the class covered by this Proposition. To show that the projection model falls in this class, we write  $\beta = Q_{xx}^{-1}Q_{xy} = g(\mu)$  where  $\mu = \mathbb{E}(z_i)$  and  $z_i = (x_i x_i', x_i y_i)$ . The class  $\mathcal{L}_2(g)$  equals the class of distributions

$$\mathcal{L}_4(\beta) = \{F : \mathbb{E}(y^4) < \infty, \mathbb{E}\|x\|^4 < \infty, \mathbb{E}(x_i x_i') > 0\}.$$

**Proposition 7.2** In the class of distributions  $F \in \mathcal{L}_4(\boldsymbol{\beta})$ , the least-squares estimator  $\hat{\boldsymbol{\beta}}$  is semiparametrically efficient for  $\boldsymbol{\beta}$ .

The least-squares estimator is an asymptotically efficient estimator of the projection coefficient because the latter is a smooth function of sample moments and the model implies no further restrictions. However, if the class of permissible distributions is restricted to a strict subset of  $\mathcal{L}_4(\boldsymbol{\beta})$  then least-squares can be inefficient. For example, the linear CEF model with heteroskedastic errors is a strict subset of  $\mathcal{L}_4(\boldsymbol{\beta})$ , and the GLS estimator has a smaller asymptotic variance than OLS. In this case, the knowledge that true conditional mean is linear allows for more efficient estimation of the unknown parameter.

From Proposition 7.1 we can also deduce that plug-in estimators  $\hat{\boldsymbol{\theta}} = \mathbf{h}(\hat{\boldsymbol{\beta}})$  are semiparametrically efficient estimators of  $\boldsymbol{\theta} = \mathbf{h}(\boldsymbol{\beta})$  when  $\mathbf{h}$  is continuously differentiable. We can also deduce that other parameters estimators are semiparametrically efficient, such as  $\hat{\sigma}^2$  for  $\sigma^2$ . To see this, note that we can write

$$\begin{aligned}\sigma^2 &= \mathbb{E}((y_i - \mathbf{x}'_i \boldsymbol{\beta})^2) \\ &= \mathbb{E}(y_i^2) - 2\mathbb{E}(y_i \mathbf{x}'_i) \boldsymbol{\beta} + \boldsymbol{\beta}' \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i) \boldsymbol{\beta} \\ &= Q_{yy} - \mathbf{Q}_{yx} \mathbf{Q}_{xx}^{-1} \mathbf{Q}_{xy}\end{aligned}$$

which is a smooth function of the moments  $Q_{yy}$ ,  $\mathbf{Q}_{yx}$  and  $\mathbf{Q}_{xx}$ . Similarly the estimator  $\hat{\sigma}^2$  equals

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n \hat{e}_i^2 \\ &= \hat{Q}_{yy} - \hat{\mathbf{Q}}_{yx} \hat{\mathbf{Q}}_{xx}^{-1} \hat{\mathbf{Q}}_{xy}.\end{aligned}$$

Since the variables  $y_i^2$ ,  $y_i \mathbf{x}'_i$  and  $\mathbf{x}_i \mathbf{x}'_i$  all have finite variances when  $F \in \mathcal{L}_4(\boldsymbol{\beta})$ , the conditions of Proposition 7.1 are satisfied. We conclude:

**Proposition 7.3** In the class of distributions  $F \in \mathcal{L}_4(\boldsymbol{\beta})$ ,  $\hat{\sigma}^2$  is semiparametrically efficient for  $\sigma^2$ .

## 7.21 Semiparametric Efficiency in the Homoskedastic Regression Model\*

In Section 7.20 we showed that the OLS estimator is semiparametrically efficient in the projection model. What if we restrict attention to the classical homoskedastic regression model? Is OLS still efficient in this class? In this section we derive the asymptotic semiparametric efficiency bound for this model, and show that it is the same as that obtained by the OLS estimator. Therefore it turns out that least-squares is efficient in this class as well.

Recall that in the homoskedastic regression model the asymptotic variance of the OLS estimator  $\hat{\boldsymbol{\beta}}$  for  $\boldsymbol{\beta}$  is  $V_{\boldsymbol{\beta}}^0 = \mathbf{Q}_{xx}^{-1} \sigma^2$ . Therefore, as described in Section 6.27, it is sufficient to find a parametric submodel whose Cramer-Rao bound for estimation of  $\boldsymbol{\beta}$  is  $V_{\boldsymbol{\beta}}^0$ . This would establish that  $V_{\boldsymbol{\beta}}^0$  is the semiparametric variance bound and the OLS estimator  $\hat{\boldsymbol{\beta}}$  is semiparametrically efficient for  $\boldsymbol{\beta}$ .

Let the joint density of  $y$  and  $\mathbf{x}$  be written as  $f(y, \mathbf{x}) = f_1(y | \mathbf{x}) f_2(\mathbf{x})$ , the product of the conditional density of  $y$  given  $\mathbf{x}$  and the marginal density of  $\mathbf{x}$ . Now consider the parametric submodel

$$f(y, \mathbf{x} | \boldsymbol{\theta}) = f_1(y | \mathbf{x}) (1 + (y - \mathbf{x}' \boldsymbol{\beta})(\mathbf{x}' \boldsymbol{\theta}) / \sigma^2) f_2(\mathbf{x}). \quad (7.39)$$

You can check that in this submodel the marginal density of  $\mathbf{x}$  is  $f_2(\mathbf{x})$  and the conditional density of  $y$  given  $\mathbf{x}$  is  $f_1(y|\mathbf{x})(1 + (y - \mathbf{x}'\boldsymbol{\beta})(\mathbf{x}'\boldsymbol{\theta})/\sigma^2)$ . To see that the latter is a valid conditional density, observe that the regression assumption implies that  $\int y f_1(y|\mathbf{x}) dy = \mathbf{x}'\boldsymbol{\beta}$  and therefore

$$\begin{aligned} & \int f_1(y|\mathbf{x})(1 + (y - \mathbf{x}'\boldsymbol{\beta})(\mathbf{x}'\boldsymbol{\theta})/\sigma^2) dy \\ &= \int f_1(y|\mathbf{x}) dy + \int f_1(y|\mathbf{x})(y - \mathbf{x}'\boldsymbol{\beta})(\mathbf{x}'\boldsymbol{\theta})/\sigma^2 dy \\ &= 1. \end{aligned}$$

In this parametric submodel the conditional mean of  $y$  given  $\mathbf{x}$  is

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}(y|\mathbf{x}) &= \int y f_1(y|\mathbf{x})(1 + (y - \mathbf{x}'\boldsymbol{\beta})(\mathbf{x}'\boldsymbol{\theta})/\sigma^2) dy \\ &= \int y f_1(y|\mathbf{x}) dy + \int y f_1(y|\mathbf{x})(y - \mathbf{x}'\boldsymbol{\beta})(\mathbf{x}'\boldsymbol{\theta})/\sigma^2 dy \\ &= \int y f_1(y|\mathbf{x}) dy + \int (y - \mathbf{x}'\boldsymbol{\beta})^2 f_1(y|\mathbf{x})(\mathbf{x}'\boldsymbol{\theta})/\sigma^2 dy \\ &\quad + \int (y - \mathbf{x}'\boldsymbol{\beta}) f_1(y|\mathbf{x}) dy (\mathbf{x}'\boldsymbol{\beta})(\mathbf{x}'\boldsymbol{\theta})/\sigma^2 \\ &= \mathbf{x}'(\boldsymbol{\beta} + \boldsymbol{\theta}), \end{aligned}$$

using the homoskedasticity assumption  $\int (y - \mathbf{x}'\boldsymbol{\beta})^2 f_1(y|\mathbf{x}) dy = \sigma^2$ . This means that in this parametric submodel, the conditional mean is linear in  $\mathbf{x}$  and the regression coefficient is  $\boldsymbol{\beta}(\boldsymbol{\theta}) = \boldsymbol{\beta} + \boldsymbol{\theta}$ .

We now calculate the score for estimation of  $\boldsymbol{\theta}$ . Since

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log f(y, \mathbf{x} | \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log(1 + (y - \mathbf{x}'\boldsymbol{\beta})(\mathbf{x}'\boldsymbol{\theta})/\sigma^2) = \frac{\mathbf{x}(y - \mathbf{x}'\boldsymbol{\beta})/\sigma^2}{1 + (y - \mathbf{x}'\boldsymbol{\beta})(\mathbf{x}'\boldsymbol{\theta})/\sigma^2}$$

the score is

$$\mathbf{s} = \frac{\partial}{\partial \boldsymbol{\theta}} \log f(y, \mathbf{x} | \boldsymbol{\theta}_0) = \mathbf{x}e/\sigma^2.$$

The Cramer-Rao bound for estimation of  $\boldsymbol{\theta}$  (and therefore  $\boldsymbol{\beta}(\boldsymbol{\theta})$  as well) is

$$(\mathbb{E}(\mathbf{s}\mathbf{s}'))^{-1} = (\sigma^{-4} \mathbb{E}((\mathbf{x}e)(\mathbf{x}e)'))^{-1} = \sigma^2 \mathbf{Q}_{\mathbf{xx}}^{-1} = \mathbf{V}_{\boldsymbol{\beta}}^0.$$

We have shown that there is a parametric submodel (7.39) whose Cramer-Rao bound for estimation of  $\boldsymbol{\beta}$  is identical to the asymptotic variance of the least-squares estimator, which therefore is the semiparametric variance bound.

**Theorem 7.16** In the homoskedastic regression model, the semiparametric variance bound for estimation of  $\boldsymbol{\beta}$  is  $\mathbf{V}^0 = \sigma^2 \mathbf{Q}_{\mathbf{xx}}^{-1}$  and the OLS estimator is semiparametrically efficient.

This result is similar to the Gauss-Markov theorem, in that it asserts the efficiency of the least-squares estimator in the context of the homoskedastic regression model. The difference is that the Gauss-Markov theorem states that OLS has the smallest variance among the set of unbiased linear estimators, while Theorem 7.16 states that OLS has the smallest asymptotic variance among all regular estimators. This is a much more powerful statement.

## 7.22 Uniformly Consistent Residuals\*

It seems natural to view the residuals  $\hat{e}_i$  as estimators of the unknown errors  $e_i$ . Are they consistent? In this section we develop an appropriate convergence result. This is not a widely-used technique, and can safely be skipped by most readers.

Notice that we can write the residual as

$$\begin{aligned}\hat{e}_i &= y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}} \\ &= e_i + \mathbf{x}'_i \boldsymbol{\beta} - \mathbf{x}'_i \hat{\boldsymbol{\beta}} \\ &= e_i - \mathbf{x}'_i (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).\end{aligned}\tag{7.40}$$

Since  $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \xrightarrow{p} \mathbf{0}$  it seems reasonable to guess that  $\hat{e}_i$  will be close to  $e_i$  if  $n$  is large.

We can bound the difference in (7.40) using the Schwarz inequality (B.12) to find

$$|\hat{e}_i - e_i| = |\mathbf{x}'_i (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})| \leq \|\mathbf{x}_i\| \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|. \tag{7.41}$$

To bound (7.41) we can use  $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = O_p(n^{-1/2})$  from Theorem 7.3, but we also need to bound the random variable  $\|\mathbf{x}_i\|$ . If the regressor is bounded, that is,  $\|\mathbf{x}_i\| \leq B < \infty$ , then  $|\hat{e}_i - e_i| \leq B \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = O_p(n^{-1/2})$ . However if the regressor does not have bounded support then we have to be more careful.

The key is Theorem 6.38 which shows that  $\mathbb{E} \|\mathbf{x}_i\|^r < \infty$  implies  $\mathbf{x}_i = o_p(n^{1/r})$  uniformly in  $i$ , or

$$n^{-1/r} \max_{1 \leq i \leq n} \|\mathbf{x}_i\| \xrightarrow{p} 0.$$

Applied to (7.41) we obtain

$$\begin{aligned}\max_{1 \leq i \leq n} |\hat{e}_i - e_i| &\leq \max_{1 \leq i \leq n} \|\mathbf{x}_i\| \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| \\ &= o_p(n^{-1/2+1/r}).\end{aligned}$$

We have shown the following.

**Theorem 7.17** Under Assumption 7.2 and  $\mathbb{E} \|\mathbf{x}_i\|^r < \infty$ , then uniformly in  $1 \leq i \leq n$

$$\hat{e}_i = e_i + o_p(n^{-1/2+1/r}). \tag{7.42}$$

The rate of convergence in (7.42) depends on  $r$ . Assumption 7.2 requires  $r \geq 4$ , so the rate of convergence is at least  $o_p(n^{-1/4})$ . As  $r$  increases, the rate improves. As a limiting case, from Theorem 6.38 we see that if  $\mathbb{E}(\exp(\mathbf{t}' \mathbf{x}_i)) < \infty$  for some  $\mathbf{t} \neq 0$  then  $\mathbf{x}_i = o_p((\log n)^{1+\eta})$  uniformly in  $i$ , and thus  $\hat{e}_i = e_i + o_p(n^{-1/2} (\log n)^{1+\eta})$ .

We mentioned in Section 7.7 that there are multiple ways to prove the consistency of the covariance matrix estimator  $\hat{\Omega}$ . We now show that Theorem 7.17 provides one simple method to establish (7.23) and thus Theorem 7.6. Let  $q_n = \max_{1 \leq i \leq n} |\hat{e}_i - e_i| = o_p(n^{-1/4})$ . Since

$$\hat{e}_i^2 - e_i^2 = 2e_i(\hat{e}_i - e_i) + (\hat{e}_i - e_i)^2,$$

then

$$\begin{aligned}
\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i (\hat{e}_i^2 - e_i^2) \right\| &\leq \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i \mathbf{x}'_i\| |\hat{e}_i^2 - e_i^2| \\
&\leq \frac{2}{n} \sum_{i=1}^n \|\mathbf{x}_i\|^2 |e_i| |\hat{e}_i - e_i| + \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i\|^2 |\hat{e}_i - e_i|^2 \\
&\leq \frac{2}{n} \sum_{i=1}^n \|\mathbf{x}_i\|^2 |e_i| q_n + \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i\|^2 q_n^2 \\
&\leq o_p(n^{-1/4}).
\end{aligned}$$

## 7.23 Asymptotic Leverage\*

Recall the definition of leverage from (3.41)

$$h_{ii} = \mathbf{x}'_i (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i.$$

These are the diagonal elements of the projection matrix  $\mathbf{P}$  and appear in the formula for leave-one-out prediction errors and HC2 and HC3 covariance matrix estimators. We can show that under i.i.d. sampling the leverage values are uniformly asymptotically small.

Let  $\lambda_{\min}(\mathbf{A})$  and  $\lambda_{\max}(\mathbf{A})$  denote the smallest and largest eigenvalues of a symmetric square matrix  $\mathbf{A}$ , and note that  $\lambda_{\max}(\mathbf{A}^{-1}) = (\lambda_{\min}(\mathbf{A}))^{-1}$ .

Since  $\frac{1}{n} \mathbf{X}' \mathbf{X} \xrightarrow{p} \mathbf{Q}_{xx} > 0$  then by the CMT,  $\lambda_{\min}(\frac{1}{n} \mathbf{X}' \mathbf{X}) \xrightarrow{p} \lambda_{\min}(\mathbf{Q}_{xx}) > 0$ . (The latter is positive since  $\mathbf{Q}_{xx}$  is positive definite and thus all its eigenvalues are positive.) Then by the Quadratic Inequality (B.18)

$$\begin{aligned}
h_{ii} &= \mathbf{x}'_i (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i \\
&\leq \lambda_{\max}((\mathbf{X}' \mathbf{X})^{-1})(\mathbf{x}'_i \mathbf{x}_i) \\
&= \left( \lambda_{\min}\left(\frac{1}{n} \mathbf{X}' \mathbf{X}\right) \right)^{-1} \frac{1}{n} \|\mathbf{x}_i\|^2 \\
&\leq (\lambda_{\min}(\mathbf{Q}_{xx}) + o_p(1))^{-1} \frac{1}{n} \max_{1 \leq i \leq n} \|\mathbf{x}_i\|^2.
\end{aligned} \tag{7.43}$$

Theorem 6.38 shows that  $\mathbb{E}(\|\mathbf{x}_i\|^r) < \infty$  implies  $\max_{1 \leq i \leq n} \|\mathbf{x}_i\|^2 = (\max_{1 \leq i \leq n} \|\mathbf{x}_i\|)^2 = o_p(n^{2/r})$  and thus (7.43) is  $o_p(n^{2/r-1})$ .

**Theorem 7.18** If  $\mathbf{x}_i$  is i.i.d.,  $\mathbf{Q}_{xx} > 0$ , and  $\mathbb{E}(\|\mathbf{x}_i\|^r) < \infty$  for some  $r \geq 2$ , then uniformly in  $1 \leq i \leq n$ ,  $h_{ii} = o_p(n^{2/r-1})$ .

For any  $r \geq 2$  then  $h_{ii} = o_p(1)$  (uniformly in  $i \leq n$ ). Larger  $r$  implies a stronger rate of convergence, for example  $r = 4$  implies  $h_{ii} = o_p(n^{-1/2})$ .

Theorem (7.18) implies that under random sampling with finite variances and large samples, no individual observation should have a large leverage value. Consequently individual observations should not be influential, unless one of these conditions is violated.

## Exercises

**Exercise 7.1** Take the model  $y_i = \mathbf{x}'_{1i}\boldsymbol{\beta}_1 + \mathbf{x}'_{2i}\boldsymbol{\beta}_2 + e_i$  with  $\mathbb{E}(\mathbf{x}_i e_i) = \mathbf{0}$ . Suppose that  $\boldsymbol{\beta}_1$  is estimated by regressing  $y_i$  on  $\mathbf{x}_{1i}$  only. Find the probability limit of this estimator. In general, is it consistent for  $\boldsymbol{\beta}_1$ ? If not, under what conditions is this estimator consistent for  $\boldsymbol{\beta}_1$ ?

**Exercise 7.2** Let  $\mathbf{y}$  be  $n \times 1$ ,  $\mathbf{X}$  be  $n \times k$  (rank  $k$ ).  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$  with  $\mathbb{E}(\mathbf{x}_i e_i) = \mathbf{0}$ . Define the *ridge regression* estimator

$$\hat{\boldsymbol{\beta}} = \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i + \lambda \mathbf{I}_k \right)^{-1} \left( \sum_{i=1}^n \mathbf{x}_i y_i \right) \quad (7.44)$$

here  $\lambda > 0$  is a fixed constant. Find the probability limit of  $\hat{\boldsymbol{\beta}}$  as  $n \rightarrow \infty$ . Is  $\hat{\boldsymbol{\beta}}$  consistent for  $\boldsymbol{\beta}$ ?

**Exercise 7.3** For the ridge regression estimator (7.44), set  $\lambda = cn$  where  $c > 0$  is fixed as  $n \rightarrow \infty$ . Find the probability limit of  $\hat{\boldsymbol{\beta}}$  as  $n \rightarrow \infty$ .

**Exercise 7.4** Verify some of the calculations reported in Section 7.4. Specifically, suppose that  $x_{1i}$  and  $x_{2i}$  only take the values  $\{-1, +1\}$ , symmetrically, with

$$\begin{aligned} \mathbb{P}(x_{1i} = x_{2i} = 1) &= \mathbb{P}(x_{1i} = x_{2i} = -1) = 3/8 \\ \mathbb{P}(x_{1i} = 1, x_{2i} = -1) &= \mathbb{P}(x_{1i} = -1, x_{2i} = 1) = 1/8 \\ \mathbb{E}(e_i^2 | x_{1i} = x_{2i}) &= \frac{5}{4} \\ \mathbb{E}(e_i^2 | x_{1i} \neq x_{2i}) &= \frac{1}{4}. \end{aligned}$$

Verify the following:

- (a)  $\mathbb{E}(x_{1i}) = 0$
- (b)  $\mathbb{E}(x_{1i}^2) = 1$
- (c)  $\mathbb{E}(x_{1i}x_{2i}) = \frac{1}{2}$
- (d)  $\mathbb{E}(e_i^2) = 1$
- (e)  $\mathbb{E}(x_{1i}^2 e_i^2) = 1$
- (f)  $\mathbb{E}(x_{1i}x_{2i}e_i^2) = \frac{7}{8}$ .

**Exercise 7.5** Show (7.13)-(7.16).

**Exercise 7.6** The model is

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) &= \mathbf{0} \\ \boldsymbol{\Omega} &= \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i e_i^2). \end{aligned}$$

Find the method of moments estimators  $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Omega}})$  for  $(\boldsymbol{\beta}, \boldsymbol{\Omega})$ .

- (a) In this model, are  $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Omega}})$  efficient estimators of  $(\boldsymbol{\beta}, \boldsymbol{\Omega})$ ?
- (b) If so, in what sense are they efficient?

**Exercise 7.7** Of the variables  $(y_i^*, y_i, \mathbf{x}_i)$  only the pair  $(y_i, \mathbf{x}_i)$  are observed. In this case, we say that  $y_i^*$  is a *latent* variable. Suppose

$$\begin{aligned} y_i^* &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) &= \mathbf{0} \\ y_i &= y_i^* + u_i \end{aligned}$$

where  $u_i$  is a measurement error satisfying

$$\begin{aligned} \mathbb{E}(\mathbf{x}_i u_i) &= \mathbf{0} \\ \mathbb{E}(y_i^* u_i) &= 0 \end{aligned}$$

Let  $\hat{\boldsymbol{\beta}}$  denote the OLS coefficient from the regression of  $y_i$  on  $\mathbf{x}_i$ .

- (a) Is  $\boldsymbol{\beta}$  the coefficient from the linear projection of  $y_i$  on  $\mathbf{x}_i$ ?
- (b) Is  $\hat{\boldsymbol{\beta}}$  consistent for  $\boldsymbol{\beta}$  as  $n \rightarrow \infty$ ?
- (c) Find the asymptotic distribution of  $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$  as  $n \rightarrow \infty$ .

**Exercise 7.8** Find the asymptotic distribution of  $\sqrt{n}(\hat{\sigma}^2 - \sigma^2)$  as  $n \rightarrow \infty$ .

**Exercise 7.9** The model is

$$\begin{aligned} y_i &= x_i \beta + e_i \\ \mathbb{E}(e_i | x_i) &= 0 \end{aligned}$$

where  $x_i \in \mathbb{R}$ . Consider the two estimators

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \\ \tilde{\beta} &= \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i}. \end{aligned}$$

- (a) Under the stated assumptions, are both estimators consistent for  $\beta$ ?
- (b) Are there conditions under which either estimator is efficient?

**Exercise 7.10** In the homoskedastic regression model  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$  with  $\mathbb{E}(e_i | \mathbf{x}_i) = 0$  and  $\mathbb{E}(e_i^2 | \mathbf{x}_i) = \sigma^2$ , suppose  $\hat{\boldsymbol{\beta}}$  is the OLS estimator of  $\boldsymbol{\beta}$  with covariance matrix estimator  $\hat{V}_{\hat{\boldsymbol{\beta}}}$ , based on a sample of size  $n$ . Let  $\hat{\sigma}^2$  be the estimator of  $\sigma^2$ . You wish to forecast an out-of-sample value of  $y_{n+1}$  given that  $\mathbf{x}_{n+1} = \mathbf{x}$ . Thus the available information is the sample  $(\mathbf{y}, \mathbf{X})$ , the estimates  $(\hat{\boldsymbol{\beta}}, \hat{V}_{\hat{\boldsymbol{\beta}}}, \hat{\sigma}^2)$ , the residuals  $\hat{\mathbf{e}}$ , and the out-of-sample value of the regressors,  $\mathbf{x}_{n+1}$ .

- (a) Find a point forecast of  $y_{n+1}$ .
- (b) Find an estimator of the variance of this forecast.

**Exercise 7.11** Take a regression model with i.i.d. observations  $(y_i, x_i)$  and scalar  $x_i$

$$\begin{aligned} y_i &= x_i \beta + e_i \\ \mathbb{E}(e_i | x_i) &= 0 \\ \Omega &= \mathbb{E}(x_i^2 e_i^2) \end{aligned}$$

Let  $\hat{\beta}$  be the OLS estimator of  $\beta$  with residuals  $\hat{e}_i = y_i - x_i \hat{\beta}$ . Consider the estimators of  $\Omega$

$$\begin{aligned}\tilde{\Omega} &= \frac{1}{n} \sum_{i=1}^n x_i^2 e_i^2 \\ \hat{\Omega} &= \frac{1}{n} \sum_{i=1}^n x_i^2 \hat{e}_i^2\end{aligned}$$

- (a) Find the asymptotic distribution of  $\sqrt{n}(\tilde{\Omega} - \Omega)$  as  $n \rightarrow \infty$ .
- (b) Find the asymptotic distribution of  $\sqrt{n}(\hat{\Omega} - \Omega)$  as  $n \rightarrow \infty$ .
- (c) How do you use the regression assumption  $\mathbb{E}(e_i | x_i) = 0$  in your answer to (b)?

**Exercise 7.12** Consider the model

$$\begin{aligned}y_i &= \alpha + \beta x_i + e_i \\ \mathbb{E}(e_i) &= 0 \\ \mathbb{E}(x_i e_i) &= 0\end{aligned}$$

with both  $y_i$  and  $x_i$  scalar. Assuming  $\alpha > 0$  and  $\beta < 0$ , suppose the parameter of interest is the area under the regression curve (e.g. consumer surplus), which is  $A = -\alpha^2/2\beta$ .

Let  $\hat{\theta} = (\hat{\alpha}, \hat{\beta})'$  be the least-squares estimators of  $\theta = (\alpha, \beta)'$  so that  $\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d N(0, V_\theta)$  and let  $\hat{V}_\theta$  be a standard consistent estimator for  $V_\theta$ .

- (a) Given the above, describe an estimator of  $A$ .
- (b) Construct an asymptotic  $(1 - \eta)$  confidence interval for  $A$ .

**Exercise 7.13** Consider an i.i.d. sample  $\{y_i, x_i\} i = 1, \dots, n$  where  $y_i$  and  $x_i$  are scalar. Consider the reverse projection model

$$\begin{aligned}x_i &= y_i \gamma + u_i \\ \mathbb{E}(y_i u_i) &= 0\end{aligned}$$

and define the parameter of interest as  $\theta = 1/\gamma$

- (a) Propose an estimator  $\hat{\gamma}$  of  $\gamma$ .
- (b) Propose an estimator  $\hat{\theta}$  of  $\theta$ .
- (c) Find the asymptotic distribution of  $\hat{\theta}$ .
- (d) Find an asymptotic standard error for  $\hat{\theta}$ .

**Exercise 7.14** Take the model

$$\begin{aligned}y_i &= x_{1i} \beta_1 + x_{2i} \beta_2 + e_i \\ \mathbb{E}(x_i e_i) &= 0\end{aligned}$$

with both  $\beta_1 \in \mathbb{R}$  and  $\beta_2 \in \mathbb{R}$ , and define the parameter

$$\theta = \beta_1 \beta_2$$

- (a) What is the appropriate estimator  $\hat{\theta}$  for  $\theta$ ?
- (b) Find the asymptotic distribution of  $\hat{\theta}$  under standard regularity conditions.

- (c) Show how to calculate an asymptotic 95% confidence interval for  $\theta$ .

**Exercise 7.15** Take the linear model

$$y_i = x_i \beta + e_i$$

$$\mathbb{E}(e_i | x_i) = 0$$

with  $n$  observations and  $x_i$  is scalar (real-valued). Consider the estimator

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i^3 y_i}{\sum_{i=1}^n x_i^4}$$

Find the asymptotic distribution of  $\sqrt{n}(\hat{\beta} - \beta)$  as  $n \rightarrow \infty$ .

**Exercise 7.16** Out of an i.i.d. sample  $(y_i, \mathbf{x}_i)$  of size  $n$ , you randomly take half the observations and estimate the least-squares regression of  $y_i$  on  $\mathbf{x}_i$  using only this sub-sample.

$$y_i = \mathbf{x}'_i \hat{\beta} + \hat{e}_i$$

Is the estimated slope coefficient  $\hat{\beta}$  consistent for the population projection coefficient? Explain your reasoning.

**Exercise 7.17** An economist reports a set of parameter estimates, including the coefficient estimates  $\hat{\beta}_1 = 1.0$ ,  $\hat{\beta}_2 = 0.8$ , and standard errors  $s(\hat{\beta}_1) = 0.07$  and  $s(\hat{\beta}_2) = 0.07$ . The author writes “The estimates show that  $\beta_1$  is larger than  $\beta_2$ .”

- (a) Write down the formula for an asymptotic 95% confidence interval for  $\theta = \beta_1 - \beta_2$ , expressed as a function of  $\hat{\beta}_1$ ,  $\hat{\beta}_2$ ,  $s(\hat{\beta}_1)$ ,  $s(\hat{\beta}_2)$  and  $\hat{\rho}$ , where  $\hat{\rho}$  is the estimated correlation between  $\hat{\beta}_1$  and  $\hat{\beta}_2$ .
- (b) Can  $\hat{\rho}$  be calculated from the reported information?
- (c) Is the author correct? Does the reported information support the author’s claim?

**Exercise 7.18** Suppose an economic model suggests

$$g(x) = \mathbb{E}(y_i | x_i = x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

where  $x_i \in \mathbb{R}$ . You have a random sample  $(y_i, x_i)$ ,  $i = 1, \dots, n$ .

- (a) Describe how to estimate  $g(x)$  at a given value  $x$ .
- (b) Describe (be specific) an appropriate confidence interval for  $g(x)$ .

**Exercise 7.19** Take the model

$$y_i = \mathbf{x}'_i \beta + e_i$$

$$\mathbb{E}(\mathbf{x}_i e_i) = \mathbf{0}$$

and suppose you have observations  $i = 1, \dots, 2n$ . (The number of observations is  $2n$ .) You randomly split the sample in half, (each has  $n$  observations), calculate  $\hat{\beta}_1$  by least-squares on the first sample, and  $\hat{\beta}_2$  by least-squares on the second sample. What is the asymptotic distribution of  $\sqrt{n}(\hat{\beta}_1 - \hat{\beta}_2)$ ?

**Exercise 7.20** The data  $\{y_i, \mathbf{x}_i, w_i\}$  is from a random sample,  $i = 1, \dots, n$ . The parameter  $\beta$  is estimated by minimizing the criterion function

$$S(\beta) = \sum_{i=1}^n w_i (y_i - \mathbf{x}'_i \beta)^2$$

That is  $\hat{\beta} = \operatorname{argmin}_{\beta} S(\beta)$ .

- (a) Find an explicit expression for  $\hat{\beta}$ .
- (b) What population parameter  $\beta$  is  $\hat{\beta}$  estimating? (Be explicit about any assumptions you need to impose. But don't make more assumptions than necessary.)
- (c) Find the probability limit for  $\hat{\beta}$  as  $n \rightarrow \infty$ .
- (d) Find the asymptotic distribution of  $\sqrt{n}(\hat{\beta} - \beta)$  as  $n \rightarrow \infty$ .

**Exercise 7.21** Take the model

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(e_i | \mathbf{x}_i) &= 0 \\ \mathbb{E}(e_i^2 | \mathbf{x}_i) &= \sigma_e^2 = \mathbf{z}'_i \boldsymbol{\gamma} \end{aligned}$$

where  $\mathbf{z}_i$  is a (vector) function of  $\mathbf{x}_i$ . The sample is  $i = 1, \dots, n$  with i.i.d. observations. For simplicity, assume that  $\mathbf{z}'_i \boldsymbol{\gamma} > 0$  for all  $\mathbf{z}_i$ . Suppose you are interested in forecasting  $y_{n+1}$  given  $\mathbf{x}_{n+1} = \mathbf{x}$  and  $\mathbf{z}_{n+1} = \mathbf{z}$  for some out-of-sample observation  $n+1$ . Describe how you would construct a point forecast and a forecast interval for  $y_{n+1}$ .

**Exercise 7.22** Take the model

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(e_i | \mathbf{x}_i) &= 0 \\ z_i &= (\mathbf{x}'_i \boldsymbol{\beta}) \gamma + u_i \\ \mathbb{E}(u_i | \mathbf{x}_i) &= 0 \end{aligned}$$

Your goal is to estimate  $\gamma$ . (Note that  $\gamma$  is scalar.) You use a two-step estimator:

- Estimate  $\hat{\beta}$  by least-squares of  $y_i$  on  $\mathbf{x}_i$ .
- Estimate  $\hat{\gamma}$  by least-squares of  $z_i$  on  $\mathbf{x}'_i \hat{\beta}$ .

- (a) Show that  $\hat{\gamma}$  is consistent for  $\gamma$ .
- (b) Find the asymptotic distribution of  $\hat{\gamma}$  when  $\gamma = 0$ .

**Exercise 7.23** The model is

$$\begin{aligned} y_i &= x_i \beta + e_i \\ \mathbb{E}(e_i | x_i) &= 0 \end{aligned}$$

where  $x_i \in R$ . Consider the the estimator

$$\tilde{\beta} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i}.$$

Find conditions under which  $\tilde{\beta}$  is consistent for  $\beta$  as  $n \rightarrow \infty$ .

**Exercise 7.24** Of the random variables  $(y_i^*, y_i, \mathbf{x}_i)$  only the pair  $(y_i, \mathbf{x}_i)$  are observed. (In this case, we say that  $y_i^*$  is a *latent* variable.) Suppose  $\mathbb{E}(y_i^* | \mathbf{x}_i) = \mathbf{x}'_i \boldsymbol{\beta}$  and  $y = y_i^* + u_i$ , where  $u_i$  is a measurement error satisfying  $\mathbb{E}(u_i | y_i^*, \mathbf{x}_i) = 0$ . Let  $\hat{\beta}$  denote the OLS coefficient from the regression of  $y_i$  on  $\mathbf{x}_i$ .

- (a) Find  $\mathbb{E}(y_i | \mathbf{x}_i)$ .
- (b) Is  $\hat{\beta}$  consistent for  $\boldsymbol{\beta}$  as  $n \rightarrow \infty$ ?

(c) Find the asymptotic distribution of  $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$  as  $n \rightarrow \infty$ .

**Exercise 7.25** The parameter  $\beta$  is defined in the model

$$y_i = x_i^* \beta + e_i$$

where  $e_i$  is independent of  $x_i^*$ ,  $\mathbb{E}(e_i) = 0$ ,  $\mathbb{E}(e_i^2) = \sigma^2$ . The observables are  $(y_i, x_i)$  where

$$x_i = x_i^* \nu_i$$

and  $\nu_i > 0$  is random measurement error. Assume that  $\nu_i$  is independent of  $x_i^*$  and  $e_i$ . Also assume that  $x_i$  and  $x_i^*$  are non-negative and real-valued. Consider the least-squares estimator  $\hat{\beta}$  for  $\beta$ .

(a) Find the plim of  $\hat{\beta}$ , expressed in terms of  $\beta$  and moments of  $(x_i, \nu_i, e_i)$ .

(b) Can you find a non-trivial condition under which  $\hat{\beta}$  is consistent for  $\beta$ ? (By non-trivial, we mean something other than  $\nu_i = 1$ .)

**Exercise 7.26** Take the standard model

$$\begin{aligned} y_i &= \mathbf{x}_i' \boldsymbol{\beta} + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) &= \mathbf{0} \end{aligned}$$

For a positive function  $w(\mathbf{x})$ , let  $w_i = w(\mathbf{x}_i)$ . Consider the estimator

$$\tilde{\boldsymbol{\beta}} = \left( \sum_{i=1}^n w_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left( \sum_{i=1}^n w_i \mathbf{x}_i y_i \right).$$

Find the probability limit (as  $n \rightarrow \infty$ ) of  $\tilde{\boldsymbol{\beta}}$ . (Do you need to add an assumption?) Is  $\tilde{\boldsymbol{\beta}}$  consistent for  $\boldsymbol{\beta}$ ? If not, under what assumption is  $\tilde{\boldsymbol{\beta}}$  consistent for  $\boldsymbol{\beta}$ ?

**Exercise 7.27** Take the regression model

$$\begin{aligned} y_i &= \mathbf{x}_i' \boldsymbol{\beta} + e_i \\ \mathbb{E}(e_i | \mathbf{x}_i) &= 0 \\ \mathbb{E}(e_i^2 | \mathbf{x}_i) &= \sigma_i^2 \end{aligned}$$

with  $\mathbf{x}_i \in R^k$ . Assume that  $\mathbb{P}(e_i = 0) = 0$ . Consider the infeasible estimator

$$\tilde{\boldsymbol{\beta}} = \left( \sum_{i=1}^n e_i^{-2} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left( \sum_{i=1}^n e_i^{-2} \mathbf{x}_i y_i \right).$$

This is a WLS estimator using the weights  $e_i^{-2}$ .

(a) Find the asymptotic distribution of  $\tilde{\boldsymbol{\beta}}$ .

(b) Contrast your result with the asymptotic distribution of infeasible GLS.

**Exercise 7.28** The model is

$$\begin{aligned} y_i &= \mathbf{x}_i' \boldsymbol{\beta} + e_i \\ \mathbb{E}(e_i | \mathbf{x}_i) &= 0. \end{aligned}$$

An econometrician is worried about the impact of some unusually large values of the regressors. The model is thus estimated on the subsample for which  $|\mathbf{x}_i| \leq c$ , for some fixed  $c$ . Let  $\tilde{\boldsymbol{\beta}}$  denote the OLS estimator on this subsample. It equals

$$\tilde{\boldsymbol{\beta}} = \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \mathbf{1}(|\mathbf{x}_i| \leq c) \right)^{-1} \left( \sum_{i=1}^n \mathbf{x}_i y_i \mathbf{1}(|\mathbf{x}_i| \leq c) \right)$$

where  $\mathbf{1}(\cdot)$  denotes the indicator function.

- (a) Show that  $\tilde{\beta} \rightarrow_p \beta$ .
- (b) Find the asymptotic distribution of  $\sqrt{n}(\tilde{\beta} - \beta)$ .

**Exercise 7.29** As in Exercise 3.26, use the CPS dataset and the subsample of white male Hispanics. Estimate the regression

$$\widehat{\log(Wage)} = \beta_1 \text{education} + \beta_2 \text{experience} + \beta_3 \text{experience}^2/100 + \beta_4.$$

- (a) Report the coefficients and robust standard errors.
- (b) Let  $\theta$  be the ratio of the return to one year of education to the return to one year of experience. Write  $\theta$  as a function of the regression coefficients and variables. Compute  $\hat{\theta}$  from the estimated model.
- (c) Write out the formula for the asymptotic standard error for  $\hat{\theta}$  as a function of the covariance matrix for  $\hat{\beta}$ . Compute  $\hat{s}(\hat{\theta})$  from the estimated model.
- (d) Construct a 90% asymptotic confidence interval for  $\theta$  from the estimated model.
- (e) Compute the regression function at  $edu = 12$  and  $experience=20$ . Compute a 95% confidence interval for the regression function at this point.
- (f) Consider an out-of-sample individual with 16 years of education and 5 years experience. Construct an 80% forecast interval for their log wage and wage. [To obtain the forecast interval for the wage, apply the exponential function to both endpoints.]

# Chapter 8

## Restricted Estimation

### 8.1 Introduction

In the linear projection model

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) = \mathbf{0}$$

a common task is to impose a constraint on the coefficient vector  $\boldsymbol{\beta}$ . For example, partitioning  $\mathbf{x}'_i = (\mathbf{x}'_{1i}, \mathbf{x}'_{2i})$  and  $\boldsymbol{\beta}' = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2)$ , a typical constraint is an exclusion restriction of the form  $\boldsymbol{\beta}_2 = \mathbf{0}$ . In this case the constrained model is

$$y_i = \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) = \mathbf{0}.$$

At first glance this appears the same as the linear projection model, but there is one important difference: the error  $e_i$  is uncorrelated with the entire regressor vector  $\mathbf{x}'_i = (\mathbf{x}'_{1i}, \mathbf{x}'_{2i})$  not just the included regressor  $\mathbf{x}_{1i}$ .

In general, a set of  $q$  linear constraints on  $\boldsymbol{\beta}$  takes the form

$$\mathbf{R}' \boldsymbol{\beta} = \mathbf{c} \tag{8.1}$$

where  $\mathbf{R}$  is  $k \times q$ ,  $\text{rank}(\mathbf{R}) = q < k$  and  $\mathbf{c}$  is  $q \times 1$ . The assumption that  $\mathbf{R}$  is full rank means that the constraints are linearly independent (there are no redundant or contradictory constraints). We can define the restricted parameter space  $\mathbf{B}_R$  as the set of values of  $\boldsymbol{\beta}$  which satisfy (8.1), that is

$$\mathbf{B}_R = \{\boldsymbol{\beta} : \mathbf{R}' \boldsymbol{\beta} = \mathbf{c}\}.$$

Sometimes we will call (8.1) a **constraint** and sometimes a **restriction**. They are the same thing. Similarly sometimes we will call estimators which satisfy (8.1) **constrained estimators** and sometimes **restricted estimators**. Again, they mean the same thing.

The constraint  $\boldsymbol{\beta}_2 = \mathbf{0}$  discussed above is a special case of the constraint (8.1) with

$$\mathbf{R} = \begin{pmatrix} \mathbf{0} \\ \mathbf{I}_{k_2} \end{pmatrix}, \tag{8.2}$$

a selector matrix, and  $\mathbf{c} = \mathbf{0}$ .

Another common restriction is that a set of coefficients sum to a known constant, i.e.  $\beta_1 + \beta_2 = 1$ . For example, this constraint arises in a constant-return-to-scale production function. Other common restrictions include the equality of coefficients  $\beta_1 = \beta_2$ , and equal and offsetting coefficients  $\beta_1 = -\beta_2$ .

A typical reason to impose a constraint is that we believe (or have information) that the constraint is true. By imposing the constraint we hope to improve estimation efficiency. The goal is to obtain consistent estimates with reduced variance relative to the unconstrained estimator.

The questions then arise: How should we estimate the coefficient vector  $\beta$  imposing the linear restriction (8.1)? If we impose such constraints, what is the sampling distribution of the resulting estimator? How should we calculate standard errors? These are the questions explored in this chapter.

## 8.2 Constrained Least Squares

An intuitively appealing method to estimate a constrained linear projection is to minimize the least-squares criterion subject to the constraint  $R'\beta = c$ .

The constrained least-squares estimator is

$$\tilde{\beta}_{\text{cls}} = \underset{R'\beta=c}{\operatorname{argmin}} \text{SSE}(\beta) \quad (8.3)$$

where

$$\text{SSE}(\beta) = \sum_{i=1}^n (y_i - x'_i \beta)^2 = y'y - 2y'X\beta + \beta'X'X\beta. \quad (8.4)$$

The estimator  $\tilde{\beta}_{\text{cls}}$  minimizes the sum of squared errors over all  $\beta$  such that  $\beta \in B_R$ , or equivalently such that the restriction (8.1) holds. We call  $\tilde{\beta}_{\text{cls}}$  the **constrained least-squares** (CLS) estimator. We follow the convention of using a tilde “~” rather than a hat “^” to indicate that  $\tilde{\beta}_{\text{cls}}$  is a restricted estimator in contrast to the unrestricted least-squares estimator  $\hat{\beta}$ , and write it as  $\tilde{\beta}_{\text{cls}}$  to be clear that the estimation method is CLS.

One method to find the solution to (8.3) uses the technique of Lagrange multipliers. The problem (8.3) is equivalent to the minimization of the Lagrangian

$$\mathcal{L}(\beta, \lambda) = \frac{1}{2} \text{SSE}(\beta) + \lambda' (R'\beta - c) \quad (8.5)$$

over  $(\beta, \lambda)$ , where  $\lambda$  is an  $s \times 1$  vector of Lagrange multipliers. The first-order conditions for minimization of (8.5) are

$$\frac{\partial}{\partial \beta} \mathcal{L}(\tilde{\beta}_{\text{cls}}, \tilde{\lambda}_{\text{cls}}) = -X'y + X'X\tilde{\beta}_{\text{cls}} + R\tilde{\lambda}_{\text{cls}} = \mathbf{0} \quad (8.6)$$

and

$$\frac{\partial}{\partial \lambda} \mathcal{L}(\tilde{\beta}_{\text{cls}}, \tilde{\lambda}_{\text{cls}}) = R'\tilde{\beta} - c = \mathbf{0}. \quad (8.7)$$

Premultiplying (8.6) by  $R'(X'X)^{-1}$  we obtain

$$-R'\hat{\beta} + R'\tilde{\beta}_{\text{cls}} + R'(X'X)^{-1}R\tilde{\lambda}_{\text{cls}} = \mathbf{0}$$

where  $\hat{\beta} = (X'X)^{-1}X'y$  is the unrestricted least-squares estimator. Imposing  $R'\tilde{\beta}_{\text{cls}} - c = \mathbf{0}$  from (8.7) and solving for  $\tilde{\lambda}_{\text{cls}}$  we find

$$\tilde{\lambda}_{\text{cls}} = \left[ R'(X'X)^{-1}R \right]^{-1} (R'\hat{\beta} - c).$$

Notice that  $(X'X)^{-1} > 0$  and  $R$  full rank imply that  $R'(X'X)^{-1}R > 0$  and is hence invertible. (See Section A.10.)

Substituting this expression into (8.6) and solving for  $\tilde{\beta}_{\text{cls}}$  we find the solution to the constrained minimization problem (8.3)

$$\tilde{\beta}_{\text{cls}} = \hat{\beta}_{\text{ols}} - (X'X)^{-1}R \left[ R'(X'X)^{-1}R \right]^{-1} (R'\hat{\beta}_{\text{ols}} - c). \quad (8.8)$$

(See Exercise 8.5 to verify that (8.8) satisfies (8.1).)

This is a general formula for the CLS estimator. It also can be written as

$$\tilde{\boldsymbol{\beta}}_{\text{cls}} = \hat{\boldsymbol{\beta}}_{\text{ols}} - \hat{\mathbf{Q}}_{\mathbf{xx}}^{-1} \mathbf{R} \left( \mathbf{R}' \hat{\mathbf{Q}}_{\mathbf{xx}}^{-1} \mathbf{R} \right)^{-1} (\mathbf{R}' \hat{\boldsymbol{\beta}}_{\text{ols}} - \mathbf{c}). \quad (8.9)$$

The CLS residuals are

$$\tilde{e}_i = y_i - \mathbf{x}'_i \tilde{\boldsymbol{\beta}}_{\text{cls}}$$

and the  $n \times 1$  vector of residuals are written in vector notation as  $\tilde{\mathbf{e}}$ .

To illustrate, we generated a random sample of 100 observations for the variables  $(y_i, x_{1i}, x_{2i})$  and calculated the sum of squared errors function for the regression of  $y_i$  on  $x_{1i}$  and  $x_{2i}$ . Figure 8.1 displays contour plots of the sum of squared errors function. The center of the contour plots is the least squares minimizer  $\hat{\boldsymbol{\beta}}_{\text{ols}} = (0.33, 0.26)'$ . Suppose it is desired to estimate the coefficients subject to the constraint  $\beta_1 + \beta_2 = 1$ . This constraint is displayed in the figure by the straight line. The constrained least squares estimator is the point on this straight line which yields the smallest sum of squared errors, which is the point which intersects with the lowest contour plot. The solution is the point where a contour plot is tangent to the constraint line, and marked as  $\tilde{\boldsymbol{\beta}}_{\text{cls}} = (0.52, 0.48)'$ .

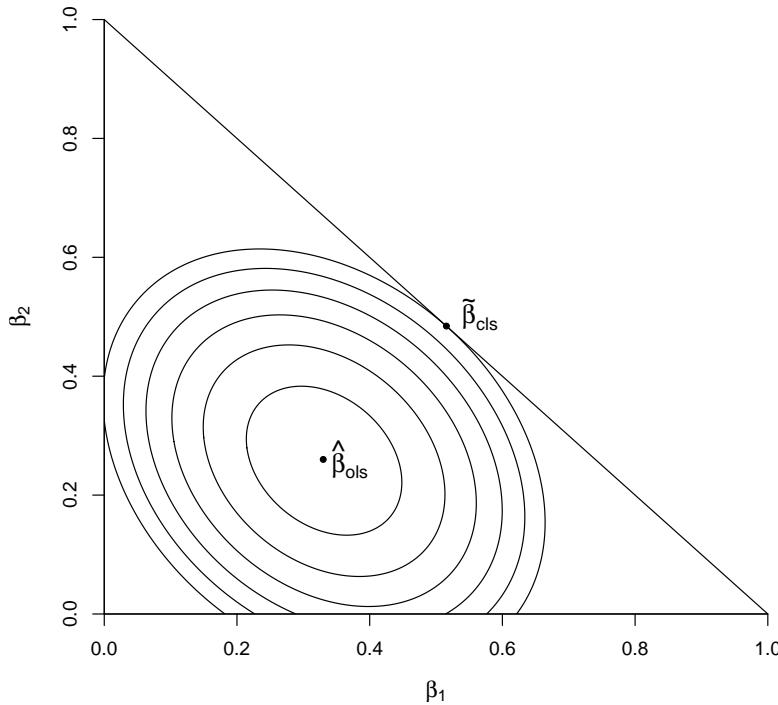


Figure 8.1: Imposing a Constraint on the Least Squares Criterion

In Stata, constrained least squares is implemented using the `cnsreg` command.

### 8.3 Exclusion Restriction

While (8.8) is a general formula for the CLS estimator, in most cases the estimator can be found by applying least-squares to a reparameterized equation. To illustrate, let us return to the first example presented at the beginning of the chapter – a simple exclusion restriction. Recall the unconstrained model is

$$y_i = \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + e_i \quad (8.10)$$

the exclusion restriction is  $\beta_2 = \mathbf{0}$ , and the constrained equation is

$$y_i = \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + e_i. \quad (8.11)$$

In this setting the CLS estimator is OLS of  $y_i$  on  $\mathbf{x}_{1i}$ . (See Exercise 8.1.) We can write this as

$$\tilde{\boldsymbol{\beta}}_1 = \left( \sum_{i=1}^n \mathbf{x}_{1i} \mathbf{x}'_{1i} \right)^{-1} \left( \sum_{i=1}^n \mathbf{x}_{1i} y_i \right). \quad (8.12)$$

The CLS estimator of the entire vector  $\boldsymbol{\beta}' = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2)$  is

$$\tilde{\boldsymbol{\beta}} = \begin{pmatrix} \tilde{\boldsymbol{\beta}}_1 \\ \mathbf{0} \end{pmatrix}. \quad (8.13)$$

It is not immediately obvious, but (8.8) and (8.13) are algebraically (and numerically) equivalent. To see this, the first component of (8.8) with (8.2) is

$$\tilde{\boldsymbol{\beta}}_1 = (\mathbf{I}_{k_2} \quad \mathbf{0}) \left[ \hat{\boldsymbol{\beta}} - \hat{\mathbf{Q}}_{\mathbf{xx}}^{-1} \begin{pmatrix} \mathbf{0} \\ \mathbf{I}_{k_2} \end{pmatrix} \left[ (\mathbf{0} \quad \mathbf{I}_{k_2}) \hat{\mathbf{Q}}_{\mathbf{xx}}^{-1} \begin{pmatrix} \mathbf{0} \\ \mathbf{I}_{k_2} \end{pmatrix} \right]^{-1} (\mathbf{0} \quad \mathbf{I}_{k_2}) \hat{\boldsymbol{\beta}} \right].$$

Using (3.40) this equals

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_1 &= \hat{\boldsymbol{\beta}}_1 - \hat{\mathbf{Q}}^{12} \left( \hat{\mathbf{Q}}^{22} \right)^{-1} \hat{\boldsymbol{\beta}}_2 \\ &= \hat{\boldsymbol{\beta}}_1 + \hat{\mathbf{Q}}_{11 \cdot 2}^{-1} \hat{\mathbf{Q}}_{12} \hat{\mathbf{Q}}_{22 \cdot 1}^{-1} \hat{\boldsymbol{\beta}}_2 \\ &= \hat{\mathbf{Q}}_{11 \cdot 2}^{-1} \left( \hat{\mathbf{Q}}_{1y} - \hat{\mathbf{Q}}_{12} \hat{\mathbf{Q}}_{22}^{-1} \hat{\mathbf{Q}}_{2y} \right) \\ &\quad + \hat{\mathbf{Q}}_{11 \cdot 2}^{-1} \hat{\mathbf{Q}}_{12} \hat{\mathbf{Q}}_{22}^{-1} \hat{\mathbf{Q}}_{22 \cdot 1} \hat{\mathbf{Q}}_{22 \cdot 1}^{-1} \left( \hat{\mathbf{Q}}_{2y} - \hat{\mathbf{Q}}_{21} \hat{\mathbf{Q}}_{11}^{-1} \hat{\mathbf{Q}}_{1y} \right) \\ &= \hat{\mathbf{Q}}_{11 \cdot 2}^{-1} \left( \hat{\mathbf{Q}}_{1y} - \hat{\mathbf{Q}}_{12} \hat{\mathbf{Q}}_{22}^{-1} \hat{\mathbf{Q}}_{21} \hat{\mathbf{Q}}_{11}^{-1} \hat{\mathbf{Q}}_{1y} \right) \\ &= \hat{\mathbf{Q}}_{11 \cdot 2}^{-1} \left( \hat{\mathbf{Q}}_{11} - \hat{\mathbf{Q}}_{12} \hat{\mathbf{Q}}_{22}^{-1} \hat{\mathbf{Q}}_{21} \right) \hat{\mathbf{Q}}_{11}^{-1} \hat{\mathbf{Q}}_{1y} \\ &= \hat{\mathbf{Q}}_{11}^{-1} \hat{\mathbf{Q}}_{1y} \end{aligned}$$

which is (8.13) as originally claimed.

## 8.4 Finite Sample Properties

In this section we explore some of the properties of the CLS estimator in the linear regression model

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + e_i \quad (8.14)$$

$$\mathbb{E}(e_i | \mathbf{x}_i) = 0. \quad (8.15)$$

First, it is useful to write the estimator and the residuals as linear functions of the error vector. These are algebraic relationships and do not rely on the linear regression assumptions.

**Theorem 8.1** Define  $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  and

$$\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\left(\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\right)^{-1}\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}.$$

Then

1.  $\mathbf{R}'\hat{\boldsymbol{\beta}} - \mathbf{c} = \mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}$
2.  $\tilde{\boldsymbol{\beta}}_{\text{cls}} - \boldsymbol{\beta} = \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{A}\mathbf{X}'\right)\mathbf{e}$
3.  $\tilde{\mathbf{e}} = (\mathbf{I} - \mathbf{P} + \mathbf{X}\mathbf{A}\mathbf{X}')\mathbf{e}$
4.  $\mathbf{I}_n - \mathbf{P} + \mathbf{X}\mathbf{A}\mathbf{X}'$  is symmetric and idempotent
5.  $\text{tr}(\mathbf{I}_n - \mathbf{P} + \mathbf{X}\mathbf{A}\mathbf{X}') = n - k + q.$

For a proof, see Exercise 8.6.

Given the linearity of Theorem 8.1.2, it is not hard to show that the CLS estimator is unbiased for  $\boldsymbol{\beta}$ .

**Theorem 8.2** In the linear regression model (8.14)-(8.15) under (8.1),  $\mathbb{E}(\tilde{\boldsymbol{\beta}}_{\text{cls}} | \mathbf{X}) = \boldsymbol{\beta}$ .

For a proof, see Exercise 8.7.

Given the linearity we can also calculate the variance matrix of  $\tilde{\boldsymbol{\beta}}_{\text{cls}}$ . For this we will add the assumption of conditional homoskedasticity to simplify the expression.

**Theorem 8.3** In the homoskedastic linear regression model (8.14)-(8.15) with  $\mathbb{E}(e_i^2 | \mathbf{x}_i) = \sigma^2$ , under (8.1),

$$\begin{aligned} V_{\tilde{\boldsymbol{\beta}}}^0 &= \text{var}(\tilde{\boldsymbol{\beta}}_{\text{cls}} | \mathbf{X}) \\ &= \left( (\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\left(\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\right)^{-1}\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1} \right) \sigma^2. \end{aligned}$$

For a proof, see Exercise 8.8.

We use the  $V_{\tilde{\boldsymbol{\beta}}}^0$  notation to emphasize that this is the variance matrix under the assumption of conditional homoskedasticity.

For inference we need an estimate of  $V_{\tilde{\boldsymbol{\beta}}}^0$ . A natural estimator is

$$\hat{V}_{\tilde{\boldsymbol{\beta}}}^0 = \left( (\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\left(\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\right)^{-1}\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1} \right) s_{\text{cls}}^2$$

where

$$s_{\text{cls}}^2 = \frac{1}{n - k + q} \sum_{i=1}^n \tilde{e}_i^2 \quad (8.16)$$

is a biased-corrected estimator of  $\sigma^2$ . Standard errors for the components of  $\boldsymbol{\beta}$  are then found by taking the squares roots of the diagonal elements of  $\widehat{V}_{\tilde{\boldsymbol{\beta}}}$ , for example

$$s(\widehat{\beta}_j) = \sqrt{\left[ \widehat{V}_{\tilde{\boldsymbol{\beta}}}^0 \right]_{jj}}.$$

The estimator (8.16) has the property that it is unbiased for  $\sigma^2$  under conditional homoskedasticity. To see this, using the properties of Theorem 8.1,

$$\begin{aligned} (n - k + q) s_{\text{cls}}^2 &= \tilde{\mathbf{e}}' \tilde{\mathbf{e}} \\ &= \mathbf{e}' (\mathbf{I}_n - \mathbf{P} + \mathbf{X}\mathbf{A}\mathbf{X}') (\mathbf{I}_n - \mathbf{P} + \mathbf{X}\mathbf{A}\mathbf{X}') \mathbf{e} \\ &= \mathbf{e}' (\mathbf{I}_n - \mathbf{P} + \mathbf{X}\mathbf{A}\mathbf{X}') \mathbf{e}. \end{aligned} \quad (8.17)$$

We defer the remainder of the proof to Exercise 8.9.

**Theorem 8.4** In the homoskedastic linear regression model (8.14)-(8.15) with  $\mathbb{E}(e_i^2 | \mathbf{x}_i) = \sigma^2$ , under (8.1),  $\mathbb{E}(s_{\text{cls}}^2 | \mathbf{X}) = \sigma^2$  and  $\mathbb{E}(\widehat{V}_{\tilde{\boldsymbol{\beta}}}^0 | \mathbf{X}) = V_{\tilde{\boldsymbol{\beta}}}^0$ .

Now consider the distributional properties in the normal regression model

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ e_i &\sim N(0, \sigma^2). \end{aligned}$$

By the linearity of Theorem 8.1.2, conditional on  $\mathbf{X}$ ,  $\tilde{\boldsymbol{\beta}}_{\text{cls}} - \boldsymbol{\beta}$  is normal. Given Theorems 8.2 and 8.3, we deduce that  $\tilde{\boldsymbol{\beta}}_{\text{cls}} \sim N(\boldsymbol{\beta}, V_{\tilde{\boldsymbol{\beta}}}^0)$ .

Similarly, from Exercise 8.1 we know  $\tilde{\mathbf{e}} = (\mathbf{I}_n - \mathbf{P} + \mathbf{X}\mathbf{A}\mathbf{X}') \mathbf{e}$  is linear in  $\mathbf{e}$  so is also conditionally normal. Furthermore, since  $(\mathbf{I}_n - \mathbf{P} + \mathbf{X}\mathbf{A}\mathbf{X}') (\mathbf{X}(\mathbf{X}')^{-1} - \mathbf{X}\mathbf{A}) = 0$ ,  $\tilde{\mathbf{e}}$  and  $\tilde{\boldsymbol{\beta}}_{\text{cls}}$  are uncorrelated and thus independent. Thus  $s_{\text{cls}}^2$  and  $\tilde{\boldsymbol{\beta}}_{\text{cls}}$  are independent.

From (8.17) and the fact that  $\mathbf{I}_n - \mathbf{P} + \mathbf{X}\mathbf{A}\mathbf{X}'$  is idempotent with rank  $n - k + q$ , it follows that

$$s_{\text{cls}}^2 \sim \sigma^2 \chi_{n-k+q}^2 / (n - k + q).$$

It follows that the t-statistic has the exact distribution

$$\begin{aligned} T &= \frac{\widehat{\beta}_j - \beta_j}{s(\widehat{\beta}_j)} \\ &\sim \frac{N(0, 1)}{\sqrt{\chi_{n-k+q}^2 / (n - k + q)}} \\ &\sim t_{n-k+q} \end{aligned}$$

a student  $t$  distribution with  $n - k + q$  degrees of freedom.

The relevance of this calculation is that the “degrees of freedom” for a CLS regression problem equal  $n - k + q$  rather than  $n - k$  as in the OLS regression problem. Essentially, the model has  $k - q$  free parameters instead of  $k$ . Another way of thinking about this is that estimation of a model with  $k$  coefficients and  $q$  restrictions is equivalent to estimation with  $k - q$  coefficients.

We summarize the properties of the normal regression model.

**Theorem 8.5** In the normal linear regression model linear regression model (8.14)-(8.15), under (8.1),

$$\begin{aligned}\tilde{\beta}_{\text{cls}} &\sim N(\beta, V_{\tilde{\beta}}^0) \\ \frac{(n-k+q)s_{\text{cls}}^2}{\sigma^2} &\sim \chi_{n-k+q}^2 \\ T &\sim t_{n-k+q}\end{aligned}$$

An interesting relationship is that in the homoskedastic regression model

$$\begin{aligned}(\hat{\beta}_{\text{ols}} - \tilde{\beta}_{\text{cls}}, \tilde{\beta}_{\text{cls}}) &= \mathbb{E}((\hat{\beta}_{\text{ols}} - \tilde{\beta}_{\text{cls}})(\tilde{\beta}_{\text{cls}} - \beta)') \\ &= \mathbb{E}((AX')(X'(X')^{-1} - XA))\sigma^2 = 0\end{aligned}$$

so  $\hat{\beta}_{\text{ols}} - \tilde{\beta}_{\text{cls}}$  and  $\tilde{\beta}_{\text{cls}}$  are uncorrelated and hence independent. One corollary is

$$\text{cov}(\hat{\beta}_{\text{ols}}, \tilde{\beta}_{\text{cls}}) = \text{var}(\tilde{\beta}_{\text{cls}}).$$

A second corollary is

$$\begin{aligned}\text{var}(\hat{\beta}_{\text{ols}} - \tilde{\beta}_{\text{cls}}) &= \text{var}(\hat{\beta}_{\text{ols}}) - \text{var}(\tilde{\beta}_{\text{cls}}) \\ &= (X'X)^{-1} R (R'(X'X)^{-1} R)^{-1} R'(X'X)^{-1} \sigma^2.\end{aligned}\tag{8.18}$$

This also shows us the difference between the CLS and OLS variances

$$\text{var}(\hat{\beta}_{\text{ols}}) - \text{var}(\tilde{\beta}_{\text{cls}}) = (X'X)^{-1} R (R'(X'X)^{-1} R)^{-1} R'(X'X)^{-1} \sigma^2 \geq 0$$

the final equality meaning positive semi-definite. It follows that  $\text{var}(\hat{\beta}_{\text{ols}}) \geq \text{var}(\tilde{\beta}_{\text{cls}})$  in the positive definite sense, and thus CLS is more efficient than OLS. Both estimators are unbiased (in the linear regression model), and CLS has a lower variance matrix (in the linear homoskedastic regression model).

The relationship (8.18) is rather interesting and will appear again. The expression says that the variance of the difference between the estimators is equal to the difference between the variances. This is rather special. It occurs (generically) when we are comparing an efficient and an inefficient estimator. We call (8.18) the **Hausmann Equality** as it was first pointed out in econometrics by Hausman (1978).

## 8.5 Minimum Distance

The previous section explored the finite sample distribution theory under the assumptions of the linear regression model, homoskedastic regression model, and normal regression model. We now return to the general projection model where we do not impose linearity, homoskedasticity, nor normality. We are interested in the question: Can we do better than CLS in this setting?

A minimum distance estimator tries to find a parameter value which satisfies the constraint which is as close as possible to the unconstrained estimate. Let  $\hat{\beta}$  be the unconstrained least-squares estimator, and for some  $k \times k$  positive definite weight matrix  $\widehat{W} > 0$  define the quadratic criterion function

$$J(\beta) = n(\hat{\beta} - \beta)' \widehat{W} (\hat{\beta} - \beta).\tag{8.19}$$

This is a (squared) weighted Euclidean distance between  $\hat{\beta}$  and  $\beta$ .  $J(\beta)$  is small if  $\beta$  is close to  $\hat{\beta}$ , and is minimized at zero only if  $\beta = \hat{\beta}$ . A **minimum distance estimator**  $\tilde{\beta}_{\text{md}}$  for  $\beta$  minimizes  $J(\beta)$  subject to the constraint (8.1), that is,

$$\tilde{\beta}_{\text{md}} = \underset{R'\beta=c}{\operatorname{argmin}} J(\beta).$$

The CLS estimator is the special case when  $\widehat{\mathbf{W}} = \widehat{\mathbf{Q}}_{xx}$ , and we write this criterion function as

$$J^0(\boldsymbol{\beta}) = n(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \widehat{\mathbf{Q}}_{xx} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \quad (8.20)$$

To see the equality of CLS and minimum distance, rewrite the least-squares criterion as follows. Substitute the unconstrained least-squares fitted equation  $y_i = \mathbf{x}'_i \widehat{\boldsymbol{\beta}} + \widehat{e}_i$  into SSE( $\boldsymbol{\beta}$ ) to obtain

$$\begin{aligned} \text{SSE}(\boldsymbol{\beta}) &= \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 \\ &= \sum_{i=1}^n (\mathbf{x}'_i \widehat{\boldsymbol{\beta}} + \widehat{e}_i - \mathbf{x}'_i \boldsymbol{\beta})^2 \\ &= \sum_{i=1}^n \widehat{e}_i^2 + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \right) (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &= n\widehat{\sigma}^2 + J^0(\boldsymbol{\beta}) \end{aligned} \quad (8.21)$$

where the third equality uses the fact that  $\sum_{i=1}^n \mathbf{x}_i \widehat{e}_i = 0$ , and the last line uses  $\sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i = n\widehat{\mathbf{Q}}_{xx}$ . The expression (8.21) only depends on  $\boldsymbol{\beta}$  through  $J^0(\boldsymbol{\beta})$ . Thus minimization of  $\text{SSE}(\boldsymbol{\beta})$  and  $J^0(\boldsymbol{\beta})$  are equivalent, and hence  $\tilde{\boldsymbol{\beta}}_{\text{md}} = \tilde{\boldsymbol{\beta}}_{\text{cls}}$  when  $\widehat{\mathbf{W}} = \widehat{\mathbf{Q}}_{xx}$ .

We can solve for  $\tilde{\boldsymbol{\beta}}_{\text{md}}$  explicitly by the method of Lagrange multipliers. The Lagrangian is

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2} J(\boldsymbol{\beta}, \widehat{\mathbf{W}}) + \boldsymbol{\lambda}' (\mathbf{R}' \boldsymbol{\beta} - \mathbf{c})$$

which is minimized over  $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ . The solution is

$$\tilde{\boldsymbol{\lambda}}_{\text{md}} = n \left( \mathbf{R}' \widehat{\mathbf{W}}^{-1} \mathbf{R} \right)^{-1} (\mathbf{R}' \widehat{\boldsymbol{\beta}} - \mathbf{c}) \quad (8.22)$$

$$\tilde{\boldsymbol{\beta}}_{\text{md}} = \widehat{\boldsymbol{\beta}} - \widehat{\mathbf{W}}^{-1} \mathbf{R} \left( \mathbf{R}' \widehat{\mathbf{W}}^{-1} \mathbf{R} \right)^{-1} (\mathbf{R}' \widehat{\boldsymbol{\beta}} - \mathbf{c}). \quad (8.23)$$

(See Exercise 8.10.) Comparing (8.23) with (8.9) we can see that  $\tilde{\boldsymbol{\beta}}_{\text{md}}$  specializes to  $\tilde{\boldsymbol{\beta}}_{\text{cls}}$  when we set  $\widehat{\mathbf{W}} = \widehat{\mathbf{Q}}_{xx}$ .

An obvious question is which weight matrix  $\widehat{\mathbf{W}}$  is best. We will address this question after we derive the asymptotic distribution for a general weight matrix.

## 8.6 Asymptotic Distribution

We first show that the class of minimum distance estimators are consistent for the population parameters when the constraints are valid.

**Assumption 8.1**  $\mathbf{R}' \boldsymbol{\beta} = \mathbf{c}$  where  $\mathbf{R}$  is  $k \times q$  with  $\text{rank}(\mathbf{R}) = q$ .

**Assumption 8.2**  $\widehat{\mathbf{W}} \xrightarrow{p} \mathbf{W} > 0$ .

**Theorem 8.6 Consistency**

Under Assumptions 7.1, 8.1, and 8.2,  $\tilde{\beta}_{\text{md}} \xrightarrow{p} \beta$  as  $n \rightarrow \infty$ .

For a proof, see Exercise 8.11.

Theorem 8.6 shows that consistency holds for any weight matrix with a positive definite limit, so the result includes the CLS estimator.

Similarly, the constrained estimators are asymptotically normally distributed.

**Theorem 8.7 Asymptotic Normality**

Under Assumptions 7.2, 8.1, and 8.2,

$$\sqrt{n}(\tilde{\beta}_{\text{md}} - \beta) \xrightarrow{d} N(\mathbf{0}, V_{\beta}(W))$$

as  $n \rightarrow \infty$ , where

$$\begin{aligned} V_{\beta}(W) = & V_{\beta} - W^{-1}R(R'W^{-1}R)^{-1}R'V_{\beta} \\ & - V_{\beta}R(R'W^{-1}R)^{-1}R'W^{-1} \\ & + W^{-1}R(R'W^{-1}R)^{-1}R'V_{\beta}R(R'W^{-1}R)^{-1}R'W^{-1} \end{aligned} \quad (8.24)$$

and  $V_{\beta} = Q_{xx}^{-1}\Omega Q_{xx}^{-1}$ .

For a proof, see Exercise 8.12.

Theorem 8.7 shows that the minimum distance estimator is asymptotically normal for all positive definite weight matrices. The asymptotic variance depends on  $W$ . The theorem includes the CLS estimator as a special case by setting  $W = Q_{xx}$ .

**Theorem 8.8 Asymptotic Distribution of CLS Estimator**

Under Assumptions 7.2 and 8.1, as  $n \rightarrow \infty$

$$\sqrt{n}(\tilde{\beta}_{\text{cls}} - \beta) \xrightarrow{d} N(\mathbf{0}, V_{\text{cls}})$$

where

$$\begin{aligned} V_{\text{cls}} = & V_{\beta} - Q_{xx}^{-1}R(R'Q_{xx}^{-1}R)^{-1}R'V_{\beta} \\ & - V_{\beta}R(R'Q_{xx}^{-1}R)^{-1}R'Q_{xx}^{-1} \\ & + Q_{xx}^{-1}R(R'Q_{xx}^{-1}R)^{-1}R'V_{\beta}R(R'Q_{xx}^{-1}R)^{-1}R'Q_{xx}^{-1}. \end{aligned}$$

For a proof, see Exercise 8.13.

## 8.7 Variance Estimation and Standard Errors

Earlier we introduced the covariance matrix estimator under the assumption of conditional homoskedasticity. We now introduce an estimator which does not impose homoskedasticity.

The asymptotic covariance matrix  $V_{\text{cls}}$  may be estimated by replacing  $V_{\beta}$  with a consistent estimator such as  $\widehat{V}_{\beta}$ . A more efficient estimator is obtained by using the restricted coefficient estimator. Given the constrained least-squares residuals  $\tilde{e}_i = y_i - \mathbf{x}'_i \tilde{\beta}_{\text{cls}}$  we can estimate the matrix  $\Omega = \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i e_i^2)$  by

$$\widetilde{\Omega} = \frac{1}{n - k + q} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \tilde{e}_i^2.$$

Notice that we have defined  $\widetilde{\Omega}$  using an adjusted degrees of freedom. This is an ad hoc adjustment designed to mimic that used for estimation of the error variance  $\sigma^2$ . Given  $\widehat{\Omega}$  the moment estimator of  $V_{\beta}$  is

$$\widetilde{V}_{\beta} = \widehat{\mathbf{Q}}_{xx}^{-1} \widetilde{\Omega} \widehat{\mathbf{Q}}_{xx}^{-1}$$

and that for  $V_{\text{cls}}$  is

$$\begin{aligned} \widetilde{V}_{\text{cls}} &= \widetilde{V}_{\beta} - \widehat{\mathbf{Q}}_{xx}^{-1} \mathbf{R} \left( \mathbf{R}' \widehat{\mathbf{Q}}_{xx}^{-1} \mathbf{R} \right)^{-1} \mathbf{R}' \widetilde{V}_{\beta} \\ &\quad - \widetilde{V}_{\beta} \mathbf{R} \left( \mathbf{R}' \widehat{\mathbf{Q}}_{xx}^{-1} \mathbf{R} \right)^{-1} \mathbf{R}' \widehat{\mathbf{Q}}_{xx}^{-1} \\ &\quad + \widehat{\mathbf{Q}}_{xx}^{-1} \mathbf{R} \left( \mathbf{R}' \widehat{\mathbf{Q}}_{xx}^{-1} \mathbf{R} \right)^{-1} \mathbf{R}' \widetilde{V}_{\beta} \mathbf{R} \left( \mathbf{R}' \widehat{\mathbf{Q}}_{xx}^{-1} \mathbf{R} \right)^{-1} \mathbf{R}' \widehat{\mathbf{Q}}_{xx}^{-1}. \end{aligned}$$

We can calculate standard errors for any linear combination  $\mathbf{h}' \widetilde{\beta}_{\text{cls}}$  so long as  $\mathbf{h}$  does not lie in the range space of  $\mathbf{R}$ . A standard error for  $\mathbf{h}' \widetilde{\beta}$  is

$$s(\mathbf{h}' \widetilde{\beta}_{\text{cls}}) = (n^{-1} \mathbf{h}' \widetilde{V}_{\text{cls}} \mathbf{h})^{1/2}.$$

## 8.8 Efficient Minimum Distance Estimator

Theorem 8.7 shows that minimum distance estimators, which include CLS as a special case, are asymptotically normal with an asymptotic covariance matrix which depends on the weight matrix  $\mathbf{W}$ . The asymptotically optimal weight matrix is the one which minimizes the asymptotic variance  $V_{\beta}(\mathbf{W})$ . This turns out to be  $\mathbf{W} = V_{\beta}^{-1}$  as is shown in Theorem 8.9 below. Since  $V_{\beta}^{-1}$  is unknown this weight matrix cannot be used for a feasible estimator, but we can replace  $V_{\beta}^{-1}$  with a consistent estimate  $\widehat{V}_{\beta}^{-1}$  and the asymptotic distribution (and efficiency) are unchanged. We call the minimum distance estimator setting  $\widehat{\mathbf{W}} = \widehat{V}_{\beta}^{-1}$  the **efficient minimum distance estimator** and takes the form

$$\widetilde{\beta}_{\text{emd}} = \widehat{\beta} - \widehat{V}_{\beta} \mathbf{R} (\mathbf{R}' \widehat{V}_{\beta} \mathbf{R})^{-1} (\mathbf{R}' \widehat{\beta} - \mathbf{c}). \quad (8.25)$$

The asymptotic distribution of (8.25) can be deduced from Theorem 8.7. (See Exercises 8.14 and 8.15, and the proof in Section 8.16.)

**Theorem 8.9 Efficient Minimum Distance Estimator**

Under Assumptions 7.2 and 8.1,

$$\sqrt{n}(\tilde{\beta}_{\text{emd}} - \beta) \xrightarrow{d} N(\mathbf{0}, V_{\beta, \text{emd}})$$

as  $n \rightarrow \infty$ , where

$$V_{\beta, \text{emd}} = V_{\beta} - V_{\beta} R (R' V_{\beta} R)^{-1} R' V_{\beta}. \quad (8.26)$$

Since

$$V_{\beta, \text{emd}} \leq V_{\beta} \quad (8.27)$$

the estimator (8.25) has lower asymptotic variance than the unrestricted estimator. Furthermore, for any  $W$ ,

$$V_{\beta, \text{emd}} \leq V_{\beta}(W) \quad (8.28)$$

so (8.25) is asymptotically efficient in the class of minimum distance estimators.

Theorem 8.9 shows that the minimum distance estimator with the smallest asymptotic variance is (8.25). One implication is that the constrained least squares estimator is generally inefficient. The interesting exception is the case of conditional homoskedasticity, in which case the optimal weight matrix is  $W = (V_{\beta}^0)^{-1}$  so in this case CLS is an efficient minimum distance estimator. Otherwise when the error is conditionally heteroskedastic, there are asymptotic efficiency gains by using minimum distance rather than least squares.

The fact that CLS is generally inefficient is counter-intuitive and requires some reflection to understand. Standard intuition suggests to apply the same estimation method (least squares) to the unconstrained and constrained models, and this is the most common empirical practice. But Theorem 8.9 shows that this is not the efficient estimation method. Instead, the efficient minimum distance estimator has a smaller asymptotic variance. Why? The reason is that the least-squares estimator does not make use of the regressor  $x_{2i}$ . It ignores the information  $\mathbb{E}(x_{2i} e_i) = \mathbf{0}$ . This information is relevant when the error is heteroskedastic and the excluded regressors are correlated with the included regressors.

Inequality (8.27) shows that the efficient minimum distance estimator  $\tilde{\beta}_{\text{emd}}$  has a smaller asymptotic variance than the unrestricted least squares estimator  $\hat{\beta}$ . This means that efficient estimation is attained by imposing correct restrictions when we use the minimum distance method.

## 8.9 Exclusion Restriction Revisited

We return to the example of estimation with a simple exclusion restriction. The model is

$$y_i = x'_{1i} \beta_1 + x'_{2i} \beta_2 + e_i$$

with the exclusion restriction  $\beta_2 = \mathbf{0}$ . We have introduced three estimators of  $\beta_1$ . The first is unconstrained least-squares applied to (8.10), which can be written as

$$\hat{\beta}_1 = \hat{Q}_{11 \cdot 2}^{-1} \hat{Q}_{1y \cdot 2}.$$

From Theorem 7.25 and equation (7.14) its asymptotic variance is

$$\text{avar}(\hat{\beta}_1) = Q_{11 \cdot 2}^{-1} (\Omega_{11} - Q_{12} Q_{22}^{-1} \Omega_{21} - \Omega_{12} Q_{22}^{-1} Q_{21} + Q_{12} Q_{22}^{-1} \Omega_{22} Q_{22}^{-1} Q_{21}) Q_{11 \cdot 2}^{-1}.$$

The second estimator of  $\beta_1$  is the CLS estimator, which can be written as

$$\tilde{\beta}_1 = \hat{Q}_{11}^{-1} \hat{Q}_{1y}.$$

Its asymptotic variance can be deduced from Theorem 8.8, but it is simpler to apply the CLT directly to show that

$$\text{avar}(\tilde{\beta}_1) = Q_{11}^{-1} \Omega_{11} Q_{11}^{-1}. \quad (8.29)$$

The third estimator of  $\beta_1$  is the efficient minimum distance estimator. Applying (8.25), it equals

$$\bar{\beta}_1 = \hat{\beta}_1 - \hat{V}_{12} \hat{V}_{22}^{-1} \hat{\beta}_2 \quad (8.30)$$

where we have partitioned

$$\hat{V}_\beta = \begin{bmatrix} \hat{V}_{11} & \hat{V}_{12} \\ \hat{V}_{21} & \hat{V}_{22} \end{bmatrix}.$$

From Theorem 8.9 its asymptotic variance is

$$\text{avar}(\bar{\beta}_1) = V_{11} - V_{12} V_{22}^{-1} V_{21}. \quad (8.31)$$

See Exercise 8.16 to verify equations (8.29), (8.30), and (8.31).

In general, the three estimators are different, and they have different asymptotic variances. It is instructive to compare the variances to assess whether or not the constrained estimator is necessarily more efficient than the unconstrained estimator.

First, consider the case of conditional homoskedasticity. In this case the two covariance matrices simplify to

$$\text{avar}(\hat{\beta}_1) = \sigma^2 Q_{11,2}^{-1}$$

and

$$\text{avar}(\tilde{\beta}_1) = \sigma^2 Q_{11}^{-1}.$$

If  $Q_{12} = 0$  (so  $x_{1i}$  and  $x_{2i}$  are orthogonal) then these two variance matrices are equal and the two estimators have equal asymptotic efficiency. Otherwise, since  $Q_{12} Q_{22}^{-1} Q_{21} \geq 0$ , then  $Q_{11} \geq Q_{11} - Q_{12} Q_{22}^{-1} Q_{21}$ , and consequently

$$Q_{11}^{-1} \sigma^2 \leq (Q_{11} - Q_{12} Q_{22}^{-1} Q_{21})^{-1} \sigma^2.$$

This means that under conditional homoskedasticity,  $\tilde{\beta}_1$  has a lower asymptotic variance matrix than  $\hat{\beta}_1$ . Therefore in this context, constrained least-squares is more efficient than unconstrained least-squares. This is consistent with our intuition that imposing a correct restriction (excluding an irrelevant regressor) improves estimation efficiency.

However, in the general case of conditional heteroskedasticity this ranking is not guaranteed. In fact what is really amazing is that the variance ranking can be reversed. The CLS estimator can have a larger asymptotic variance than the unconstrained least squares estimator.

To see this let's use the simple heteroskedastic example from Section 7.4. In that example,  $Q_{11} = Q_{22} = 1$ ,  $Q_{12} = \frac{1}{2}$ ,  $\Omega_{11} = \Omega_{22} = 1$ , and  $\Omega_{12} = \frac{7}{8}$ . We can calculate (see Exercise 8.17) that  $Q_{11,2} = \frac{3}{4}$  and

$$\text{avar}(\hat{\beta}_1) = \frac{2}{3} \quad (8.32)$$

$$\text{avar}(\tilde{\beta}_1) = 1 \quad (8.33)$$

$$\text{avar}(\bar{\beta}_1) = \frac{5}{8}. \quad (8.34)$$

Thus the restricted least-squares estimator  $\tilde{\beta}_1$  has a larger variance than the unrestricted least-squares estimator  $\hat{\beta}_1$ ! The minimum distance estimator has the smallest variance of the three, as expected.

What we have found is that when the estimation method is least-squares, deleting the irrelevant variable  $x_{2i}$  can actually increase estimation variance, or equivalently, adding an irrelevant variable can decrease the estimation variance.

To repeat this unexpected finding, we have shown in a very simple example that it is possible for least-squares applied to the short regression (8.11) to be less efficient for estimation of  $\beta_1$  than least-squares applied to the long regression (8.10), even though the constraint  $\beta_2 = 0$  is valid! This result is strongly counter-intuitive. It seems to contradict our initial motivation for pursuing constrained estimation – to improve estimation efficiency.

It turns out that a more refined answer is appropriate. Constrained estimation is desirable, but not constrained least-squares estimation. While least-squares is asymptotically efficient for estimation of the unconstrained projection model, it is not an efficient estimator of the constrained projection model.

## 8.10 Variance and Standard Error Estimation

We have discussed covariance matrix estimation for the CLS estimator, but not yet for the EMD estimator.

The asymptotic covariance matrix (8.26) may be estimated by replacing  $V_\beta$  with a consistent estimate. It is best to construct the variance estimate using  $\tilde{\beta}_{\text{emd}}$ . The EMD residuals are  $\tilde{e}_i = y_i - \mathbf{x}'_i \tilde{\beta}_{\text{emd}}$ . Using these we can estimate the matrix  $\Omega = \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i e_i^2)$  by

$$\tilde{\Omega} = \frac{1}{n - k + q} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \tilde{e}_i^2.$$

Following the formula for CLS we recommend an adjusted degrees of freedom. Given  $\tilde{\Omega}$  the moment estimator of  $V_\beta$  is

$$\tilde{V}_\beta = \hat{Q}_{xx}^{-1} \tilde{\Omega} \hat{Q}_{xx}^{-1}$$

Given this, we construct the variance estimator

$$\tilde{V}_{\beta, \text{emd}} = \tilde{V}_\beta - \tilde{V}_\beta \mathbf{R} (\mathbf{R}' \tilde{V}_\beta \mathbf{R})^{-1} \mathbf{R}' \tilde{V}_\beta. \quad (8.35)$$

A standard error for  $\mathbf{h}' \tilde{\beta}$  is then

$$s(\mathbf{h}' \tilde{\beta}) = (n^{-1} \mathbf{h}' \tilde{V}_{\beta, \text{emd}} \mathbf{h})^{1/2}. \quad (8.36)$$

## 8.11 Hausman Equality

From (8.25) we have

$$\begin{aligned} \sqrt{n}(\hat{\beta}_{\text{ols}} - \tilde{\beta}_{\text{emd}}) &= \hat{V}_\beta \mathbf{R} (\mathbf{R}' \hat{V}_\beta \mathbf{R})^{-1} \sqrt{n}(\mathbf{R}' \hat{\beta}_{\text{ols}} - \mathbf{c}) \\ &\xrightarrow{d} \mathcal{N}\left(\mathbf{0}, V_\beta \mathbf{R} (\mathbf{R}' V_\beta \mathbf{R})^{-1} \mathbf{R}' V_\beta\right). \end{aligned}$$

It follows that the asymptotic variances of the estimators satisfy the relationship

$$\text{avar}(\hat{\beta}_{\text{ols}} - \tilde{\beta}_{\text{emd}}) = \text{avar}(\hat{\beta}_{\text{ols}}) - \text{avar}(\tilde{\beta}_{\text{emd}}). \quad (8.37)$$

We call (8.37) the Hausman Equality: the asymptotic variance of the difference between an efficient and inefficient estimator is the difference in the asymptotic variances.

## 8.12 Example: Mankiw, Romer and Weil (1992)

We illustrate the methods by replicating some of the estimates reported in a well-known paper by Mankiw, Romer, and Weil (1992). The paper investigates the implications of the Solow growth model using cross-country regressions. A key equation in their paper regresses the change between 1960 and 1985 in log GDP per capita on (1) log GDP in 1960, (2) the log of the ratio of aggregate investment to GDP, (3) the log of the sum of the population growth rate  $n$ , the technological growth rate  $g$ , and the rate of depreciation  $\delta$ , and (4) the log of the percentage of the working-age population that is in secondary school ( $School$ ), the latter a proxy for human-capital accumulation.

Table 8.1: Estimates of Solow Growth Model

	$\hat{\beta}_{ols}$	$\hat{\beta}_{cls}$	$\hat{\beta}_{emd}$
$\log GDP_{1960}$	-0.29 (0.05)	-0.30 (0.05)	-0.30 (0.05)
$\log \frac{I}{GDP}$	0.52 (0.11)	0.50 (0.09)	0.46 (0.08)
$\log(n + g + \delta)$	-0.51 (0.24)	-0.74 (0.08)	-0.71 (0.07)
$\log School$	0.23 (0.07)	0.24 (0.07)	0.25 (0.06)
Intercept	3.02 (0.74)	2.46 (0.44)	2.48 (0.44)

Standard errors are heteroskedasticity-consistent

The data is available on the textbook webpage in the file MRW1992.

The sample is 98 non-oil-producing countries, and the data was reported in the published paper. As  $g$  and  $\delta$  were unknown the authors set  $g + \delta = 0.05$ . We report least-squares estimates in the first column of Table 8.1. The estimates are consistent with the Solow theory due to the positive coefficients on investment and human capital and negative coefficient for population growth. The estimates are also consistent with the convergence hypothesis (that income levels tend towards a common mean over time) as the coefficient on initial GDP is negative.

The authors show that in the Solow model the 2<sup>nd</sup>, 3<sup>rd</sup> and 4<sup>th</sup> coefficients sum to zero. They reestimated the equation imposing this constraint. We present constrained least-squares estimates in the second column of Table 8.1, and efficient minimum distance estimates in the third column. Most of the coefficients and standard errors only exhibit small changes by imposing the constraint. The one exception is the coefficient on log population growth, which increases in magnitude and its standard error decreases substantially. The differences between the CLS and EMD estimates are modest.

We now present Stata, R and MATLAB code which implements these estimates.

You may notice that the Stata code has a section which uses the Mata matrix programming language. This is used because Stata does not implement the efficient minimum distance estimator, so needs to be separately programmed. As illustrated here, the Mata language allows a Stata user to implement methods using commands which are quite similar to MATLAB.

**Stata do File**

```

use "MRW1992.dta", clear
gen lndY = log(Y85)-log(Y60)
gen lnY60 = log(Y60)
gen lnI = log(invest/100)
gen lnG = log(pop_growth/100+0.05)
gen lnS = log(school/100)
// Unrestricted regression
reg lndY lnY60 lnI lnG lnS if N==1, r
// Store result for efficient minimum distance
mat b = e(b)'
scalar k = e(rank)
mat V = e(V)
// Constrained regression
constraint define 1 lnI+lnG+lnS=0
cnsreg lndY lnY60 lnI lnG lnS if N==1, constraints(1) r
// Efficient minimum distance
mata{
    data = st_data(.,("lnY60","lnI","lnG","lnS","lndY","N"))
    data_select = select(data,data[.,6]==1)
    y = data_select[.,5]
    n = rows(y)
    x = (data_select[.,1..4],J(n,1,1))
    k = cols(x)
    invx = invsym(x'*x)
    b_ols = st_matrix("b")
    V_ols = st_matrix("V")
    R = (0 \ 1 \ 1 \ 1 \ 0)
    b_emd = b_ols-V_ols*R*invsym(R'*V_ols*R)*R'*b_ols
    e_emd = J(1,k,y-x*b_emd)
    xe_emd = x:*e_emd
    xe_emd'*xe_emd
    V2 = (n/(n-k+1))*invx*(xe_emd'*xe_emd)*invx
    V_emd = V2 - V2*R*invsym(R'*V2*R)*R'*V2
    se_emd = diagonal(sqrt(V_emd))
    st_matrix("b_emd",b_emd)
    st_matrix("se_emd",se_emd)}
mat list b_emd
mat list se_emd

```

**R Program File**

```

data <- read.table("MRW1992.txt",header=TRUE)
N <- matrix(data$N,ncol=1)
lnDY <- matrix(log(data$Y85)-log(data$Y60),ncol=1)
lnY60 <- matrix(log(data$Y60),ncol=1)
lnI <- matrix(log(data$invest/100),ncol=1)
lnG <- matrix(log(data$pop_growth/100+0.05),ncol=1)
lnS <- matrix(log(data$school/100),ncol=1)
xx <- as.matrix(cbind(lnY60,lnI,lnG,lnS,matrix(1,nrow(lnDY),1)))
x <- xx[N==1,]
y <- lnDY[N==1]
n <- nrow(x)
k <- ncol(x)
# Unrestricted regression
invx <- solve(t(x)%*%x)
b_ols <- solve((t(x)%*%x),(t(x)%*%y))
e_ols <- rep((y-x%*%beta_ols),times=k)
xe_ols <- x*e_ols
V_ols <- (n/(n-k))*invx%*%(t(xe_ols)%*%xe_ols)%*%invx
se_ols <- sqrt(diag(V_ols))
print(beta_ols)
print(se_ols)
# Constrained regression
R <- c(0,1,1,1,0)
iR <- invx%*%R%*%solve(t(R)%*%invx%*%R)%*%t(R)
b_cls <- b_ols - iR%*%b_ols
e_cls <- rep((y-x%*%b_cls),times=k)
xe_cls <- x*e_cls
V_tilde <- (n/(n-k+1))*invx%*%(t(xe_cls)%*%xe_cls)%*%invx
V_cls <- V_tilde - iR%*%V_tilde - V_tilde%*%t(iR) +iR%*%V_tilde%*%t(iR)
print(b_cls)print(se_cls)
# Efficient minimum distance
Vr <- V_ols%*%R%*%solve(t(R)%*%V_ols%*%R)%*%t(R)
b_emd <- b_ols - Vr%*%b_ols
e_emd <- rep((y-x%*%b_emd),times=k)
xe_emd <- x*e_emd
V2 <- (n/(n-k+1))*invx%*%(t(xe_emd)%*%xe_emd)%*%invx
V_emd <- V2 - V2%*%R%*%solve(t(R)%*%V2%*%R)%*%t(R)%*%V2
se_emd <- sqrt(diag(V_emd))

```

**MATLAB Program File**

```

data = xlsread('MRW1992.xlsx');
N = data(:,1);
Y60 = data(:,4);
Y85 = data(:,5);
pop_growth = data(:,7);
invest = data(:,8);
school = data(:,9);
lnDY = log(Y85)-log(Y60);
lnY60 = log(Y60);
lnI = log(invest/100);
lnG = log(pop_growth/100+0.05);
lnS = log(school/100);
xx = [lnY60,lnI,lnG,lnS,ones(size(lnDy,1),1)];
x = xx(N==1,:);
y = lnDy(N==1);
[n,k] = size(x);
% Unrestricted regression
invx = inv(x'*x);
beta_ols = (x'*x)\(x'*y);
e_ols = repmat((y-x*beta_ols),1,k);
xe_ols = x.*e_ols;
V_ols = (n/(n-k))*invx*(xe_ols'*xe_ols)*invx;
se_ols = sqrt(diag(V_ols));
display(beta_ols);
display(se_ols);
% Constrained regression
R = [0;1;1;1;0];
iR = invx*R*inv(R'*invx*R)*R';
beta_cls = beta_ols - iR*beta_ols;
e_cls = repmat((y-x*beta_cls),1,k);
xe_cls = x.*e_cls;
V_tilde = (n/(n-k+1))*invx*(xe_cls'*xe_cls)*invx;
V_cls = V_tilde - iR*V_tilde - V_tilde*(iR')...
+ iR*V_tilde*(iR');
se_cls = sqrt(diag(V_cls));
display(beta_cls);
display(se_cls);
% Efficient minimum distance
beta_emd = beta_ols-V_ols*R*inv(R'*V_ols*R)*R'*beta_ols;
e_emd = repmat((y-x*beta_emd),1,k);
xe_emd = x.*e_emd;
V2 = (n/(n-k+1))*invx*(xe_emd'*xe_emd)*invx;
V_emd = V2 - V2*R*inv(R'*V2*R)*R'*V2;
se_emd = sqrt(diag(V_emd));
display(beta_emd);
display(se_emd);

```

## 8.13 Misspecification

What are the consequences for a constrained estimator  $\tilde{\beta}$  if the constraint (8.1) is incorrect? To be specific, suppose that the truth is

$$\mathbf{R}'\boldsymbol{\beta} = \mathbf{c}^*$$

where  $\mathbf{c}^*$  is not necessarily equal to  $\mathbf{c}$ .

This situation is a generalization of the analysis of “omitted variable bias” from Section 2.24, where we found that the short regression (e.g. (8.12)) is estimating a different projection coefficient than the long regression (e.g. (8.10)).

One mechanical answer is that we can use the formula (8.23) for the minimum distance estimator to find that

$$\tilde{\beta}_{\text{md}} \xrightarrow{P} \boldsymbol{\beta}_{\text{md}}^* = \boldsymbol{\beta} - \mathbf{W}^{-1}\mathbf{R}(\mathbf{R}'\mathbf{W}^{-1}\mathbf{R})^{-1}(\mathbf{c}^* - \mathbf{c}). \quad (8.38)$$

The second term,  $\mathbf{W}^{-1}\mathbf{R}(\mathbf{R}'\mathbf{W}^{-1}\mathbf{R})^{-1}(\mathbf{c}^* - \mathbf{c})$ , shows that imposing an incorrect constraint leads to inconsistency – an asymptotic bias. We can call the limiting value  $\boldsymbol{\beta}_{\text{md}}^*$  the minimum-distance projection coefficient or the pseudo-true value implied by the restriction.

However, we can say more.

For example, we can describe some characteristics of the approximating projections. The CLS estimator projection coefficient has the representation

$$\boldsymbol{\beta}_{\text{cls}}^* = \underset{\mathbf{R}'\boldsymbol{\beta}=\mathbf{c}}{\operatorname{argmin}} \mathbb{E}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2,$$

the best linear predictor subject to the constraint (8.1). The minimum distance estimator converges in probability to

$$\boldsymbol{\beta}_{\text{md}}^* = \underset{\mathbf{R}'\boldsymbol{\beta}=\mathbf{c}}{\operatorname{argmin}} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' \mathbf{W} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)$$

where  $\boldsymbol{\beta}_0$  is the true coefficient. That is,  $\boldsymbol{\beta}_{\text{md}}^*$  is the coefficient vector satisfying (8.1) closest to the true value in the weighted Euclidean norm. These calculations show that the constrained estimators are still reasonable in the sense that they produce good approximations to the true coefficient, conditional on being required to satisfy the constraint.

We can also show that  $\tilde{\beta}_{\text{md}}$  has an asymptotic normal distribution. The trick is to define the pseudo-true value

$$\boldsymbol{\beta}_n^* = \boldsymbol{\beta} - \widehat{\mathbf{W}}^{-1}\mathbf{R}(\mathbf{R}'\widehat{\mathbf{W}}^{-1}\mathbf{R})^{-1}(\mathbf{c}^* - \mathbf{c}). \quad (8.39)$$

(Note that (8.38) and (8.39) are different!) Then

$$\begin{aligned} \sqrt{n}(\tilde{\beta}_{\text{md}} - \boldsymbol{\beta}_n^*) &= \sqrt{n}(\tilde{\beta} - \boldsymbol{\beta}) - \widehat{\mathbf{W}}^{-1}\mathbf{R}(\mathbf{R}'\widehat{\mathbf{W}}^{-1}\mathbf{R})^{-1}\sqrt{n}(\mathbf{R}'\tilde{\beta} - \mathbf{c}^*) \\ &= \left( \mathbf{I} - \widehat{\mathbf{W}}^{-1}\mathbf{R}(\mathbf{R}'\widehat{\mathbf{W}}^{-1}\mathbf{R})^{-1}\mathbf{R}' \right) \sqrt{n}(\tilde{\beta} - \boldsymbol{\beta}) \\ &\xrightarrow{d} \left( \mathbf{I} - \mathbf{W}^{-1}\mathbf{R}(\mathbf{R}'\mathbf{W}^{-1}\mathbf{R})^{-1}\mathbf{R}' \right) \mathcal{N}(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}}) \\ &= \mathcal{N}(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}}(\mathbf{W})). \end{aligned} \quad (8.40)$$

In particular

$$\sqrt{n}(\tilde{\beta}_{\text{md}} - \boldsymbol{\beta}_n^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}}^*).$$

This means that even when the constraint (8.1) is misspecified, the conventional covariance matrix estimator (8.35) and standard errors (8.36) are appropriate measures of the sampling variance, though the distributions are centered at the pseudo-true values (projections)  $\boldsymbol{\beta}_n^*$  rather than  $\boldsymbol{\beta}$ . The fact that the estimators are biased is an unavoidable consequence of misspecification.

An alternative approach to the asymptotic distribution theory under misspecification uses the concept of local alternatives. It is a technical device which might seem a bit artificial, but it is a powerful

method to derive useful distributional approximations in a wide variety of contexts. The idea is to index the true coefficient  $\beta_n$  by  $n$  via the relationship

$$\mathbf{R}'\boldsymbol{\beta}_n = \mathbf{c} + \boldsymbol{\delta} n^{-1/2}. \quad (8.41)$$

Equation (8.41) specifies that  $\boldsymbol{\beta}_n$  violates (8.1) and thus the constraint is misspecified. However, the constraint is “close” to correct, as the difference  $\mathbf{R}'\boldsymbol{\beta}_n - \mathbf{c} = \boldsymbol{\delta} n^{-1/2}$  is “small” in the sense that it decreases with the sample size  $n$ . We call (8.41) **local misspecification**.

The asymptotic theory is then derived as  $n \rightarrow \infty$  under the sequence of probability distributions with the coefficients  $\boldsymbol{\beta}_n$ . The way to think about this is that the true value of the parameter is  $\boldsymbol{\beta}_n$ , and it is “close” to satisfying (8.1). The reason why the deviation is proportional to  $n^{-1/2}$  is because this is the only choice under which the localizing parameter  $\boldsymbol{\delta}$  appears in the asymptotic distribution but does not dominate it. The best way to see this is to work through the asymptotic approximation.

Since  $\boldsymbol{\beta}_n$  is the true coefficient value, then  $y_i = \mathbf{x}'_i \boldsymbol{\beta}_n + e_i$  and we have the standard representation for the unconstrained estimator, namely

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n) &= \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_i e_i \right) \\ &\xrightarrow{d} N(\mathbf{0}, V_{\boldsymbol{\beta}}). \end{aligned} \quad (8.42)$$

There is no difference under fixed (classical) or local asymptotics, since the right-hand-side is independent of the coefficient  $\boldsymbol{\beta}_n$ .

A difference arises for the constrained estimator. Using (8.41),  $\mathbf{c} = \mathbf{R}'\boldsymbol{\beta}_n - \boldsymbol{\delta} n^{-1/2}$ , so

$$\mathbf{R}'\hat{\boldsymbol{\beta}} - \mathbf{c} = \mathbf{R}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n) + \boldsymbol{\delta} n^{-1/2}$$

and

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_{\text{md}} &= \hat{\boldsymbol{\beta}} - \widehat{\mathbf{W}}^{-1} \mathbf{R} \left( \mathbf{R}' \widehat{\mathbf{W}}^{-1} \mathbf{R} \right)^{-1} (\mathbf{R}' \hat{\boldsymbol{\beta}} - \mathbf{c}) \\ &= \hat{\boldsymbol{\beta}} - \widehat{\mathbf{W}}^{-1} \mathbf{R} \left( \mathbf{R}' \widehat{\mathbf{W}}^{-1} \mathbf{R} \right)^{-1} \mathbf{R}' (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n) + \widehat{\mathbf{W}}^{-1} \mathbf{R} \left( \mathbf{R}' \widehat{\mathbf{W}}^{-1} \mathbf{R} \right)^{-1} \boldsymbol{\delta} n^{-1/2}. \end{aligned}$$

It follows that

$$\begin{aligned} \sqrt{n}(\tilde{\boldsymbol{\beta}}_{\text{md}} - \boldsymbol{\beta}_n) &= \left( \mathbf{I} - \widehat{\mathbf{W}}^{-1} \mathbf{R} \left( \mathbf{R}' \widehat{\mathbf{W}}^{-1} \mathbf{R} \right)^{-1} \mathbf{R}' \right) \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n) \\ &\quad + \widehat{\mathbf{W}}^{-1} \mathbf{R} \left( \mathbf{R}' \widehat{\mathbf{W}}^{-1} \mathbf{R} \right)^{-1} \boldsymbol{\delta}. \end{aligned}$$

The first term is asymptotically normal (from 8.42)). The second term converges in probability to a constant. This is because the  $n^{-1/2}$  local scaling in (8.41) is exactly balanced by the  $\sqrt{n}$  scaling of the estimator. No alternative rate would have produced this result.

Consequently, we find that the asymptotic distribution equals

$$\begin{aligned} \sqrt{n}(\tilde{\boldsymbol{\beta}}_{\text{md}} - \boldsymbol{\beta}_n) &\xrightarrow{d} N(\mathbf{0}, V_{\boldsymbol{\beta}}) + \mathbf{W}^{-1} \mathbf{R} \left( \mathbf{R}' \mathbf{W}^{-1} \mathbf{R} \right)^{-1} \boldsymbol{\delta} \\ &= N(\boldsymbol{\delta}^*, V_{\boldsymbol{\beta}}(\mathbf{W})) \end{aligned} \quad (8.43)$$

where

$$\boldsymbol{\delta}^* = \mathbf{W}^{-1} \mathbf{R} \left( \mathbf{R}' \mathbf{W}^{-1} \mathbf{R} \right)^{-1} \boldsymbol{\delta}.$$

The asymptotic distribution (8.43) is an approximation of the sampling distribution of the restricted estimator under misspecification. The distribution (8.43) contains an asymptotic bias component  $\boldsymbol{\delta}^*$ . The approximation is not fundamentally different from (8.40) – they both have the same asymptotic variances, and both reflect the bias due to misspecification. The difference is that (8.40) puts the bias on the left-side of the convergence arrow, while (8.43) has the bias on the right-side. There is no substantive difference between the two, but (8.43) is more convenient for some purposes, such as the analysis of the power of tests, as we will explore in the next chapter.

## 8.14 Nonlinear Constraints

In some cases it is desirable to impose nonlinear constraints on the parameter vector  $\beta$ . They can be written as

$$\mathbf{r}(\beta) = \mathbf{0} \quad (8.44)$$

where  $\mathbf{r} : \mathbb{R}^k \rightarrow \mathbb{R}^q$ . This includes the linear constraints (8.1) as a special case. An example of (8.44) which cannot be written as (8.1) is  $\beta_1\beta_2 = 1$ , which is (8.44) with  $r(\beta) = \beta_1\beta_2 - 1$ .

The constrained least-squares and minimum distance estimators of  $\beta$  subject to (8.44) solve the minimization problems

$$\tilde{\beta}_{\text{cls}} = \underset{\mathbf{r}(\beta)=\mathbf{0}}{\operatorname{argmin}} \text{SSE}(\beta) \quad (8.45)$$

$$\tilde{\beta}_{\text{md}} = \underset{\mathbf{r}(\beta)=\mathbf{0}}{\operatorname{argmin}} J(\beta) \quad (8.46)$$

where  $\text{SSE}(\beta)$  and  $J(\beta)$  are defined in (8.4) and (8.19), respectively. The solutions minimize the Lagrangians

$$\mathcal{L}(\beta, \lambda) = \frac{1}{2} \text{SSE}(\beta) + \lambda' \mathbf{r}(\beta)$$

or

$$\mathcal{L}(\beta, \lambda) = \frac{1}{2} J(\beta) + \lambda' \mathbf{r}(\beta) \quad (8.47)$$

over  $(\beta, \lambda)$ .

Computationally, there is no general closed-form solution for the estimator so they must be found numerically. Algorithms to numerically solve (8.45) and (8.46) are known as **constrained optimization** methods, and are available in programming languages including MATLAB, GAUSS and R.

**Assumption 8.3**  $\mathbf{r}(\beta) = \mathbf{0}$ ,  $\mathbf{r}(\beta)$  is continuously differentiable at the true  $\beta$ , and  $\text{rank}(\mathbf{R}) = q$ , where  $\mathbf{R} = \frac{\partial}{\partial \beta} \mathbf{r}(\beta)'$ .

The asymptotic distribution is a simple generalization of the case of a linear constraint, but the proof is more delicate.

**Theorem 8.10** Under Assumptions 7.2, 8.2, and 8.3, for  $\tilde{\beta} = \tilde{\beta}_{\text{md}}$  and  $\tilde{\beta} = \tilde{\beta}_{\text{cls}}$  defined in (8.45) and (8.46),

$$\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{d} N(\mathbf{0}, V_{\beta}(W))$$

as  $n \rightarrow \infty$ , where  $V_{\beta}(W)$  is defined in (8.24). For  $\tilde{\beta}_{\text{cls}}$ ,  $W = \mathbf{Q}_{xx}$  and  $V_{\beta}(W) = V_{\text{cls}}$  as defined in Theorem 8.8.  $V_{\beta}(W)$  is minimized with  $W = V_{\beta}^{-1}$ , in which case the asymptotic variance is

$$V_{\beta}^* = V_{\beta} - V_{\beta} \mathbf{R} (\mathbf{R}' V_{\beta} \mathbf{R})^{-1} \mathbf{R}' V_{\beta}.$$

The asymptotic variance matrix for the efficient minimum distance estimator can be estimated by

$$\hat{V}_{\beta}^* = \hat{V}_{\beta} - \hat{V}_{\beta} \hat{\mathbf{R}} (\hat{\mathbf{R}}' \hat{V}_{\beta} \hat{\mathbf{R}})^{-1} \hat{\mathbf{R}}' \hat{V}_{\beta}$$

where

$$\hat{\mathbf{R}} = \frac{\partial}{\partial \beta} \mathbf{r}(\tilde{\beta}_{\text{md}})'. \quad (8.48)$$

Standard errors for the elements of  $\tilde{\beta}_{\text{md}}$  are the square roots of the diagonal elements of  $\hat{V}_{\beta}^* = n^{-1} \hat{V}_{\beta}^*$ .

## 8.15 Inequality Restrictions

Inequality constraints on the parameter vector  $\beta$  take the form

$$\mathbf{r}(\beta) \geq \mathbf{0} \quad (8.49)$$

for some function  $\mathbf{r} : \mathbb{R}^k \rightarrow \mathbb{R}^q$ . The most common example is a non-negative constraint

$$\beta_1 \geq 0.$$

The constrained least-squares and minimum distance estimators can be written as

$$\tilde{\beta}_{\text{cls}} = \underset{\mathbf{r}(\beta) \geq \mathbf{0}}{\operatorname{argmin}} \text{SSE}(\beta) \quad (8.50)$$

and

$$\tilde{\beta}_{\text{md}} = \underset{\mathbf{r}(\beta) \geq \mathbf{0}}{\operatorname{argmin}} J(\beta). \quad (8.51)$$

Except in special cases the constrained estimators do not have simple algebraic solutions. An important exception is when there is a single non-negativity constraint, e.g.  $\beta_1 \geq 0$  with  $q = 1$ . In this case the constrained estimator can be found by two-step approach. First compute the unconstrained estimator  $\hat{\beta}$ . If  $\hat{\beta}_1 \geq 0$  then  $\tilde{\beta} = \hat{\beta}$ . Second, if  $\hat{\beta}_1 < 0$  then impose  $\beta_1 = 0$  (eliminate the regressor  $X_1$ ) and re-estimate. This yields the constrained least-squares estimator. While this method works when there is a single non-negativity constraint, it does not immediately generalize to other contexts.

The computational problems (8.50) and (8.51) are examples of **quadratic programming** problems. Quick and easy computer algorithms are available in programming languages including MATLAB, GAUSS and R.

Inference on inequality-constrained estimators is unfortunately quite challenging. The conventional asymptotic theory gives rise to the following dichotomy. If the true parameter satisfies the strict inequality  $\mathbf{r}(\beta) > \mathbf{0}$ , then asymptotically the estimator is not subject to the constraint and the inequality-constrained estimator has an asymptotic distribution equal to the unconstrained case. However if the true parameter is on the boundary, e.g.  $\mathbf{r}(\beta) = \mathbf{0}$ , then the estimator has a truncated structure. This is easiest to see in the one-dimensional case. If we have an estimator  $\hat{\beta}$  which satisfies  $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} Z = N(0, V_\beta)$  and  $\beta = 0$ , then the constrained estimator  $\tilde{\beta} = \max[\hat{\beta}, 0]$  will have the asymptotic distribution  $\sqrt{n}\tilde{\beta} \xrightarrow{d} \max[Z, 0]$ , a “half-normal” distribution.

## 8.16 Technical Proofs\*

**Proof of Theorem 8.9, Equation (8.28).** Let  $\mathbf{R}_\perp$  be a full rank  $k \times (k - q)$  matrix satisfying  $\mathbf{R}'_\perp \mathbf{V}_\beta \mathbf{R} = \mathbf{0}$  and then set  $\mathbf{C} = [\mathbf{R}, \mathbf{R}_\perp]$  which is full rank and invertible. Then we can calculate that

$$\begin{aligned} \mathbf{C}' \mathbf{V}_\beta^* \mathbf{C} &= \begin{bmatrix} \mathbf{R}' \mathbf{V}_\beta^* \mathbf{R} & \mathbf{R}' \mathbf{V}_\beta^* \mathbf{R}_\perp \\ \mathbf{R}'_\perp \mathbf{V}_\beta^* \mathbf{R} & \mathbf{R}'_\perp \mathbf{V}_\beta^* \mathbf{R}_\perp \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}'_\perp \mathbf{V}_\beta \mathbf{R}_\perp \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} \mathbf{C}' \mathbf{V}_\beta(\mathbf{W}) \mathbf{C} &= \\ &= \begin{bmatrix} \mathbf{R}' \mathbf{V}_\beta^*(\mathbf{W}) \mathbf{R} & \mathbf{R}' \mathbf{V}_\beta^*(\mathbf{W}) \mathbf{R}_\perp \\ \mathbf{R}'_\perp \mathbf{V}_\beta^*(\mathbf{W}) \mathbf{R} & \mathbf{R}'_\perp \mathbf{V}_\beta^*(\mathbf{W}) \mathbf{R}_\perp \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}'_\perp \mathbf{V}_\beta \mathbf{R}_\perp + \mathbf{R}'_\perp \mathbf{W} \mathbf{R} (\mathbf{R}' \mathbf{W} \mathbf{R})^{-1} \mathbf{R}' \mathbf{V}_\beta \mathbf{R} (\mathbf{R}' \mathbf{W} \mathbf{R})^{-1} \mathbf{R}' \mathbf{W} \mathbf{R}_\perp \end{bmatrix}. \end{aligned}$$

Thus

$$\begin{aligned} & \mathbf{C}' \left( V_{\beta}(\mathbf{W}) - V_{\beta}^* \right) \mathbf{C} \\ &= \mathbf{C}' V_{\beta}(\mathbf{W}) \mathbf{C} - \mathbf{C}' V_{\beta}^* \mathbf{C} \\ &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}'_{\perp} \mathbf{W} \mathbf{R} (\mathbf{R}' \mathbf{W} \mathbf{R})^{-1} \mathbf{R}' V_{\beta} \mathbf{R} (\mathbf{R}' \mathbf{W} \mathbf{R})^{-1} \mathbf{R}' \mathbf{W} \mathbf{R}_{\perp} \end{bmatrix} \\ &\geq 0 \end{aligned}$$

Since  $\mathbf{C}$  is invertible it follows that  $V_{\beta}(\mathbf{W}) - V_{\beta}^* \geq 0$  which is (8.28). ■

**Proof of Theorem 8.10.** We show the result for the minimum distance estimator  $\tilde{\beta} = \tilde{\beta}_{\text{md}}$ , as the proof for the constrained least-squares estimator is similar. For simplicity we assume that the constrained estimator is consistent  $\tilde{\beta} \xrightarrow{p} \beta$ . This can be shown with more effort, but requires a deeper treatment than appropriate for this textbook.

For each element  $r_j(\beta)$  of the  $q$ -vector  $\mathbf{r}(\beta)$ , by the mean value theorem there exists a  $\beta_j^*$  on the line segment joining  $\tilde{\beta}$  and  $\beta$  such that

$$\mathbf{r}_j(\tilde{\beta}) = \mathbf{r}_j(\beta) + \frac{\partial}{\partial \beta} \mathbf{r}_j(\beta_j^*)' (\tilde{\beta} - \beta). \quad (8.52)$$

Let  $\mathbf{R}_n^*$  be the  $k \times q$  matrix

$$\mathbf{R}^* = \begin{bmatrix} \frac{\partial}{\partial \beta} \mathbf{r}_1(\beta_1^*) & \frac{\partial}{\partial \beta} \mathbf{r}_2(\beta_2^*) & \cdots & \frac{\partial}{\partial \beta} \mathbf{r}_q(\beta_q^*) \end{bmatrix}.$$

Since  $\tilde{\beta} \xrightarrow{p} \beta$  it follows that  $\beta_j^* \xrightarrow{p} \beta$ , and by the CMT,  $\mathbf{R}^* \xrightarrow{p} \mathbf{R}$ . Stacking the (8.52), we obtain

$$\mathbf{r}(\tilde{\beta}) = \mathbf{r}(\beta) + \mathbf{R}^{*'} (\tilde{\beta} - \beta).$$

Since  $\mathbf{r}(\tilde{\beta}) = \mathbf{0}$  by construction and  $\mathbf{r}(\beta) = \mathbf{0}$  by Assumption 8.1, this implies

$$\mathbf{0} = \mathbf{R}^{*'} (\tilde{\beta} - \beta). \quad (8.53)$$

The first-order condition for (8.47) is

$$\widehat{\mathbf{W}} (\widehat{\beta} - \tilde{\beta}) = \widehat{\mathbf{R}} \tilde{\lambda}.$$

where  $\widehat{\mathbf{R}}$  is defined in (8.48).

Premultiplying by  $\mathbf{R}^{*'} \widehat{\mathbf{W}}^{-1}$ , inverting, and using (8.53), we find

$$\tilde{\lambda} = \left( \mathbf{R}^{*'} \widehat{\mathbf{W}}^{-1} \widehat{\mathbf{R}} \right)^{-1} \mathbf{R}^{*'} (\widehat{\beta} - \tilde{\beta}) = \left( \mathbf{R}^{*'} \widehat{\mathbf{W}}^{-1} \widehat{\mathbf{R}} \right)^{-1} \mathbf{R}^{*'} (\widehat{\beta} - \beta).$$

Thus

$$\tilde{\beta} - \beta = \left( \mathbf{I} - \widehat{\mathbf{W}}^{-1} \widehat{\mathbf{R}} \left( \mathbf{R}_n^{*'} \widehat{\mathbf{W}}^{-1} \widehat{\mathbf{R}} \right)^{-1} \mathbf{R}_n^{*'} \right) (\widehat{\beta} - \beta). \quad (8.54)$$

From Theorem 7.3 and Theorem 7.6 we find

$$\begin{aligned} \sqrt{n} (\tilde{\beta} - \beta) &= \left( \mathbf{I} - \widehat{\mathbf{W}}^{-1} \widehat{\mathbf{R}} \left( \mathbf{R}_n^{*'} \widehat{\mathbf{W}}^{-1} \widehat{\mathbf{R}} \right)^{-1} \mathbf{R}_n^{*'} \right) \sqrt{n} (\widehat{\beta} - \beta) \\ &\xrightarrow{d} \left( \mathbf{I} - \mathbf{W}^{-1} \mathbf{R} (\mathbf{R}' \mathbf{W}^{-1} \mathbf{R})^{-1} \mathbf{R}' \right) \mathbf{N}(\mathbf{0}, \mathbf{V}_{\beta}) \\ &= \mathbf{N}(\mathbf{0}, \mathbf{V}_{\beta}(\mathbf{W})). \end{aligned}$$

■

## Exercises

**Exercise 8.1** In the model  $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$ , show directly from definition (8.3) that the CLS estimate of  $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$  subject to the constraint that  $\boldsymbol{\beta}_2 = \mathbf{0}$  is the OLS regression of  $\mathbf{y}$  on  $\mathbf{X}_1$ .

**Exercise 8.2** In the model  $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$ , show directly from definition (8.3) that the CLS estimate of  $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ , subject to the constraint that  $\boldsymbol{\beta}_1 = \mathbf{c}$  (where  $\mathbf{c}$  is some given vector) is the OLS regression of  $\mathbf{y} - \mathbf{X}_1\mathbf{c}$  on  $\mathbf{X}_2$ .

**Exercise 8.3** In the model  $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$ , with  $\mathbf{X}_1$  and  $\mathbf{X}_2$  each  $n \times k$ , find the CLS estimate of  $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ , subject to the constraint that  $\boldsymbol{\beta}_1 = -\boldsymbol{\beta}_2$ .

**Exercise 8.4** In the linear projection model  $y_i = \alpha + \mathbf{x}'_i \boldsymbol{\beta} + e_i$ , consider the restriction  $\boldsymbol{\beta} = \mathbf{0}$ .

- (a) Find the constrained least-squares (CLS) estimator of  $\alpha$  under the restriction  $\boldsymbol{\beta} = \mathbf{0}$ .
- (b) Find an expression for the efficient minimum distance estimator of  $\alpha$  under the restriction  $\boldsymbol{\beta} = \mathbf{0}$ .

**Exercise 8.5** Verify that for  $\tilde{\boldsymbol{\beta}}_{\text{cls}}$  defined in (8.8) that  $\mathbf{R}'\tilde{\boldsymbol{\beta}}_{\text{cls}} = \mathbf{c}$ .

**Exercise 8.6** Prove Theorem 8.1.

**Exercise 8.7** Prove Theorem 8.2, that is,  $\mathbb{E}(\tilde{\boldsymbol{\beta}}_{\text{cls}} | \mathbf{X}) = \boldsymbol{\beta}$ , under the assumptions of the linear regression regression model and (8.1).

Hint: Use Theorem 8.1.

**Exercise 8.8** Prove Theorem 8.3.

**Exercise 8.9** Prove Theorem 8.4, that is,  $\mathbb{E}(s_{\text{cls}}^2 | \mathbf{X}) = \sigma^2$ , under the assumptions of the homoskedastic regression model and (8.1).

**Exercise 8.10** Verify (8.22) and (8.23), and that the minimum distance estimator  $\tilde{\boldsymbol{\beta}}_{\text{md}}$  with  $\widehat{\mathbf{W}} = \widehat{\mathbf{Q}}_{xx}$  equals the CLS estimator.

**Exercise 8.11** Prove Theorem 8.6.

**Exercise 8.12** Prove Theorem 8.7.

**Exercise 8.13** Prove Theorem 8.8. (Hint: Use that CLS is a special case of Theorem 8.7.)

**Exercise 8.14** Verify that (8.26) is  $V_{\boldsymbol{\beta}}(\mathbf{W})$  with  $\mathbf{W} = \mathbf{V}_{\boldsymbol{\beta}}^{-1}$ .

**Exercise 8.15** Prove (8.27). Hint: Use (8.26).

**Exercise 8.16** Verify (8.29), (8.30) and (8.31).

**Exercise 8.17** Verify (8.32), (8.33), and (8.34).

**Exercise 8.18** Suppose you have two independent samples

$$y_{1i} = \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + e_{1i}$$

and

$$y_{2i} = \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + e_{2i}$$

both of sample size  $n$ , and both  $\mathbf{x}_{1i}$  and  $\mathbf{x}_{2i}$  are  $k \times 1$ . You estimate  $\boldsymbol{\beta}_1$  and  $\boldsymbol{\beta}_2$  by OLS on each sample,  $\hat{\boldsymbol{\beta}}_1$  and  $\hat{\boldsymbol{\beta}}_2$ , say, with asymptotic covariance matrix estimators  $\hat{V}_{\boldsymbol{\beta}_1}$  and  $\hat{V}_{\boldsymbol{\beta}_2}$  (which are consistent for the asymptotic covariance matrices  $V_{\boldsymbol{\beta}_1}$  and  $V_{\boldsymbol{\beta}_2}$ ). Consider efficient minimum distance estimation under the restriction  $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ .

- (a) Find the estimator  $\tilde{\boldsymbol{\beta}}$  of  $\boldsymbol{\beta} = \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ .
- (b) Find the asymptotic distribution of  $\tilde{\boldsymbol{\beta}}$ .
- (c) How would you approach the problem if the sample sizes are different, say  $n_1$  and  $n_2$ ?

**Exercise 8.19** As in Exercise 7.29 and 3.26, use the CPS dataset and the subsample of white male Hispanics.

- (a) Estimate the regression

$$\widehat{\log(Wage)} = \beta_1 education + \beta_2 experience + \beta_3 experience^2/100 + \beta_4 Married_1 + \beta_5 Married_2 + \beta_6 Married_3 + \beta_7 Widowed + \beta_8 Divorced + \beta_9 Separated + \beta_{10}$$

where  $Married_1$ ,  $Married_2$ , and  $Married_3$  are the first three marital status codes as listed in Section 3.22.

- (b) Estimate the equation using constrained least-squares, imposing the constraints  $\beta_4 = \beta_7$  and  $\beta_8 = \beta_9$ , and report the estimates and standard errors.
- (c) Estimate the equation using efficient minimum distance, imposing the same constraints, and report the estimates and standard errors.
- (d) Under what constraint on the coefficients is the wage equation non-decreasing in experience for experience up to 50?
- (e) Estimate the equation imposing  $\beta_4 = \beta_7$ ,  $\beta_8 = \beta_9$ , and the inequality from part (d).

**Exercise 8.20** Take the model

$$\begin{aligned} y_i &= m(x_i) + e_i \\ m(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p \\ \mathbb{E}(z_i e_i) &= 0 \\ z_i &= (1, x_i, \dots, x_i^p)' \\ g(x) &= \frac{d}{dx} m(x) \end{aligned}$$

with i.i.d. observations  $(y_i, x_i)$ ,  $i = 1, \dots, n$ . The order of the polynomial  $p$  is known.

- (a) How should we interpret the function  $m(x)$  given the projection assumption? How should we interpret  $g(x)$ ? (Briefly)
- (b) Describe an estimator  $\hat{g}(x)$  of  $g(x)$ .
- (c) Find the asymptotic distribution of  $\sqrt{n}(\hat{g}(x) - g(x))$  as  $n \rightarrow \infty$ .
- (d) Show how to construct an asymptotic 95% confidence interval for  $g(x)$  (for a single  $x$ ).
- (e) Assume  $p = 2$ . Describe how to estimate  $g(x)$  imposing the constraint that  $m(x)$  is concave.
- (f) Assume  $p = 2$ . Describe how to estimate  $g(x)$  imposing the constraint that  $m(u)$  is increasing on the region  $u \in [x_L, x_U]$ .

**Exercise 8.21** Take the linear model with restrictions

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) &= \mathbf{0} \\ \mathbf{R}' \boldsymbol{\beta} &= \mathbf{c} \end{aligned}$$

with  $n$  observations. Consider three estimators for  $\boldsymbol{\beta}$

- $\hat{\beta}$ , the unconstrained least squares estimator
- $\tilde{\beta}$ , the constrained least squares estimator
- $\bar{\beta}$ , the constrained efficient minimum distance estimator

For each estimator, define the residuals  $\hat{e}_i = y_i - \mathbf{x}'_i \hat{\beta}$ ,  $\tilde{e}_i = y_i - \mathbf{x}'_i \tilde{\beta}$ ,  $\bar{e}_i = y_i - \mathbf{x}'_i \bar{\beta}$ , and variance estimators  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{e}_i^2$ ,  $\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \tilde{e}_i^2$ , and  $\bar{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \bar{e}_i^2$ .

(a) As  $\bar{\beta}$  is the most efficient estimator and  $\hat{\beta}$  the least, do you expect that  $\bar{\sigma}^2 < \tilde{\sigma}^2 < \hat{\sigma}^2$ , in large samples?

(b) Consider the statistic

$$T_n = \hat{\sigma}^{-2} \sum_{i=1}^n (\hat{e}_i - \tilde{e}_i)^2.$$

Find the asymptotic distribution for  $T_n$  when  $\mathbf{R}'\beta = \mathbf{c}$  is true.

(c) Does the result of the previous question simplify when the error  $e_i$  is homoskedastic?

**Exercise 8.22** Take the linear model

$$\begin{aligned} y_i &= x_{1i}\beta_1 + x_{2i}\beta_2 + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) &= \mathbf{0} \end{aligned}$$

with  $n$  observations. Consider the restriction

$$\frac{\beta_1}{\beta_2} = 2.$$

- (a) Find an explicit expression for the constrained least-squares (CLS) estimator  $\tilde{\beta} = (\tilde{\beta}_1, \tilde{\beta}_2)$  of  $\beta = (\beta_1, \beta_2)$  under the restriction. Your answer should be specific to the restriction, it should not be a generic formula for an abstract general restriction.
- (b) Derive the asymptotic distribution of  $\tilde{\beta}_1$  under the assumption that the restriction is true.

# Chapter 9

## Hypothesis Testing

In Chapter 5 we briefly introduced hypothesis testing in the context of the normal regression model. In this chapter we explore hypothesis testing in greater detail, with a particular emphasis on asymptotic inference.

### 9.1 Hypotheses

In Chapter 8 we discussed estimation subject to restrictions, including linear restrictions (8.1), non-linear restrictions (8.44), and inequality restrictions (8.49). In this chapter we discuss **tests** of such restrictions.

Hypothesis tests attempt to assess whether there is evidence to contradict a proposed parametric restriction. Let

$$\boldsymbol{\theta} = \mathbf{r}(\boldsymbol{\beta})$$

be a  $q \times 1$  parameter of interest where  $\mathbf{r} : \mathbb{R}^k \rightarrow \Theta \subset \mathbb{R}^q$  is some transformation. For example,  $\boldsymbol{\theta}$  may be a single coefficient, e.g.  $\boldsymbol{\theta} = \beta_j$ , the difference between two coefficients, e.g.  $\boldsymbol{\theta} = \beta_j - \beta_\ell$ , or the ratio of two coefficients, e.g.  $\boldsymbol{\theta} = \beta_j / \beta_\ell$ .

A point hypothesis concerning  $\boldsymbol{\theta}$  is a proposed restriction such as

$$\boldsymbol{\theta} = \boldsymbol{\theta}_0 \tag{9.1}$$

where  $\boldsymbol{\theta}_0$  is a hypothesized (known) value.

More generally, letting  $\boldsymbol{\beta} \in \mathcal{B} \subset \mathbb{R}^k$  be the parameter space, a hypothesis is a restriction  $\boldsymbol{\beta} \in \mathcal{B}_0$  where  $\mathcal{B}_0$  is a proper subset of  $\mathcal{B}$ . This specializes to (9.1) by setting  $\mathcal{B}_0 = \{\boldsymbol{\beta} \in \mathcal{B} : \mathbf{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0\}$ .

In this chapter we will focus exclusively on point hypotheses of the form (9.1) as they are the most common and relatively simple to handle.

The hypothesis to be tested is called the null hypothesis.

**Definition 9.1** The **null hypothesis**, written  $\mathbb{H}_0$ , is the restriction  $\boldsymbol{\theta} = \boldsymbol{\theta}_0$  or  $\boldsymbol{\beta} \in \mathcal{B}_0$ .

We often write the null hypothesis as  $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$  or  $\mathbb{H}_0 : \mathbf{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$ .

The complement of the null hypothesis (the collection of parameter values which do not satisfy the null hypothesis) is called the alternative hypothesis.

**Definition 9.2** The **alternative hypothesis**, written  $\mathbb{H}_1$ , is the set  $\{\boldsymbol{\theta} \in \Theta : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0\}$  or  $\{\boldsymbol{\beta} \in \mathcal{B} : \boldsymbol{\beta} \notin \mathcal{B}_0\}$ .

We often write the alternative hypothesis as  $H_1 : \theta \neq \theta_0$  or  $H_1 : r(\beta) \neq \theta_0$ . For simplicity, we often refer to the hypotheses as “the null” and “the alternative”. Figure 9.1 illustrates the division of the parameter space into null and alternative hypotheses.

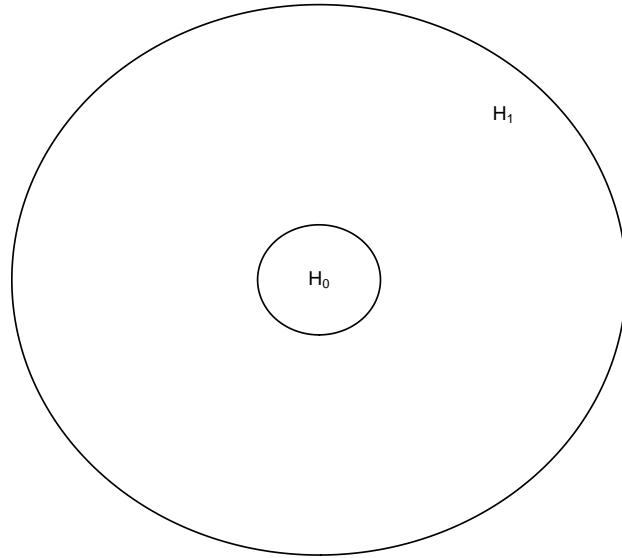


Figure 9.1: Null and Alternative Hypotheses

In hypothesis testing, we assume that there is a true (but unknown) value of  $\theta$  and this value either satisfies  $H_0$  or does not satisfy  $H_0$ . The goal of hypothesis testing is to assess whether or not  $H_0$  is true, by asking if  $H_0$  is consistent with the observed data.

To be specific, take our example of wage determination and consider the question: Does union membership affect wages? We can turn this into a hypothesis test by specifying the null as the restriction that a coefficient on union membership is zero in a wage regression. Consider, for example, the estimates reported in Table 4.1. The coefficient for “Male Union Member” is 0.095 (a wage premium of 9.5%) and the coefficient for “Female Union Member” is 0.022 (a wage premium of 2.2%). These are estimates, not the true values. The question is: Are the true coefficients zero? To answer this question, the testing method asks the question: Are the observed estimates compatible with the hypothesis, in the sense that the deviation from the hypothesis can be reasonably explained by stochastic variation? Or are the observed estimates incompatible with the hypothesis, in the sense that the observed estimates would be highly unlikely if the hypothesis were true?

## 9.2 Acceptance and Rejection

A hypothesis test either accepts the null hypothesis or rejects the null hypothesis in favor of the alternative hypothesis. We can describe these two decisions as “Accept  $H_0$ ” and “Reject  $H_0$ ”. In the example given in the previous section, the decision would be either to accept the hypothesis that union membership does not affect wages, or to reject the hypothesis in favor of the alternative that union membership does affect wages.

The decision is based on the data, and so is a mapping from the sample space to the decision set.

This splits the sample space into two regions  $S_0$  and  $S_1$  such that if the observed sample falls into  $S_0$  we accept  $H_0$ , while if the sample falls into  $S_1$  we reject  $H_0$ . The set  $S_0$  is called the **acceptance region** and the set  $S_1$  the **rejection or critical region**.

It is convenient to express this mapping as a real-valued function called a **test statistic**

$$T = T((y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n))$$

relative to a **critical value**  $c$ . The hypothesis test then consists of the decision rule

1. Accept  $H_0$  if  $T \leq c$ .
2. Reject  $H_0$  if  $T > c$ .

Figure 9.2 illustrates the division of the sample space into acceptance and rejection regions.

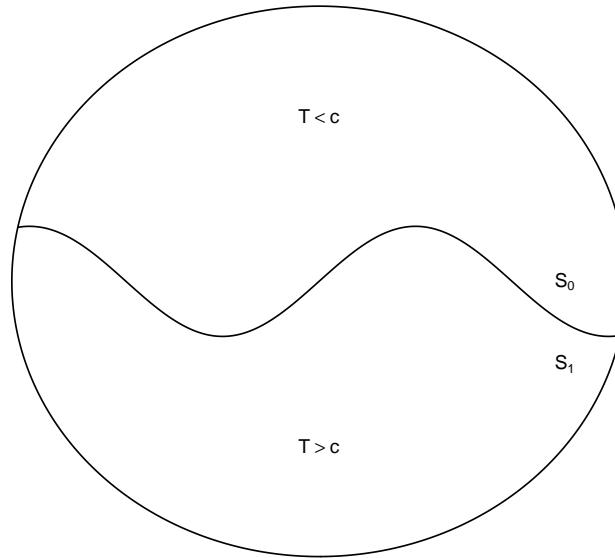


Figure 9.2: Acceptance and Rejection Regions

A test statistic  $T$  should be designed so that small values are likely when  $H_0$  is true and large values are likely when  $H_1$  is true. There is a well developed statistical theory concerning the design of optimal tests. We will not review that theory here, but instead refer the reader to Lehmann and Romano (2005). In this chapter we will summarize the main approaches to the design of test statistics.

The most commonly used test statistic is the absolute value of the t-statistic

$$T = |T(\theta_0)| \tag{9.2}$$

where

$$T(\theta) = \frac{\hat{\theta} - \theta}{s(\hat{\theta})} \tag{9.3}$$

is the t-statistic from (7.33),  $\hat{\theta}$  is a point estimate, and  $s(\hat{\theta})$  its standard error.  $T$  is an appropriate statistic when testing hypotheses on individual coefficients or real-valued parameters  $\theta = h(\boldsymbol{\beta})$  and  $\theta_0$  is the hypothesized value. Quite typically,  $\theta_0 = 0$ , as interest focuses on whether or not a coefficient equals zero, but this is not the only possibility. For example, interest may focus on whether an elasticity  $\theta$  equals 1, in which case we may wish to test  $H_0 : \theta = 1$ .

### 9.3 Type I Error

A false rejection of the null hypothesis  $\mathbb{H}_0$  (rejecting  $\mathbb{H}_0$  when  $\mathbb{H}_0$  is true) is called a **Type I error**. The probability of a Type I error is

$$\mathbb{P}(\text{Reject } \mathbb{H}_0 \mid \mathbb{H}_0 \text{ true}) = \mathbb{P}(T > c \mid \mathbb{H}_0 \text{ true}). \quad (9.4)$$

The finite sample **size** of the test is defined as the supremum of (9.4) across all data distributions which satisfy  $\mathbb{H}_0$ . A primary goal of test construction is to limit the incidence of Type I error by bounding the size of the test.

For the reasons discussed in Chapter 7, in typical econometric models the exact sampling distributions of estimators and test statistics are unknown and hence we cannot explicitly calculate (9.4). Instead, we typically rely on asymptotic approximations. Suppose that the test statistic has an asymptotic distribution under  $\mathbb{H}_0$ . That is, when  $\mathbb{H}_0$  is true

$$T \xrightarrow{d} \xi \quad (9.5)$$

as  $n \rightarrow \infty$  for some continuously-distributed random variable  $\xi$ . This is not a substantive restriction as most conventional econometric tests satisfy (9.5). Let  $G(u) = \mathbb{P}(\xi \leq u)$  denote the distribution of  $\xi$ . We call  $\xi$  (or  $G$ ) the **asymptotic null distribution**.

It is generally desirable to design test statistics  $T$  whose asymptotic null distribution  $G$  is known and does not depend on unknown parameters. In this case we say that the statistic  $T$  is **asymptotically pivotal**.

For example, if the test statistic equals the absolute t-statistic from (9.2), then we know from Theorem 7.11 that if  $\theta = \theta_0$  (that is, the null hypothesis holds), then  $T \xrightarrow{d} |Z|$  as  $n \rightarrow \infty$  where  $Z \sim N(0, 1)$ . This means that  $G(u) = \mathbb{P}(|Z| \leq u) = 2\Phi(u) - 1$ , the distribution of the absolute value of the standard normal as shown in (7.34). This distribution does not depend on unknowns and is pivotal.

We define the **asymptotic size** of the test as the asymptotic probability of a Type I error:

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P}(T > c \mid \mathbb{H}_0 \text{ true}) &= \mathbb{P}(\xi > c) \\ &= 1 - G(c). \end{aligned}$$

We see that the asymptotic size of the test is a simple function of the asymptotic null distribution  $G$  and the critical value  $c$ . For example, the asymptotic size of a test based on the absolute t-statistic with critical value  $c$  is  $2(1 - \Phi(c))$ .

In the dominant approach to hypothesis testing, the researcher pre-selects a **significance level**  $\alpha \in (0, 1)$  and then selects  $c$  so that the (asymptotic) size is no larger than  $\alpha$ . When the asymptotic null distribution  $G$  is pivotal, we can accomplish this by setting  $c$  equal to the  $1 - \alpha$  quantile of the distribution  $G$ . (If the distribution  $G$  is not pivotal, more complicated methods must be used, pointing out the great convenience of using asymptotically pivotal test statistics.) We call  $c$  the **asymptotic critical value** because it has been selected from the asymptotic null distribution. For example, since  $2(1 - \Phi(1.96)) = 0.05$ , it follows that the 5% asymptotic critical value for the absolute t-statistic is  $c = 1.96$ . Calculation of normal critical values is done numerically in statistical software. For example, in MATLAB the command is `norminv(1-alpha/2)`.

### 9.4 t tests

As we mentioned earlier, the most common test of the one-dimensional hypothesis

$$\mathbb{H}_0 : \theta = \theta_0$$

against the alternative

$$\mathbb{H}_1 : \theta \neq \theta_0$$

is the absolute value of the t-statistic (9.3). We now formally state its asymptotic null distribution, which is a simple application of Theorem 7.11.

**Theorem 9.1** Under Assumptions 7.2, 7.3, and  $\mathbb{H}_0 : \theta = \theta_0$ ,

$$T(\theta_0) \xrightarrow{d} Z.$$

For  $c$  satisfying  $\alpha = 2(1 - \Phi(c))$ ,

$$\mathbb{P}(|T(\theta_0)| > c | \mathbb{H}_0) \longrightarrow \alpha,$$

and the test “Reject  $\mathbb{H}_0$  if  $|T(\theta_0)| > c$ ” has asymptotic size  $\alpha$ .

The theorem shows that asymptotic critical values can be taken from the normal distribution. As in our discussion of asymptotic confidence intervals (Section 7.13), the critical value could alternatively be taken from the student  $t$  distribution, which would be the exact test in the normal regression model (Section 5.15). Indeed,  $t$  critical values are the default in packages such as Stata. Since the critical values from the student  $t$  distribution are (slightly) larger than those from the normal distribution, using student  $t$  critical values decreases the rejection probability of the test. In practical applications the difference is typically unimportant unless the sample size is quite small (in which case the asymptotic approximation should be questioned as well).

The alternative hypothesis  $\theta \neq \theta_0$  is sometimes called a “two-sided” alternative. In contrast, sometimes we are interested in testing for one-sided alternatives such as

$$\mathbb{H}_1 : \theta > \theta_0$$

or

$$\mathbb{H}_1 : \theta < \theta_0.$$

Tests of  $\theta = \theta_0$  against  $\theta > \theta_0$  or  $\theta < \theta_0$  are based on the signed t-statistic  $T = T(\theta_0)$ . The hypothesis  $\theta = \theta_0$  is rejected in favor of  $\theta > \theta_0$  if  $T > c$  where  $c$  satisfies  $\alpha = 1 - \Phi(c)$ . Negative values of  $T$  are not taken as evidence against  $\mathbb{H}_0$ , as point estimates  $\hat{\theta}$  less than  $\theta_0$  do not point to  $\theta > \theta_0$ . Since the critical values are taken from the single tail of the normal distribution, they are smaller than for two-sided tests. Specifically, the asymptotic 5% critical value is  $c = 1.645$ . Thus, we reject  $\theta = \theta_0$  in favor of  $\theta > \theta_0$  if  $T > 1.645$ .

Conversely, tests of  $\theta = \theta_0$  against  $\theta < \theta_0$  reject  $\mathbb{H}_0$  for negative t-statistics, e.g. if  $T \leq -c$ . For this alternative large positive values of  $T$  are not evidence against  $\mathbb{H}_0$ . An asymptotic 5% test rejects if  $T < -1.645$ .

There seems to be an ambiguity. Should we use the two-sided critical value 1.96 or the one-sided critical value 1.645? The answer is that in most cases the two-sided critical value is appropriate. We should use the one-sided critical values only when the parameter space is known to satisfy a one-sided restriction such as  $\theta \geq \theta_0$ . This is when the test of  $\theta = \theta_0$  against  $\theta > \theta_0$  makes sense. If the restriction  $\theta \geq \theta_0$  is not known *a priori*, then imposing this restriction to test  $\theta = \theta_0$  against  $\theta > \theta_0$  does not make sense. Since linear regression coefficients typically do not have *a priori* sign restrictions, the standard convention is to use two-sided critical values.

This may seem contrary to the way testing is presented in statistical textbooks, which often focus on one-sided alternative hypotheses. The latter focus is primarily for pedagogy, as the one-sided theoretical problem is cleaner and easier to understand.

## 9.5 Type II Error and Power

A false acceptance of the null hypothesis  $H_0$  (accepting  $H_0$  when  $H_1$  is true) is called a **Type II error**. The rejection probability under the alternative hypothesis is called the **power** of the test, and equals 1 minus the probability of a Type II error:

$$\pi(\theta) = \mathbb{P}(\text{Reject } H_0 | H_1 \text{ true}) = \mathbb{P}(T > c | H_1 \text{ true}).$$

We call  $\pi(\theta)$  the **power function** and is written as a function of  $\theta$  to indicate its dependence on the true value of the parameter  $\theta$ .

In the dominant approach to hypothesis testing, the goal of test construction is to have high power subject to the constraint that the size of the test is lower than the pre-specified significance level. Generally, the power of a test depends on the true value of the parameter  $\theta$ , and for a well behaved test the power is increasing both as  $\theta$  moves away from the null hypothesis  $\theta_0$  and as the sample size  $n$  increases.

Given the two possible states of the world ( $H_0$  or  $H_1$ ) and the two possible decisions (Accept  $H_0$  or Reject  $H_0$ ), there are four possible pairings of states and decisions as is depicted in Table 9.1.

Table 9.1: Hypothesis Testing Decisions

	Accept $H_0$	Reject $H_0$
$H_0$ true	Correct Decision	Type I Error
$H_1$ true	Type II Error	Correct Decision

Given a test statistic  $T$ , increasing the critical value  $c$  increases the acceptance region  $S_0$  while decreasing the rejection region  $S_1$ . This decreases the likelihood of a Type I error (decreases the size) but increases the likelihood of a Type II error (decreases the power). Thus the choice of  $c$  involves a trade-off between size and the power. This is why the significance level  $\alpha$  of the test cannot be set arbitrarily small. (Otherwise the test will not have meaningful power.)

It is important to consider the power of a test when interpreting hypothesis tests, as an overly narrow focus on size can lead to poor decisions. For example, it is easy to design a test which has perfect size yet has trivial power. Specifically, for any hypothesis we can use the following test: Generate a random variable  $U \sim U[0, 1]$  and reject  $H_0$  if  $U < \alpha$ . This test has exact size of  $\alpha$ . Yet the test also has power precisely equal to  $\alpha$ . When the power of a test equals the size, we say that the test has **trivial power**. Nothing is learned from such a test.

## 9.6 Statistical Significance

Testing requires a pre-selected choice of significance level  $\alpha$ , yet there is no objective scientific basis for choice of  $\alpha$ . Nevertheless the common practice is to set  $\alpha = 0.05$  (5%). Alternative values are  $\alpha = 0.10$  (10%) and  $\alpha = 0.01$  (1%). These choices are somewhat the by-product of traditional tables of critical values and statistical software.

The informal reasoning behind the choice of a 5% critical value is to ensure that Type I errors should be relatively unlikely – that the decision “Reject  $H_0$ ” has scientific strength – yet the test retains power against reasonable alternatives. The decision “Reject  $H_0$ ” means that the evidence is inconsistent with the null hypothesis, in the sense that it is relatively unlikely (1 in 20) that data generated by the null hypothesis would yield the observed test result.

In contrast, the decision “Accept  $H_0$ ” is not a strong statement. It does not mean that the evidence supports  $H_0$ , only that there is insufficient evidence to reject  $H_0$ . Because of this, it is more accurate to use the label “Do not Reject  $H_0$ ” instead of “Accept  $H_0$ ”.

When a test rejects  $H_0$  at the 5% significance level it is common to say that the statistic is **statistically significant** and if the test accepts  $H_0$  it is common to say that the statistic is **not statistically significant** or that it is **statistically insignificant**. It is helpful to remember that this is simply a compact way of

saying “Using the statistic  $T$ , the hypothesis  $H_0$  can [cannot] be rejected at the asymptotic 5% level.” Furthermore, when the null hypothesis  $H_0 : \theta = 0$  is rejected it is common to say that the coefficient  $\theta$  is statistically significant, because the test has rejected the hypothesis that the coefficient is equal to zero.

Let us return to the example about the union wage premium as measured in Table 4.1. The absolute t-statistic for the coefficient on “Male Union Member” is  $0.095/0.020 = 4.7$ , which is greater than the 5% asymptotic critical value of 1.96. Therefore we reject the hypothesis that union membership does not affect wages for men. In this case, we can say that union membership is statistically significant for men. However, the absolute t-statistic for the coefficient on “Female Union Member” is  $0.023/0.020 = 1.2$ , which is less than 1.96 and therefore we do not reject the hypothesis that union membership does not affect wages for women. In this case we find that membership for women is not statistically significant.

When a test accepts a null hypothesis (when a test is not statistically significant) a common misinterpretation is that this is evidence that the null hypothesis is true. This is incorrect. Failure to reject is by itself not evidence. Without an analysis of power, we do not know the likelihood of making a Type II error, and thus are uncertain. In our wage example, it would be a mistake to write that “the regression finds that female union membership has no effect on wages”. This is an incorrect and most unfortunate interpretation. The test has failed to reject the hypothesis that the coefficient is zero, but that does not mean that the coefficient is actually zero.

When a test rejects a null hypothesis (when a test is statistically significant) it is strong evidence against the hypothesis (since if the hypothesis were true then rejection is an unlikely event). Rejection should be taken as evidence against the null hypothesis. However, we can never conclude that the null hypothesis is indeed false, as we cannot exclude the possibility that we are making a Type I error.

Perhaps more importantly, there is an important distinction between statistical and economic significance. If we correctly reject the hypothesis  $H_0 : \theta = 0$  it means that the true value of  $\theta$  is non-zero. This includes the possibility that  $\theta$  may be non-zero but close to zero in magnitude. This only makes sense if we interpret the parameters in the context of their relevant models. In our wage regression example, we might consider wage effects of 1% magnitude or less as being “close to zero”. In a log wage regression this corresponds to a dummy variable with a coefficient less than 0.01. If the standard error is sufficiently small (less than 0.005) then a coefficient estimate of 0.01 will be statistically significant but not economically significant. This occurs frequently in applications with very large sample sizes where standard errors can be quite small.

The solution is to focus whenever possible on confidence intervals and the economic meaning of the coefficients. For example, if the coefficient estimate is 0.005 with a standard error of 0.002 then a 95% confidence interval would be  $[0.001, 0.009]$  indicating that the true effect is likely between 0% and 1%, and hence is slightly positive but small. This is much more informative than the misleading statement “the effect is statistically positive”.

## 9.7 P-Values

Continuing with the wage regression estimates reported in Table 4.1, consider another question: Does marriage status affect wages? To test the hypothesis that marriage status has no effect on wages, we examine the t-statistics for the coefficients on “Married Male” and “Married Female” in Table 4.1, which are  $0.211/0.010 = 22$  and  $0.016/0.010 = 1.7$ , respectively. The first exceeds the asymptotic 5% critical value of 1.96, so we reject the hypothesis for men. The second is smaller than 1.96, so we fail to reject the hypothesis for women. Taking a second look at the statistics, we see that the statistic for men (22) is exceptionally high, and that for women (1.7) is only slightly below the critical value. Suppose that the t-statistic for women were slightly increased to 2.0. This is larger than the critical value so would lead to the decision “Reject  $H_0$ ” rather than “Accept  $H_0$ ”. Should we really be making a different decision if the t-statistic is 2.0 rather than 1.7? The difference in values is small, shouldn’t the difference in the decision be also small? Thinking through these examples it seems unsatisfactory to simply report “Accept  $H_0$ ” or “Reject  $H_0$ ”. These two decisions do not summarize the evidence. Instead, the magnitude of the statistic  $T$  suggests a “degree of evidence” against  $H_0$ . How can we take this into account?

The answer is to report what is known as the **asymptotic p-value**

$$p = 1 - G(T).$$

Since the distribution function  $G$  is monotonically increasing, the p-value is a monotonically decreasing function of  $T$  and is an equivalent test statistic. Instead of rejecting  $H_0$  at the significance level  $\alpha$  if  $T > c$ , we can reject  $H_0$  if  $p < \alpha$ . Thus it is sufficient to report  $p$ , and let the reader decide. In practice, the p-value is calculated numerically. For example, in MATLAB the command is `2*(1-normalcdf(abs(t)))`.

It is instructive to interpret  $p$  as the **marginal significance level**: the smallest value of  $\alpha$  for which the test  $T$  “rejects” the null hypothesis. That is,  $p = 0.11$  means that  $T$  rejects  $H_0$  for all significance levels greater than 0.11, but fails to reject  $H_0$  for significance levels less than 0.11.

Furthermore, the asymptotic p-value has a very convenient asymptotic null distribution. Since  $T \xrightarrow{d} \xi$  under  $H_0$ , then  $p = 1 - G(T) \xrightarrow{d} 1 - G(\xi)$ , which has the distribution

$$\begin{aligned} \mathbb{P}(1 - G(\xi) \leq u) &= \mathbb{P}(1 - u \leq G(\xi)) \\ &= 1 - \mathbb{P}(\xi \leq G^{-1}(1 - u)) \\ &= 1 - G(G^{-1}(1 - u)) \\ &= 1 - (1 - u) \\ &= u, \end{aligned}$$

which is the uniform distribution on  $[0, 1]$ . (This calculation assumes that  $G(u)$  is strictly increasing which is true for conventional asymptotic distributions such as the normal.) Thus  $p \xrightarrow{d} U[0, 1]$ . This means that the “unusualness” of  $p$  is easier to interpret than the “unusualness” of  $T$ .

An important caveat is that the p-value  $p$  should not be interpreted as the probability that either hypothesis is true. A common mis-interpretation is that  $p$  is the probability “that the null hypothesis is true.” This is incorrect. Rather,  $p$  is the marginal significance level – a measure of the strength of information against the null hypothesis.

For a t-statistic, the p-value can be calculated either using the normal distribution or the student  $t$  distribution, the latter presented in Section 5.15. p-values calculated using the student  $t$  will be slightly larger, though the difference is small when the sample size is large.

Returning to our empirical example, for the test that the coefficient on “Married Male” is zero, the p-value is 0.000. This means that it would be nearly impossible to observe a t-statistic as large as 22 when the true value of the coefficient is zero. When presented with such evidence we can say that we “strongly reject” the null hypothesis, that the test is “highly significant”, or that “the test rejects at any conventional critical value”. In contrast, the p-value for the coefficient on “Married Female” is 0.094. In this context it is typical to say that the test is “close to significant”, meaning that the p-value is larger than 0.05, but not too much larger.

A related (but inferior) empirical practice is to append asterisks (\*) to coefficient estimates or test statistics to indicate the level of significance. A common practice is to append a single asterisk (\*) for an estimate or test statistic which exceeds the 10% critical value (i.e., is significant at the 10% level), append a double asterisk (\*\*) for a test which exceeds the 5% critical value, and append a triple asterisk (\*\*\*) for a test which exceeds the 1% critical value. Such a practice can be better than a table of raw test statistics as the asterisks permit a quick interpretation of significance. On the other hand, asterisks are inferior to p-values, which are also easy and quick to interpret. The goal is essentially the same; it seems wiser to report p-values whenever possible and avoid the use of asterisks.

Our recommendation is that the best empirical practice is to compute and report the asymptotic p-value  $p$  rather than simply the test statistic  $T$ , the binary decision Accept/Reject, or appending asterisks. The p-value is a simple statistic, easy to interpret, and contains more information than the other choices.

We now summarize the main features of hypothesis testing.

1. Select a significance level  $\alpha$ .

2. Select a test statistic  $T$  with asymptotic distribution  $T \xrightarrow{d} \xi$  under  $\mathbb{H}_0$ .
3. Set the asymptotic critical value  $c$  so that  $1 - G(c) = \alpha$ , where  $G$  is the distribution function of  $\xi$ .
4. Calculate the asymptotic p-value  $p = 1 - G(T)$ .
5. Reject  $\mathbb{H}_0$  if  $T > c$ , or equivalently  $p < \alpha$ .
6. Accept  $\mathbb{H}_0$  if  $T \leq c$ , or equivalently  $p \geq \alpha$ .
7. Report  $p$  to summarize the evidence concerning  $\mathbb{H}_0$  versus  $\mathbb{H}_1$ .

## 9.8 t-ratios and the Abuse of Testing

In Section 4.19, we argued that a good applied practice is to report coefficient estimates  $\hat{\theta}$  and standard errors  $s(\hat{\theta})$  for all coefficients of interest in estimated models. With  $\hat{\theta}$  and  $s(\hat{\theta})$  the reader can easily construct confidence intervals  $[\hat{\theta} \pm 2s(\hat{\theta})]$  and t-statistics  $(\hat{\theta} - \theta_0) / s(\hat{\theta})$  for hypotheses of interest.

Some applied papers (especially older ones) report t-ratios  $T = \hat{\theta} / s(\hat{\theta})$  instead of standard errors. This is poor econometric practice. While the same information is being reported (you can back out standard errors by division, e.g.  $s(\hat{\theta}) = \hat{\theta} / T$ ), standard errors are generally more helpful to readers than t-ratios. Standard errors help the reader focus on the estimation precision and confidence intervals, while t-ratios focus attention on statistical significance. While statistical significance is important, it is less important that the parameter estimates themselves and their confidence intervals. The focus should be on the meaning of the parameter estimates, their magnitudes, and their interpretation, not on listing which variables have significant (e.g. non-zero) coefficients. In many modern applications, sample sizes are very large so standard errors can be very small. Consequently t-ratios can be large even if the coefficient estimates are economically small. In such contexts it may not be interesting to announce “The coefficient is non-zero!” Instead, what is interesting to announce is that “The coefficient estimate is economically interesting!”

In particular, some applied papers report coefficient estimates and t-ratios, and limit their discussion of the results to describing which variables are “significant” (meaning that their t-ratios exceed 2) and the signs of the coefficient estimates. This is very poor empirical work, and should be studiously avoided. It is also a recipe for banishment of your work to lower tier economics journals.

Fundamentally, the common t-ratio is a test for the hypothesis that a coefficient equals zero. This should be reported and discussed when this is an interesting economic hypothesis of interest. But if this is not the case, it is distracting.

One problem is that standard packages, such as Stata, by default report t-statistics and p-values for every estimated coefficient. While this can be useful (as a user doesn’t need to explicitly ask to test a desired coefficient) it can be misleading as it may unintentionally suggest that the entire list of t-statistics and p-values are important. Instead, a user should focus on tests of scientifically motivated hypotheses.

In general, when a coefficient  $\theta$  is of interest, it is constructive to focus on the point estimate, its standard error, and its confidence interval. The point estimate gives our “best guess” for the value. The standard error is a measure of precision. The confidence interval gives us the range of values consistent with the data. If the standard error is large then the point estimate is not a good summary about  $\theta$ . The endpoints of the confidence interval describe the bounds on the likely possibilities. If the confidence interval embraces too broad a set of values for  $\theta$ , then the dataset is not sufficiently informative to render useful inferences about  $\theta$ . On the other hand if the confidence interval is tight, then the data have produced an accurate estimate, and the focus should be on the value and interpretation of this estimate. In contrast, the statement “the t-ratio is highly significant” has little interpretive value.

The above discussion requires that the researcher knows what the coefficient  $\theta$  means (in terms of the economic problem) and can interpret values and magnitudes, not just signs. This is critical for good applied econometric practice.

For example, consider the question about the effect of marriage status on mean log wages. We had found that the effect is “highly significant” for men and “close to significant” for women. Now, let’s construct asymptotic 95% confidence intervals for the coefficients. The one for men is [0.19, 0.23] and that for women is [-0.00, 0.03]. This shows that average wages for married men are about 19-23% higher than for unmarried men, which is substantial, while the difference for women is about 0-3%, which is small. These *magnitudes* are more informative than the results of the hypothesis tests.

## 9.9 Wald Tests

The t-test is appropriate when the null hypothesis is a real-valued restriction. More generally, there may be multiple restrictions on the coefficient vector  $\beta$ . Suppose that we have  $q > 1$  restrictions which can be written in the form (9.1). It is natural to estimate  $\theta = r(\beta)$  by the plug-in estimator  $\hat{\theta} = r(\hat{\beta})$ . To test  $H_0 : \theta = \theta_0$  against  $H_1 : \theta \neq \theta_0$  one approach is to measure the magnitude of the discrepancy  $\hat{\theta} - \theta_0$ . As this is a vector, there is more than one measure of its length. One simple measure is the weighted quadratic form known as the **Wald statistic**. This is (7.37) evaluated at the null hypothesis

$$W = W(\theta_0) = (\hat{\theta} - \theta_0)' \hat{V}_{\hat{\theta}}^{-1} (\hat{\theta} - \theta_0) \quad (9.6)$$

where  $\hat{V}_{\hat{\theta}} = \hat{R}' \hat{V}_{\hat{\beta}} \hat{R}$  is an estimator of  $V_{\hat{\theta}}$  and  $\hat{R} = \frac{\partial}{\partial \beta} r(\hat{\beta})'$ . Notice that we can write  $W$  alternatively as

$$W = n (\hat{\theta} - \theta_0)' \hat{V}_{\theta}^{-1} (\hat{\theta} - \theta_0)$$

using the asymptotic variance estimator  $\hat{V}_{\theta}$ , or we can write it directly as a function of  $\hat{\beta}$  as

$$W = (r(\hat{\beta}) - \theta_0)' (\hat{R}' \hat{V}_{\hat{\beta}} \hat{R})^{-1} (r(\hat{\beta}) - \theta_0).$$

Also, when  $r(\beta) = R'\beta$  is a linear function of  $\beta$ , then the Wald statistic simplifies to

$$W = (R'\hat{\beta} - \theta_0)' (R' \hat{V}_{\hat{\beta}} R)^{-1} (R'\hat{\beta} - \theta_0).$$

The Wald statistic  $W$  is a weighted Euclidean measure of the length of the vector  $\hat{\theta} - \theta_0$ . When  $q = 1$  then  $W = T^2$ , the square of the t-statistic, so hypothesis tests based on  $W$  and  $|T|$  are equivalent. The Wald statistic (9.6) is a generalization of the t-statistic to the case of multiple restrictions. As the Wald statistic is symmetric in the argument  $\hat{\theta} - \theta_0$  it treats positive and negative alternatives symmetrically. Thus the inherent alternative is always two-sided.

As shown in Theorem 7.13, when  $\beta$  satisfies  $r(\beta) = \theta_0$  then  $W \xrightarrow{d} \chi_q^2$ , a chi-square random variable with  $q$  degrees of freedom. Let  $G_q(u)$  denote the  $\chi_q^2$  distribution function. For a given significance level  $\alpha$ , the asymptotic critical value  $c$  satisfies  $\alpha = 1 - G_q(c)$ . For example, the 5% critical values for  $q = 1$ ,  $q = 2$ , and  $q = 3$  are 3.84, 5.99, and 7.82, respectively, and in general the level  $\alpha$  critical value can be calculated in MATLAB as `chi2inv(1-alpha, q)`. An asymptotic test rejects  $H_0$  in favor of  $H_1$  if  $W > c$ . As with t-tests, it is conventional to describe a Wald test as “significant” if  $W$  exceeds the 5% asymptotic critical value.

**Theorem 9.2** Under Assumptions 7.2, 7.3, 7.4, and  $H_0 : \theta = \theta_0$ , then

$$W \xrightarrow{d} \chi_q^2,$$

and for  $c$  satisfying  $\alpha = 1 - G_q(c)$ ,

$$\mathbb{P}(W > c | H_0) \longrightarrow \alpha$$

so the test “Reject  $H_0$  if  $W > c$ ” has asymptotic size  $\alpha$ .

Notice that the asymptotic distribution in Theorem 9.2 depends solely on  $q$ , the number of restrictions being tested. It does not depend on  $k$ , the number of parameters estimated.

The asymptotic p-value for  $W$  is  $p = 1 - G_q(W)$ , and this is particularly useful when testing multiple restrictions. For example, if you write that a Wald test on eight restrictions ( $q = 8$ ) has the value  $W = 11.2$ , it is difficult for a reader to assess the magnitude of this statistic unless they have quick access to a statistical table or software. Instead, if you write that the p-value is  $p = 0.19$  (as is the case for  $W = 11.2$  and  $q = 8$ ) then it is simple for a reader to interpret its magnitude as “insignificant”. To calculate the asymptotic p-value for a Wald statistic in MATLAB, use the command `1-chi2c df(w, q)`.

Some packages (including Stata) and papers report  $F$  versions of Wald statistics. That is, for any Wald statistic  $W$  which tests a  $q$ -dimensional restriction, the  $F$  version of the test is

$$F = W/q.$$

When  $F$  is reported, it is conventional to use  $F_{q,n-k}$  critical values and p-values rather than  $\chi^2_q$  values. The connection between Wald and F statistics is demonstrated in Section 9.14 we show that when Wald statistics are calculated using a homoskedastic covariance matrix, then  $F = W/q$  is identical to the F statistic of (5.22). While there is no formal justification to using the  $F_{q,n-k}$  distribution for non-homoskedastic covariance matrices, the  $F_{q,n-k}$  distribution provides continuity with the exact distribution theory under normality and is a bit more conservative than the  $\chi^2_q$  distribution. (Furthermore, the difference is small when  $n - k$  is moderately large.)

To implement a test of zero restrictions in Stata, an easy method is to use the command “test X1 X2” where X1 and X2 are the names of the variables whose coefficients are hypothesized to equal zero. This command should be executed after executing a regression command. The  $F$  version of the Wald statistic is reported, using the covariance matrix calculated using the method specified in the regression command. A p-value is reported, calculated using the  $F_{q,n-k}$  distribution.

To illustrate, consider the empirical results presented in Table 4.1. The hypothesis “Union membership does not affect wages” is the joint restriction that both coefficients on “Male Union Member” and “Female Union Member” are zero. We calculate the Wald statistic for this joint hypothesis and find  $W = 23$  (or  $F = 12.5$ ) with a p-value of  $p = 0.000$ . Thus we reject the null hypothesis in favor of the alternative that at least one of the coefficients is non-zero. This does not mean that both coefficients are non-zero, just that one of the two is non-zero. Therefore examining both the joint Wald statistic and the individual t-statistics is useful for interpretation.

As a second example from the same regression, take the hypothesis that married status has no effect on mean wages for women. This is the joint restriction that the coefficients on “Married Female” and “Formerly Married Female” are zero. The Wald statistic for this hypothesis is  $W = 6.4$  ( $F = 3.2$ ) with a p-value of 0.04. Such a p-value is typically called “marginally significant”, in the sense that it is slightly smaller than 0.05.

The Wald statistic was proposed by Wald (1943).

### Abraham Wald

The Hungarian mathematician/statistician/econometrician Abraham Wald (1902-1950) developed an optimality property for the Wald test in terms of weighted average power. He also developed the field of sequential testing and the design of experiments.

## 9.10 Homoskedastic Wald Tests

If the error is known to be homoskedastic, then it is appropriate to use the homoskedastic Wald statistic (7.38) which replaces  $\widehat{V}_{\hat{\theta}}$  with the homoskedastic estimator  $\widehat{V}_{\hat{\theta}}^0$ . This statistic equals

$$\begin{aligned} W^0 &= (\hat{\theta} - \theta_0)' (\widehat{V}_{\hat{\theta}}^0)^{-1} (\hat{\theta} - \theta_0) \\ &= (\mathbf{r}(\hat{\beta}) - \theta_0)' (\mathbf{R}' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{R})^{-1} (\mathbf{r}(\hat{\beta}) - \theta_0) / s^2. \end{aligned}$$

In the case of linear hypotheses  $\mathbb{H}_0 : \mathbf{R}' \beta = \theta_0$  we can write this as

$$W^0 = (\mathbf{R}' \hat{\beta} - \theta_0)' (\mathbf{R}' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{R})^{-1} (\mathbf{R}' \hat{\beta} - \theta_0) / s^2. \quad (9.7)$$

We call either a **homoskedastic Wald statistic** as it is an appropriate test when the errors are conditionally homoskedastic.

As for  $W$ , when  $q = 1$  then  $W^0 = T^2$ , the square of the t-statistic where the latter is computed with a homoskedastic standard error.

**Theorem 9.3** Under Assumptions 7.2 and 7.3,  $\mathbb{E}(e_i^2 | \mathbf{x}_i) = \sigma^2 > 0$ , and  $\mathbb{H}_0 : \theta = \theta_0$ , then

$$W^0 \xrightarrow{d} \chi_q^2,$$

and for  $c$  satisfying  $\alpha = 1 - G_q(c)$ ,

$$\mathbb{P}(W^0 > c | \mathbb{H}_0) \longrightarrow \alpha$$

so the test “Reject  $\mathbb{H}_0$  if  $W^0 > c$ ” has asymptotic size  $\alpha$ .

## 9.11 Criterion-Based Tests

The Wald statistic is based on the length of the vector  $\hat{\theta} - \theta_0$ : the discrepancy between the estimate  $\hat{\theta} = \mathbf{r}(\hat{\beta})$  and the hypothesized value  $\theta_0$ . An alternative class of tests is based on the discrepancy between the criterion function minimized with and without the restriction.

Criterion-based testing applies when we have a criterion function, say  $J(\beta)$  with  $\beta \in \mathcal{B}$ , which is minimized for estimation, and the goal is to test  $\mathbb{H}_0 : \beta \in \mathcal{B}_0$  versus  $\mathbb{H}_1 : \beta \notin \mathcal{B}_0$  where  $\mathcal{B}_0 \subset \mathcal{B}$ . Minimizing the criterion function over  $\mathcal{B}$  and  $\mathcal{B}_0$  we obtain the unrestricted and restricted estimators

$$\begin{aligned} \hat{\beta} &= \underset{\beta \in \mathcal{B}}{\text{argmin}} J(\beta) \\ \tilde{\beta} &= \underset{\beta \in \mathcal{B}_0}{\text{argmin}} J(\beta). \end{aligned}$$

The **criterion-based statistic** for  $\mathbb{H}_0$  versus  $\mathbb{H}_1$  is proportional to

$$\begin{aligned} J &= \min_{\beta \in \mathcal{B}_0} J(\beta) - \min_{\beta \in \mathcal{B}} J(\beta) \\ &= J(\tilde{\beta}) - J(\hat{\beta}). \end{aligned}$$

The criterion-based statistic  $J$  is sometimes called a **distance** statistic, a **minimum-distance** statistic, or a **likelihood-ratio-like** statistic.

Since  $\mathcal{B}_0$  is a subset of  $\mathcal{B}$ ,  $J(\tilde{\beta}) \geq J(\hat{\beta})$  and thus  $J \geq 0$ . The statistic  $J$  measures the cost (on the criterion) of imposing the null restriction  $\beta \in \mathcal{B}_0$ .

## 9.12 Minimum Distance Tests

The minimum distance test is a criterion-based test where  $J(\boldsymbol{\beta})$  is the minimum distance criterion (8.19)

$$J(\boldsymbol{\beta}) = n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \widehat{\mathbf{W}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \quad (9.8)$$

with  $\hat{\boldsymbol{\beta}}$  the unrestricted (LS) estimator. The restricted estimator  $\tilde{\boldsymbol{\beta}}_{\text{md}}$  minimizes (9.8) subject to  $\boldsymbol{\beta} \in \mathbf{B}_0$ . Observing that  $J(\hat{\boldsymbol{\beta}}) = 0$ , the minimum distance statistic simplifies to

$$J = J(\tilde{\boldsymbol{\beta}}_{\text{md}}) = n(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\text{md}})' \widehat{\mathbf{W}}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\text{md}}). \quad (9.9)$$

The efficient minimum distance estimator  $\tilde{\boldsymbol{\beta}}_{\text{emd}}$  is obtained by setting  $\widehat{\mathbf{W}} = \widehat{\mathbf{V}}_{\boldsymbol{\beta}}^{-1}$  in (9.8) and (9.9). The efficient minimum distance statistic for  $\mathbb{H}_0 : \boldsymbol{\beta} \in \mathbf{B}_0$  is therefore

$$J^* = n(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\text{emd}})' \widehat{\mathbf{V}}_{\boldsymbol{\beta}}^{-1}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\text{emd}}). \quad (9.10)$$

Consider the class of linear hypotheses  $\mathbb{H}_0 : \mathbf{R}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$ . In this case we know from (8.25) that the efficient minimum distance estimator  $\tilde{\boldsymbol{\beta}}_{\text{emd}}$  subject to the constraint  $\mathbf{R}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$  is

$$\tilde{\boldsymbol{\beta}}_{\text{emd}} = \hat{\boldsymbol{\beta}} - \widehat{\mathbf{V}}_{\boldsymbol{\beta}} \mathbf{R} (\mathbf{R}' \widehat{\mathbf{V}}_{\boldsymbol{\beta}} \mathbf{R})^{-1} (\mathbf{R}' \hat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0)$$

and thus

$$\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\text{emd}} = \widehat{\mathbf{V}}_{\boldsymbol{\beta}} \mathbf{R} (\mathbf{R}' \widehat{\mathbf{V}}_{\boldsymbol{\beta}} \mathbf{R})^{-1} (\mathbf{R}' \hat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0).$$

Substituting into (9.10) we find

$$\begin{aligned} J^* &= n(\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0)' (\mathbf{R}' \widehat{\mathbf{V}}_{\boldsymbol{\beta}} \mathbf{R})^{-1} \mathbf{R}' \widehat{\mathbf{V}}_{\boldsymbol{\beta}} \widehat{\mathbf{V}}_{\boldsymbol{\beta}}^{-1} \widehat{\mathbf{V}}_{\boldsymbol{\beta}} \mathbf{R} (\mathbf{R}' \widehat{\mathbf{V}}_{\boldsymbol{\beta}} \mathbf{R})^{-1} (\mathbf{R}' \hat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0) \\ &= n(\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0)' (\mathbf{R}' \widehat{\mathbf{V}}_{\boldsymbol{\beta}} \mathbf{R})^{-1} (\mathbf{R}' \hat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0) \\ &= W, \end{aligned}$$

which is the Wald statistic (9.6).

Thus for linear hypotheses  $\mathbb{H}_0 : \mathbf{R}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$ , the efficient minimum distance statistic  $J^*$  is identical to the Wald statistic (9.6). For non-linear hypotheses, however, the Wald and minimum distance statistics are different.

Newey and West (1987a) established the asymptotic null distribution of  $J^*$  for linear and non-linear hypotheses.

**Theorem 9.4** Under Assumptions 7.2, 7.3, 7.4, and  $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ , then  $J^* \xrightarrow{d} \chi_q^2$ .

Testing using the minimum distance statistic  $J^*$  is similar to testing using the Wald statistic  $W$ . Critical values and p-values are computed using the  $\chi_q^2$  distribution.  $\mathbb{H}_0$  is rejected in favor of  $\mathbb{H}_1$  if  $J^*$  exceeds the level  $\alpha$  critical value, which can be calculated in MATLAB as `chi2inv(1-alpha, q)`. The asymptotic p-value is  $p = 1 - G_q(J^*)$ . In MATLAB, use the command `1-chi2cdf(J, q)`.

We now demonstrate Theorem 9.4. The conditions of Theorem 8.10 hold, since  $\mathbb{H}_0$  implies Assumption 8.1. From (8.54) with  $\widehat{\mathbf{W}} = \widehat{\mathbf{V}}_{\boldsymbol{\beta}}$ , we see that

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\text{emd}}) &= \widehat{\mathbf{V}}_{\boldsymbol{\beta}} \widehat{\mathbf{R}} (\mathbf{R}_n^{*\prime} \widehat{\mathbf{V}}_{\boldsymbol{\beta}} \widehat{\mathbf{R}})^{-1} \mathbf{R}_n^{*\prime} \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\xrightarrow{d} \mathbf{V}_{\boldsymbol{\beta}} \mathbf{R} (\mathbf{R}' \mathbf{V}_{\boldsymbol{\beta}} \mathbf{R})^{-1} \mathbf{R}' \mathbf{N}(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}}) \\ &= \mathbf{V}_{\boldsymbol{\beta}} \mathbf{R} \mathbf{Z} \end{aligned}$$

where  $Z \sim N(\mathbf{0}, (\mathbf{R}' V_{\beta} \mathbf{R})^{-1})$ . Thus

$$\begin{aligned} J^* &= n(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}})' \widehat{V}_{\beta}^{-1} (\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}) \\ &\xrightarrow{d} Z' \mathbf{R}' V_{\beta} V_{\beta}^{-1} V_{\beta} \mathbf{R} Z \\ &= Z' (\mathbf{R}' V_{\beta} \mathbf{R}) Z = \chi_q^2 \end{aligned}$$

as claimed.

### 9.13 Minimum Distance Tests Under Homoskedasticity

If we set  $\widehat{W} = \widehat{\mathbf{Q}}_{xx}/s^2$  in (9.8) we obtain the criterion (8.20)

$$J^0(\boldsymbol{\beta}) = n(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \widehat{\mathbf{Q}}_{xx} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) / s^2.$$

A minimum distance statistic for  $H_0 : \boldsymbol{\beta} \in \mathcal{B}_0$  is

$$J^0 = \min_{\boldsymbol{\beta} \in \mathcal{B}_0} J^0(\boldsymbol{\beta}).$$

Equation (8.21) showed that

$$\text{SSE}(\boldsymbol{\beta}) = n\widehat{\sigma}^2 + s^2 J^0(\boldsymbol{\beta})$$

and so the minimizers of  $\text{SSE}(\boldsymbol{\beta})$  and  $J^0(\boldsymbol{\beta})$  are identical. Thus the constrained minimizer of  $J^0(\boldsymbol{\beta})$  is constrained least-squares

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \underset{\boldsymbol{\beta} \in \mathcal{B}_0}{\operatorname{argmin}} J^0(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta} \in \mathcal{B}_0}{\operatorname{argmin}} \text{SSE}(\boldsymbol{\beta}) \quad (9.11)$$

and therefore

$$\begin{aligned} J_n^0 &= J_n^0(\widetilde{\boldsymbol{\beta}}_{\text{cls}}) \\ &= n(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{cls}})' \widehat{\mathbf{Q}}_{xx} (\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{cls}}) / s^2. \end{aligned}$$

In the special case of linear hypotheses  $H_0 : \mathbf{R}' \boldsymbol{\beta} = \boldsymbol{\theta}_0$ , the constrained least-squares estimator subject to  $\mathbf{R}' \boldsymbol{\beta} = \boldsymbol{\theta}_0$  has the solution (8.9)

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \widehat{\boldsymbol{\beta}} - \widehat{\mathbf{Q}}_{xx}^{-1} \mathbf{R} \left( \mathbf{R}' \widehat{\mathbf{Q}}_{xx}^{-1} \mathbf{R} \right)^{-1} (\mathbf{R}' \widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0)$$

and solving we find

$$J^0 = n(\mathbf{R}' \widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0)' \left( \mathbf{R}' \widehat{\mathbf{Q}}_{xx}^{-1} \mathbf{R} \right)^{-1} (\mathbf{R}' \widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0) / s^2 = W^0.$$

This is the homoskedastic Wald statistic (9.7). Thus for testing linear hypotheses, homoskedastic minimum distance and Wald statistics agree.

For nonlinear hypotheses they disagree, but have the same null asymptotic distribution.

**Theorem 9.5** Under Assumptions 7.2 and 7.3,  $\mathbb{E}(e_i^2 | \mathbf{x}_i) = \sigma^2 > 0$ , and  $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ , then  $J^0 \xrightarrow{d} \chi_q^2$ .

## 9.14 F Tests

In Section 5.16 we introduced the  $F$  test for exclusion restrictions in the normal regression model. More generally, the  $F$  statistic for testing  $\mathbb{H}_0 : \boldsymbol{\beta} \in \mathbf{B}_0$  is

$$F = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2) / q}{\hat{\sigma}^2 / (n - k)} \quad (9.12)$$

where

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}})^2$$

and  $\hat{\boldsymbol{\beta}}$  are the unconstrained estimators of  $\boldsymbol{\beta}$  and  $\sigma^2$ ,

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}'_i \tilde{\boldsymbol{\beta}}_{\text{cls}})^2$$

and  $\tilde{\boldsymbol{\beta}}_{\text{cls}}$  are the constrained least-squares estimators from (9.11),  $q$  is the number of restrictions, and  $k$  is the number of unconstrained coefficients.

We can alternatively write

$$F = \frac{\text{SSE}(\tilde{\boldsymbol{\beta}}_{\text{cls}}) - \text{SSE}(\hat{\boldsymbol{\beta}})}{qs^2} \quad (9.13)$$

where

$$\text{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2$$

is the sum-of-squared errors. Thus  $F$  is a criterion-based statistic. Using (8.21) we can also write

$$F = J^0 / q,$$

so the  $F$  statistic is identical to the homoskedastic minimum distance statistic divided by the number of restrictions  $q$ .

As we discussed in the previous section, in the special case of linear hypotheses  $\mathbb{H}_0 : \mathbf{R}' \boldsymbol{\beta} = \boldsymbol{\theta}_0$ ,  $J^0 = W^0$ . It follows that in this case  $F = W^0 / q$ . Thus for linear restrictions the  $F$  statistic equals the homoskedastic Wald statistic divided by  $q$ . It follows that they are equivalent tests for  $\mathbb{H}_0$  against  $\mathbb{H}_1$ .

**Theorem 9.6** For tests of linear hypotheses  $\mathbb{H}_0 : \mathbf{R}' \boldsymbol{\beta} = \boldsymbol{\theta}_0$ ,

$$F = W^0 / q$$

the  $F$  statistic equals the homoskedastic Wald statistic divided by the degrees of freedom. Thus under 7.2,  $\mathbb{E}(e_i^2 | \mathbf{x}_i) = \sigma^2 > 0$ , and  $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ , then

$$F \xrightarrow{d} \chi_q^2 / q.$$

When using an  $F$  statistic, it is conventional to use the  $F_{q,n-k}$  distribution for critical values and p-values. Critical values are given in MATLAB by `finv(1-alpha, q, n-k)`, and p-values by `1-fcdf(F, q, n-k)`. Alternatively, the  $\chi_q^2 / q$  distribution can be used, using `chi2inv(1-alpha, q) / q` and `1-chi2cdf(F*q, q)`, respectively. Using the  $F_{q,n-k}$  distribution is a prudent small sample adjustment which yields exact answers if the errors are normal, and otherwise slightly increasing the critical values and p-values relative to the asymptotic approximation. Once again, if the sample size is small enough that the choice makes a difference, then probably we shouldn't be trusting the asymptotic approximation anyway!

An elegant feature about (9.12) or (9.13) is that they are directly computable from the standard output from two simple OLS regressions, as the sum of squared errors (or regression variance) is a typical printed output from statistical packages, and is often reported in applied tables. Thus  $F$  can be calculated by hand from standard reported statistics even if you don't have the original data (or if you are sitting in a seminar and listening to a presentation!).

If you are presented with an  $F$  statistic (or a Wald statistic, as you can just divide by  $q$ ) but don't have access to critical values, a useful rule of thumb is to know that for large  $n$ , the 5% asymptotic critical value is decreasing as  $q$  increases, and is less than 2 for  $q \geq 7$ .

A word of warning: In many statistical packages, when an OLS regression is estimated an " $F$ -statistic" is automatically reported, even though no hypothesis test was requested. What the package is reporting is an  $F$  statistic of the hypothesis that all slope coefficients<sup>1</sup> are zero. This was a popular statistic in the early days of econometric reporting when sample sizes were very small and researchers wanted to know if there was "any explanatory power" to their regression. This is rarely an issue today, as sample sizes are typically sufficiently large that this  $F$  statistic is nearly always highly significant. While there are special cases where this  $F$  statistic is useful, these cases are not typical. As a general rule, there is no reason to report this  $F$  statistic.

## 9.15 Hausman Tests

Hausman (1978) introduced a general idea about how to test a hypothesis  $\mathbb{H}_0$ . If you have two estimators, one which is efficient under  $\mathbb{H}_0$  but inconsistent under  $\mathbb{H}_1$ , and another which is consistent under  $\mathbb{H}_1$ , then construct a test as a quadratic form in the differences of the estimators. In the case of testing a hypothesis  $\mathbb{H}_0 : \mathbf{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$  let  $\hat{\boldsymbol{\beta}}_{\text{ols}}$  denote the unconstrained least-squares estimator and let  $\tilde{\boldsymbol{\beta}}_{\text{emd}}$  denote the efficient minimum distance estimator which imposes  $\mathbf{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$ . Both estimators are consistent under  $\mathbb{H}_0$ , but  $\tilde{\boldsymbol{\beta}}_{\text{emd}}$  is asymptotically efficient. Under  $\mathbb{H}_1$ ,  $\hat{\boldsymbol{\beta}}_{\text{ols}}$  is consistent for  $\boldsymbol{\beta}$  but  $\tilde{\boldsymbol{\beta}}_{\text{emd}}$  is inconsistent. The difference has the asymptotic distribution

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{ols}} - \tilde{\boldsymbol{\beta}}_{\text{emd}}) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}} \mathbf{R} (\mathbf{R}' \mathbf{V}_{\boldsymbol{\beta}} \mathbf{R})^{-1} \mathbf{R}' \mathbf{V}_{\boldsymbol{\beta}}\right).$$

Let  $\mathbf{A}^-$  denote the Moore-Penrose generalized inverse. The Hausman statistic for  $\mathbb{H}_0$  is

$$\begin{aligned} H &= (\hat{\boldsymbol{\beta}}_{\text{ols}} - \tilde{\boldsymbol{\beta}}_{\text{emd}})' \widehat{\text{avar}}(\hat{\boldsymbol{\beta}}_{\text{ols}} - \tilde{\boldsymbol{\beta}}_{\text{emd}})^- (\hat{\boldsymbol{\beta}}_{\text{ols}} - \tilde{\boldsymbol{\beta}}_{\text{emd}}) \\ &= n(\hat{\boldsymbol{\beta}}_{\text{ols}} - \tilde{\boldsymbol{\beta}}_{\text{emd}})' \left( \hat{\mathbf{V}}_{\boldsymbol{\beta}} \hat{\mathbf{R}} (\hat{\mathbf{R}}' \hat{\mathbf{V}}_{\boldsymbol{\beta}} \hat{\mathbf{R}})^{-1} \hat{\mathbf{R}}' \hat{\mathbf{V}}_{\boldsymbol{\beta}} \right)^- (\hat{\boldsymbol{\beta}}_{\text{ols}} - \tilde{\boldsymbol{\beta}}_{\text{emd}}). \end{aligned}$$

The matrix  $\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{1/2} \hat{\mathbf{R}} (\hat{\mathbf{R}}' \hat{\mathbf{V}}_{\boldsymbol{\beta}} \hat{\mathbf{R}})^{-1} \hat{\mathbf{R}}' \hat{\mathbf{V}}_{\boldsymbol{\beta}}^{1/2}$  idempotent so its generalized inverse is itself. (See Section A.11.) It follows that

$$\begin{aligned} \left( \hat{\mathbf{V}}_{\boldsymbol{\beta}} \hat{\mathbf{R}} (\hat{\mathbf{R}}' \hat{\mathbf{V}}_{\boldsymbol{\beta}} \hat{\mathbf{R}})^{-1} \hat{\mathbf{R}}' \hat{\mathbf{V}}_{\boldsymbol{\beta}} \right)^- &= \hat{\mathbf{V}}_{\boldsymbol{\beta}}^{-1/2} \left( \hat{\mathbf{V}}_{\boldsymbol{\beta}}^{1/2} \hat{\mathbf{R}} (\hat{\mathbf{R}}' \hat{\mathbf{V}}_{\boldsymbol{\beta}} \hat{\mathbf{R}})^{-1} \hat{\mathbf{R}}' \hat{\mathbf{V}}_{\boldsymbol{\beta}}^{1/2} \right)^- \hat{\mathbf{V}}_{\boldsymbol{\beta}}^{-1/2} \\ &= \hat{\mathbf{V}}_{\boldsymbol{\beta}}^{-1/2} \hat{\mathbf{V}}_{\boldsymbol{\beta}}^{1/2} \hat{\mathbf{R}} (\hat{\mathbf{R}}' \hat{\mathbf{V}}_{\boldsymbol{\beta}} \hat{\mathbf{R}})^{-1} \hat{\mathbf{R}}' \hat{\mathbf{V}}_{\boldsymbol{\beta}}^{1/2} \hat{\mathbf{V}}_{\boldsymbol{\beta}}^{-1/2} \\ &= \hat{\mathbf{R}} (\hat{\mathbf{R}}' \hat{\mathbf{V}}_{\boldsymbol{\beta}} \hat{\mathbf{R}})^{-1} \hat{\mathbf{R}}'. \end{aligned}$$

Thus the Hausman statistic is

$$H = n(\hat{\boldsymbol{\beta}}_{\text{ols}} - \tilde{\boldsymbol{\beta}}_{\text{emd}})' \hat{\mathbf{R}} (\hat{\mathbf{R}}' \hat{\mathbf{V}}_{\boldsymbol{\beta}} \hat{\mathbf{R}})^{-1} \hat{\mathbf{R}}' (\hat{\boldsymbol{\beta}}_{\text{ols}} - \tilde{\boldsymbol{\beta}}_{\text{emd}}).$$

In the context of linear restrictions,  $\hat{\mathbf{R}} = \mathbf{R}$  and  $\mathbf{R}' \tilde{\boldsymbol{\beta}} = \boldsymbol{\theta}_0$  so the statistic takes the form

$$H = n(\mathbf{R}' \hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\theta}_0)' \hat{\mathbf{R}} (\mathbf{R}' \hat{\mathbf{V}}_{\boldsymbol{\beta}} \mathbf{R})^{-1} (\mathbf{R}' \hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\theta}_0),$$

---

<sup>1</sup>All coefficients except the intercept.

which is precisely the Wald statistic. With nonlinear restrictions  $W$  and  $H$  can differ.

In either case we see that that the asymptotic null distribution of the Hausman statistic  $H$  is  $\chi_q^2$ , so the appropriate test is to reject  $\mathbb{H}_0$  in favor of  $\mathbb{H}_1$  if  $H > c$  where  $c$  is a critical value taken from the  $\chi_q^2$  distribution.

**Theorem 9.7** For general hypotheses the Hausman test statistic is

$$H = n(\hat{\beta}_{\text{ols}} - \tilde{\beta}_{\text{emd}})' \hat{\mathbf{R}} (\hat{\mathbf{R}}' \hat{\mathbf{V}}_{\beta} \hat{\mathbf{R}})^{-1} \hat{\mathbf{R}}' (\hat{\beta}_{\text{ols}} - \tilde{\beta}_{\text{emd}}).$$

Under Assumptions 7.2, 7.3, 7.4, and  $\mathbb{H}_0 : \mathbf{r}(\beta) = \theta_0$ ,

$$H \xrightarrow{d} \chi_q^2.$$

## 9.16 Score Tests

Score tests are traditionally derived in likelihood analysis, but can more generally be constructed from first-order conditions evaluated at restricted estimates. We focus on the likelihood derivation.

Given the log likelihood function  $\log L(\beta, \sigma^2)$ , a restriction  $\mathbb{H}_0 : \mathbf{r}(\beta) = \theta_0$ , and restricted estimators  $\tilde{\beta}$  and  $\tilde{\sigma}^2$ , the **score statistic** for  $\mathbb{H}_0$  is defined as

$$S = \left( \frac{\partial}{\partial \beta} \log L(\tilde{\beta}, \tilde{\sigma}^2) \right)' \left( -\frac{\partial^2}{\partial \beta \partial \beta'} \log L(\tilde{\beta}, \tilde{\sigma}^2) \right)^{-1} \left( \frac{\partial}{\partial \beta} \log L(\tilde{\beta}, \tilde{\sigma}^2) \right).$$

The idea is that if the restriction is true, then the restricted estimators should be close to the maximum of the log-likelihood where the derivative should be small. However if the restriction is false then the restricted estimators should be distant from the maximum and the derivative should be large. Hence small values of  $S$  are expected under  $\mathbb{H}_0$  and large values under  $\mathbb{H}_1$ . Tests of  $\mathbb{H}_0$  thus reject for large values of  $S$ .

We explore the score statistic in the context of the normal regression model and linear hypotheses  $\mathbf{r}(\beta) = \mathbf{R}'\beta$ . Recall that in the normal regression log-likelihood function is

$$\log L(\beta, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}'_i \beta)^2.$$

The constrained MLE under linear hypotheses is constrained least squares

$$\begin{aligned} \tilde{\beta} &= \hat{\beta} - (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R} \left[ \mathbf{R}' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R} \right]^{-1} (\mathbf{R}' \hat{\beta} - \mathbf{c}) \\ \tilde{e}_i &= y_i - \mathbf{x}'_i \tilde{\beta} \\ \tilde{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n \tilde{e}_i^2. \end{aligned}$$

We can calculate that the derivative and Hessian are

$$\begin{aligned} \frac{\partial}{\partial \beta} \log L(\tilde{\beta}, \tilde{\sigma}^2) &= \frac{1}{\tilde{\sigma}^2} \sum_{i=1}^n \mathbf{x}_i (y_i - \mathbf{x}'_i \tilde{\beta}) = \frac{1}{\tilde{\sigma}^2} \mathbf{X}' \tilde{\mathbf{e}} \\ -\frac{\partial^2}{\partial \beta \partial \beta'} \log L(\tilde{\beta}, \tilde{\sigma}^2) &= \frac{1}{\tilde{\sigma}^2} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i = \frac{1}{\tilde{\sigma}^2} \mathbf{X}' \mathbf{X}. \end{aligned}$$

Since  $\tilde{\mathbf{e}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$  we can further calculate that

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}} \log L(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2) &= \frac{1}{\tilde{\sigma}^2} (\mathbf{X}' \mathbf{X}) \left( (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} - (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{X} \tilde{\boldsymbol{\beta}} \right) \\ &= \frac{1}{\tilde{\sigma}^2} (\mathbf{X}' \mathbf{X}) (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) \\ &= \frac{1}{\tilde{\sigma}^2} \mathbf{R} \left[ \mathbf{R}' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{R} \right]^{-1} (\mathbf{R}' \hat{\boldsymbol{\beta}} - \mathbf{c}).\end{aligned}$$

Together we find that

$$S = (\mathbf{R}' \hat{\boldsymbol{\beta}} - \mathbf{c})' \left( \mathbf{R}' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{R} \right)^{-1} (\mathbf{R}' \hat{\boldsymbol{\beta}} - \mathbf{c}) / \tilde{\sigma}^2.$$

This is identical to the homoskedastic Wald statistic, with  $s^2$  replaced by  $\tilde{\sigma}^2$ . We can also write  $S$  as a monotonic transformation of the  $F$  statistic, since

$$S = n \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)}{\hat{\sigma}^2} = n \left( 1 - \frac{\hat{\sigma}^2}{\tilde{\sigma}^2} \right) = n \left( 1 - \frac{1}{1 + \frac{q}{n-k} F} \right).$$

The test “Reject  $H_0$  for large values of  $S$ ” is identical to the test “Reject  $H_0$  for large values of  $F$ ”, so they are identical tests. Since for the normal regression model the exact distribution of  $F$  is known, it is better to use the  $F$  statistic with  $F$  p-values.

In more complicated settings a potential advantage of score tests is that they are calculated using the restricted parameter estimates  $\tilde{\boldsymbol{\beta}}$  rather than the unrestricted estimates  $\hat{\boldsymbol{\beta}}$ . Thus when  $\tilde{\boldsymbol{\beta}}$  is relatively easy to calculate there can be a preference for score statistics. This is not a concern for linear restrictions.

More generally, score and score-like statistics can be constructed from first-order conditions evaluated at restricted parameter estimates. Also, when test statistics are constructed using covariance matrix estimators which are calculated using restricted parameter estimates (e.g. restricted residuals) then these are often described as score tests.

An example of the latter is the Wald-type statistic

$$W = (\mathbf{r}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0)' \left( \hat{\mathbf{R}}' \tilde{\mathbf{V}}_{\hat{\boldsymbol{\beta}}} \hat{\mathbf{R}} \right)^{-1} (\mathbf{r}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0)$$

where the covariance matrix estimate  $\tilde{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$  is calculated using the restricted residuals  $\tilde{e}_i = y_i - \mathbf{x}'_i \tilde{\boldsymbol{\beta}}$ . This may be done when  $\boldsymbol{\beta}$  and  $\boldsymbol{\theta}$  are high-dimensional, so there is worry that the estimator  $\tilde{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$  is imprecise.

## 9.17 Problems with Tests of Nonlinear Hypotheses

While the  $t$  and Wald tests work well when the hypothesis is a linear restriction on  $\boldsymbol{\beta}$ , they can work quite poorly when the restrictions are nonlinear. This can be seen by a simple example introduced by Lafontaine and White (1986). Take the model

$$\begin{aligned}y_i &= \beta + e_i \\ e_i &\sim N(0, \sigma^2)\end{aligned}$$

and consider the hypothesis

$$H_0 : \beta = 1.$$

Let  $\hat{\beta}$  and  $\hat{\sigma}^2$  be the sample mean and variance of  $y_i$ . The standard Wald test for  $H_0$  is

$$W = n \frac{(\hat{\beta} - 1)^2}{\hat{\sigma}^2}.$$

Now notice that  $H_0$  is equivalent to the hypothesis

$$H_0(s) : \beta^s = 1$$

for any positive integer  $s$ . Letting  $r(\beta) = \beta^s$ , and noting  $\mathbf{R} = s\beta^{s-1}$ , we find that the standard Wald test for  $\mathbb{H}_0(s)$  is

$$W(s) = n \frac{(\hat{\beta}^s - 1)^2}{\hat{\sigma}^2 s^2 \hat{\beta}^{2s-2}}.$$

While the hypothesis  $\beta^s = 1$  is unaffected by the choice of  $s$ , the statistic  $W(s)$  varies with  $s$ . This is an unfortunate feature of the Wald statistic.

To demonstrate this effect, we have plotted in Figure 9.3 the Wald statistic  $W(s)$  as a function of  $s$ , setting  $n/\hat{\sigma}^2 = 10$ . The increasing solid line is for the case  $\hat{\beta} = 0.8$ . The decreasing dashed line is for the case  $\hat{\beta} = 1.6$ . It is easy to see that in each case there are values of  $s$  for which the test statistic is significant relative to asymptotic critical values, while there are other values of  $s$  for which the test statistic is insignificant. This is distressing since the choice of  $s$  is arbitrary and irrelevant to the actual hypothesis.

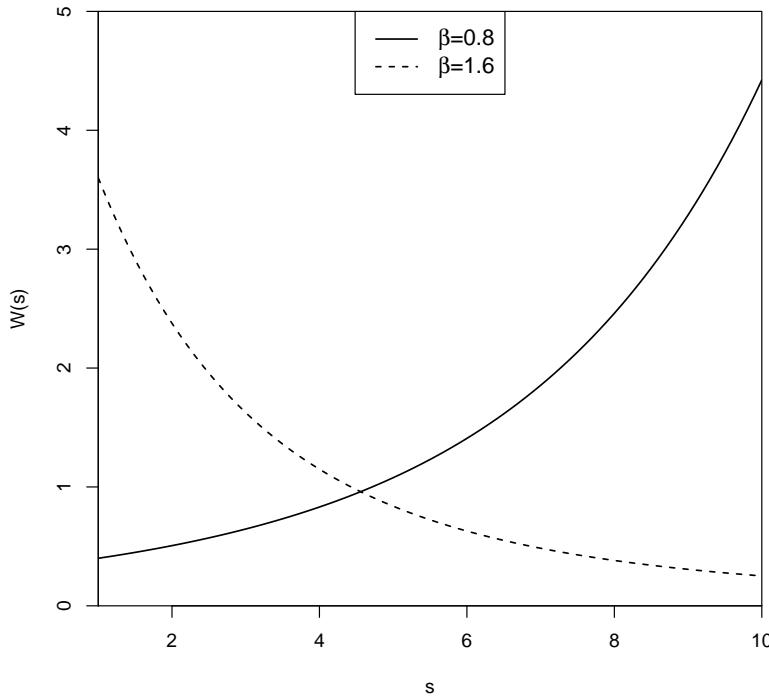


Figure 9.3: Wald Statistic as a Function of  $s$

Our first-order asymptotic theory is not useful to help pick  $s$ , as  $W(s) \xrightarrow{d} \chi_1^2$  under  $\mathbb{H}_0$  for any  $s$ . This is a context where **Monte Carlo simulation** can be quite useful as a tool to study and compare the exact distributions of statistical procedures in finite samples. The method uses random simulation to create artificial datasets, to which we apply the statistical tools of interest. This produces random draws from the statistic's sampling distribution. Through repetition, features of this distribution can be calculated.

In the present context of the Wald statistic, one feature of importance is the Type I error of the test using the asymptotic 5% critical value 3.84 – the probability of a false rejection,  $\mathbb{P}(W(s) > 3.84 | \beta = 1)$ . Given the simplicity of the model, this probability depends only on  $s$ ,  $n$ , and  $\sigma^2$ . In Table 9.2 we report the results of a Monte Carlo simulation where we vary these three parameters. The value of  $s$  is varied from 1 to 10,  $n$  is varied among 20, 100 and 500, and  $\sigma$  is varied among 1 and 3. The Table reports the simulation estimate of the Type I error probability from 50,000 random samples. Each row of the table corresponds to a different value of  $s$  – and thus corresponds to a particular choice of test statistic. The second through seventh columns contain the Type I error probabilities for different combinations of  $n$  and  $\sigma$ . These probabilities are calculated as the percentage of the 50,000 simulated Wald statistics  $W(s)$

which are larger than 3.84. The null hypothesis  $\beta^s = 1$  is true, so these probabilities are Type I error.

To interpret the table, remember that the ideal Type I error probability is 5% (.05) with deviations indicating distortion. Type I error rates between 3% and 8% are considered reasonable. Error rates above 10% are considered excessive. Rates above 20% are unacceptable. When comparing statistical procedures, we compare the rates row by row, looking for tests for which rejection rates are close to 5% and rarely fall outside of the 3%-8% range. For this particular example the only test which meets this criterion is the conventional  $W = W(1)$  test. Any other choice of  $s$  leads to a test with unacceptable Type I error probabilities.

Table 9.2: Type I Error Probability of Asymptotic 5%  $W(s)$  Test

$s$	$\sigma = 1$			$\sigma = 3$		
	$n = 20$	$n = 100$	$n = 500$	$n = 20$	$n = 100$	$n = 500$
1	0.05	0.05	0.05	0.05	0.05	0.05
2	0.07	0.06	0.05	0.14	0.08	0.06
3	0.09	0.06	0.05	0.21	0.12	0.07
4	0.12	0.07	0.05	0.25	0.15	0.08
5	0.14	0.08	0.06	0.27	0.18	0.10
6	0.16	0.09	0.06	0.30	0.20	0.12
7	0.18	0.10	0.06	0.32	0.22	0.13
8	0.20	0.12	0.07	0.33	0.24	0.14
9	0.21	0.13	0.07	0.34	0.25	0.16
10	0.23	0.14	0.08	0.35	0.26	0.17

Rejection frequencies from 50,000 simulated random samples.

In Table 9.2 you can also see the impact of variation in sample size. In each case, the Type I error probability improves towards 5% as the sample size  $n$  increases. There is, however, no magic choice of  $n$  for which all tests perform uniformly well. Test performance deteriorates as  $s$  increases, which is not surprising given the dependence of  $W(s)$  on  $s$  as shown in Figure 9.3.

In this example it is not surprising that the choice  $s = 1$  yields the best test statistic. Other choices are arbitrary and would not be used in practice. While this is clear in this particular example, in other examples natural choices are not always obvious and the best choices may in fact appear counter-intuitive at first.

This point can be illustrated through another example which is similar to one developed in Gregory and Veall (1985). Take the model

$$\begin{aligned} y_i &= \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + e_i \\ \mathbb{E}(x_i e_i) &= \mathbf{0} \end{aligned} \tag{9.14}$$

and the hypothesis

$$\mathbb{H}_0: \frac{\beta_1}{\beta_2} = \theta_0$$

where  $\theta_0$  is a known constant. Equivalently, define  $\theta = \beta_1/\beta_2$ , so the hypothesis can be stated as  $\mathbb{H}_0: \theta = \theta_0$ .

Let  $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$  be the least-squares estimator of (9.14), let  $\hat{V}_{\hat{\boldsymbol{\beta}}}$  be an estimator of the covariance

matrix for  $\hat{\beta}$  and set  $\hat{\theta} = \hat{\beta}_1 / \hat{\beta}_2$ . Define

$$\hat{\mathbf{R}}_1 = \begin{pmatrix} 0 \\ 1 \\ \frac{1}{\hat{\beta}_2} \\ -\frac{\hat{\beta}_1}{\hat{\beta}_2^2} \end{pmatrix}$$

so that the standard error for  $\hat{\theta}$  is  $s(\hat{\theta}) = (\hat{\mathbf{R}}_1' \hat{\mathbf{V}}_{\hat{\beta}} \hat{\mathbf{R}}_1)^{1/2}$ . In this case a t-statistic for  $H_0$  is

$$T_1 = \frac{\left(\frac{\hat{\beta}_1}{\hat{\beta}_2} - \theta_0\right)}{s(\hat{\theta})}.$$

An alternative statistic can be constructed through reformulating the null hypothesis as

$$H_0: \beta_1 - \theta_0 \beta_2 = 0.$$

A t-statistic based on this formulation of the hypothesis is

$$T_2 = \frac{\hat{\beta}_1 - \theta_0 \hat{\beta}_2}{(\mathbf{R}_2' \hat{\mathbf{V}}_{\hat{\beta}} \mathbf{R}_2)^{1/2}}$$

where

$$\mathbf{R}_2 = \begin{pmatrix} 0 \\ 1 \\ -\theta_0 \end{pmatrix}.$$

To compare  $T_1$  and  $T_2$  we perform another simple Monte Carlo simulation. We let  $x_{1i}$  and  $x_{2i}$  be mutually independent  $N(0, 1)$  variables,  $e_i$  be an independent  $N(0, \sigma^2)$  draw with  $\sigma = 3$ , and normalize  $\beta_0 = 0$  and  $\beta_1 = 1$ . This leaves  $\beta_2$  as a free parameter, along with sample size  $n$ . We vary  $\beta_2$  among .1, .25, .50, .75, and 1.0 and  $n$  among 100 and 500.

Table 9.3: Type I Error Probability of Asymptotic 5% t-tests

$\beta_2$	$n = 100$				$n = 500$			
	$\mathbb{P}(T < -1.645)$		$\mathbb{P}(T > 1.645)$		$\mathbb{P}(T < -1.645)$		$\mathbb{P}(T > 1.645)$	
	$T_1$	$T_2$	$T_1$	$T_2$	$T_1$	$T_2$	$T_1$	$T_2$
0.10	0.47	0.05	0.00	0.05	0.28	0.05	0.00	0.05
0.25	0.27	0.05	0.00	0.05	0.16	0.05	0.00	0.05
0.50	0.14	0.05	0.00	0.05	0.12	0.05	0.00	0.05
0.75	0.03	0.05	0.00	0.05	0.08	0.05	0.01	0.05
1.00	0.00	0.05	0.00	0.05	0.03	0.05	0.03	0.05

Rejection frequencies from 50,000 simulated random samples.

The one-sided Type I error probabilities  $\mathbb{P}(T < -1.645)$  and  $\mathbb{P}(T > 1.645)$  are calculated from 50,000 simulated samples. The results are presented in Table 9.3. Ideally, the entries in the table should be 0.05. However, the rejection rates for the  $T_1$  statistic diverge greatly from this value, especially for small values of  $\beta_2$ . The left tail probabilities  $\mathbb{P}(T_1 < -1.645)$  greatly exceed 5%, while the right tail probabilities  $\mathbb{P}(T_1 > 1.645)$  are close to zero in most cases. In contrast, the rejection rates for the linear  $T_2$  statistic are invariant to the value of  $\beta_2$ , and equal 5% for both sample sizes. The implication of Table 9.3 is that the two t-ratios have dramatically different sampling behavior.

The common message from both examples is that Wald statistics are sensitive to the algebraic formulation of the null hypothesis.

A simple solution is to use the minimum distance statistic  $J$ , which equals  $W$  with  $r = 1$  in the first example, and  $|T_2|$  in the second example. The minimum distance statistic is invariant to the algebraic formulation of the null hypothesis, so is immune to this problem. Whenever possible, the Wald statistic should not be used to test nonlinear hypotheses.

Theoretical investigations of these issues include Park and Phillips (1988) and Dufour (1997).

## 9.18 Monte Carlo Simulation

In Section 9.17 we introduced the method of Monte Carlo simulation to illustrate the small sample problems with tests of nonlinear hypotheses. In this section we describe the method in more detail.

Recall, our data consist of observations  $(y_i, \mathbf{x}_i)$  which are random draws from a population distribution  $F$ . Let  $\boldsymbol{\theta}$  be a parameter and let  $T = T((y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n), \boldsymbol{\theta})$  be a statistic of interest, for example an estimator  $\hat{\theta}$  or a t-statistic  $(\hat{\theta} - \theta) / s(\hat{\theta})$ . The exact distribution of  $T$  is

$$G(u, F) = \mathbb{P}(T \leq u | F).$$

While the asymptotic distribution of  $T$  might be known, the exact (finite sample) distribution  $G$  is generally unknown.

Monte Carlo simulation uses numerical simulation to compute  $G(u, F)$  for selected choices of  $F$ . This is useful to investigate the performance of the statistic  $T$  in reasonable situations and sample sizes. The basic idea is that for any given  $F$ , the distribution function  $G(u, F)$  can be calculated numerically through simulation. The name Monte Carlo derives from the famous Mediterranean gambling resort where games of chance are played.

The method of Monte Carlo is quite simple to describe. The researcher chooses  $F$  (the distribution of the data) and the sample size  $n$ . A “true” value of  $\boldsymbol{\theta}$  is implied by this choice, or equivalently the value  $\boldsymbol{\theta}$  is selected directly by the researcher which implies restrictions on  $F$ .

Then the following experiment is conducted by computer simulation:

1.  $n$  independent random pairs  $(y_i^*, \mathbf{x}_i^*)$ ,  $i = 1, \dots, n$ , are drawn from the distribution  $F$  using the computer’s random number generator.
2. The statistic  $T = T((y_1^*, \mathbf{x}_1^*), \dots, (y_n^*, \mathbf{x}_n^*), \boldsymbol{\theta})$  is calculated on this pseudo data.

For step 1, computer packages have built-in random number procedures including  $U[0, 1]$  and  $N(0, 1)$ . From these most random variables can be constructed. (For example, a chi-square can be generated by sums of squares of normals.)

For step 2, it is important that the statistic be evaluated at the “true” value of  $\boldsymbol{\theta}$  corresponding to the choice of  $F$ .

The above experiment creates one random draw from the distribution  $G(u, F)$ . This is one observation from an unknown distribution. Clearly, from one observation very little can be said. So the researcher repeats the experiment  $B$  times, where  $B$  is a large number. Typically, we set  $B = 1000$  or  $B = 5000$ . We will discuss this choice later.

Notationally, let the  $b^{th}$  experiment result in the draw  $T_b$ ,  $b = 1, \dots, B$ . These results are stored. After all  $B$  experiments have been calculated, these results constitute a random sample of size  $B$  from the distribution of  $G(u, F) = \mathbb{P}(T_b \leq u) = \mathbb{P}(T \leq u | F)$ .

From a random sample, we can estimate any feature of interest using (typically) a method of moments estimator. We now describe some specific examples.

Suppose we are interested in the bias, mean-squared error (MSE), and/or variance of the distribution of  $\hat{\theta} - \theta$ . We then set  $T = \hat{\theta} - \theta$ , run the above experiment, and calculate

$$\begin{aligned}\widehat{\text{bias}}(\hat{\theta}) &= \frac{1}{B} \sum_{b=1}^B T_b = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b - \theta \\ \widehat{\text{mse}}(\hat{\theta}) &= \frac{1}{B} \sum_{b=1}^B (T_b)^2 = \frac{1}{B} \sum_{b=1}^B (\hat{\theta}_b - \theta)^2 \\ \widehat{\text{var}}(\hat{\theta}) &= \widehat{\text{mse}}(\hat{\theta}) - \left( \widehat{\text{bias}}(\hat{\theta}) \right)^2\end{aligned}$$

Suppose we are interested in the Type I error associated with an asymptotic 5% two-sided t-test. We would then set  $T = |\hat{\theta} - \theta| / s(\hat{\theta})$  and calculate

$$\hat{P} = \frac{1}{B} \sum_{b=1}^B \mathbb{1}(T_b \geq 1.96), \quad (9.15)$$

the percentage of the simulated t-ratios which exceed the asymptotic 5% critical value.

Suppose we are interested in the 5% and 95% quantile of  $T = \hat{\theta}$  or  $T = (\hat{\theta} - \theta) / s(\hat{\theta})$ . We then compute the 5% and 95% sample quantiles of the sample  $\{T_b\}$ . The  $\alpha$  sample quantile is a number  $q_\alpha$  such that  $100\alpha\%$  of the sample are less than  $q_\alpha$ . A simple way to compute sample quantiles is to sort the sample  $\{T_b\}$  from low to high. Then  $q_\alpha$  is the  $N^{\text{th}}$  number in this ordered sequence, where  $N = B\alpha$ . For example, if we set  $B = 1000$ , then the 5% sample quantile is 50<sup>th</sup> sorted value and the 95% sample quantile is the 950<sup>th</sup> sorted value.

The typical purpose of a Monte Carlo simulation is to investigate the performance of a statistical procedure in realistic settings. Generally, the performance will depend on  $n$  and  $F$ . In many cases, an estimator or test may perform wonderfully for some values, and poorly for others. It is therefore useful to conduct a variety of experiments, for a selection of choices of  $n$  and  $F$ .

As discussed above, the researcher must select the number of experiments,  $B$ . Often this is called the number of **replications**. Quite simply, a larger  $B$  results in more precise estimates of the features of interest of  $G$ , but requires more computational time. In practice, therefore, the choice of  $B$  is often guided by the computational demands of the statistical procedure. Since the results of a Monte Carlo experiment are estimates computed from a random sample of size  $B$ , it is straightforward to calculate standard errors for any quantity of interest. If the standard error is too large to make a reliable inference, then  $B$  will have to be increased.

In particular, it is simple to make inferences about rejection probabilities from statistical tests, such as the percentage estimate reported in (9.15). The random variable  $\mathbb{1}(T_b \geq 1.96)$  is i.i.d. Bernoulli, equalling 1 with probability  $p = \mathbb{E}(\mathbb{1}(T_b \geq 1.96))$ . The average (9.15) is therefore an unbiased estimator of  $p$  with standard error  $s(\hat{p}) = \sqrt{p(1-p)/B}$ . As  $p$  is unknown, this may be approximated by replacing  $p$  with  $\hat{p}$  or with an hypothesized value. For example, if we are assessing an asymptotic 5% test, then we can set  $s(\hat{p}) = \sqrt{(.05)(.95)/B} \approx .22/\sqrt{B}$ . Hence, standard errors for  $B = 100$ , 1000, and 5000, are, respectively,  $s(\hat{p}) = .022$ , .007, and .003.

Most papers in econometric methods and some empirical papers include the results of Monte Carlo simulations to illustrate the performance of their methods. When extending existing results, it is good practice to start by replicating existing (published) results. This is not exactly possible in the case of simulation results, as they are inherently random. For example suppose a paper investigates a statistical test, and reports a simulated rejection probability of 0.07 based on a simulation with  $B = 100$  replications. Suppose you attempt to replicate this result, and find a rejection probability of 0.03 (again using  $B = 100$  simulation replications). Should you conclude that you have failed in your attempt? Absolutely not! Under the hypothesis that both simulations are identical, you have two independent estimates,  $\hat{p}_1 = 0.07$  and  $\hat{p}_2 = 0.03$ , of a common probability  $p$ . The asymptotic (as  $B \rightarrow \infty$ ) distribution of their difference is  $\sqrt{B}(\hat{p}_1 - \hat{p}_2) \xrightarrow{d} N(0, 2p(1-p)/B)$ , so a standard error for  $\hat{p}_1 - \hat{p}_2 = 0.04$  is  $\hat{s} = \sqrt{2p(1-p)/B} \approx 0.03$ , using the estimate  $p = (\hat{p}_1 + \hat{p}_2)/2$ . Since the t-ratio  $0.04/0.03 = 1.3$  is not statistically significant, it is incorrect

to reject the null hypothesis that the two simulations are identical. The difference between the results  $\hat{p}_1 = 0.07$  and  $\hat{p}_2 = 0.03$  is consistent with random variation.

What should be done? The first mistake was to copy the previous paper's choice of  $B = 100$ . Instead, suppose you set  $B = 10,000$ . Suppose you now obtain  $\hat{p}_2 = 0.04$ . Then  $\hat{p}_1 - \hat{p}_2 = 0.03$  and a standard error is  $\hat{s} = \sqrt{p(1-p)(1/100 + 1/10000)} \approx 0.02$ . Still we cannot reject the hypothesis that the two simulations are different. Even though the estimates (0.07 and 0.04) appear to be quite different, the difficulty is that the original simulation used a very small number of replications ( $B = 100$ ) so the reported estimate is quite imprecise. In this case, it is appropriate to conclude that your results "replicate" the previous study, as there is no statistical evidence to reject the hypothesis that they are equivalent.

Most journals have policies requiring authors to make available their data sets and computer programs required for empirical results. They do not have similar policies regarding simulations. Nevertheless, it is good professional practice to make your simulations available. The best practice is to post your simulation code on your webpage. This invites others to build on and use your results, leading to possible collaboration, citation, and/or advancement.

## 9.19 Confidence Intervals by Test Inversion

There is a close relationship between hypothesis tests and confidence intervals. We observed in Section 7.13 that the standard 95% asymptotic confidence interval for a parameter  $\theta$  is

$$\begin{aligned}\hat{C} &= [\hat{\theta} - 1.96 \cdot s(\hat{\theta}), \quad \hat{\theta} + 1.96 \cdot s(\hat{\theta})] \\ &= \{\theta : |T(\theta)| \leq 1.96\}.\end{aligned}\tag{9.16}$$

That is, we can describe  $\hat{C}$  as "The point estimate plus or minus 2 standard errors" or "The set of parameter values not rejected by a two-sided t-test." The second definition, known as **test statistic inversion**, is a general method for finding confidence intervals, and typically produces confidence intervals with excellent properties.

Given a test statistic  $T(\theta)$  and critical value  $c$ , the acceptance region "Accept if  $T(\theta) \leq c$ " is identical to the confidence interval  $\hat{C} = \{\theta : T(\theta) \leq c\}$ . Since the regions are identical, the probability of coverage  $\mathbb{P}(\theta \in \hat{C})$  equals the probability of correct acceptance  $\mathbb{P}(\text{Accept}|\theta)$  which is exactly 1 minus the Type I error probability. Thus inverting a test with good Type I error probabilities yields a confidence interval with good coverage probabilities.

Now suppose that the parameter of interest  $\theta = r(\boldsymbol{\beta})$  is a nonlinear function of the coefficient vector  $\boldsymbol{\beta}$ . In this case the standard confidence interval for  $\theta$  is the set  $\hat{C}$  as in (9.16) where  $\hat{\theta} = r(\hat{\boldsymbol{\beta}})$  is the point estimator and  $s(\hat{\theta}) = \sqrt{\hat{\mathbf{R}}' \hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}} \hat{\mathbf{R}}}$  is the delta method standard error. This confidence interval is inverting the t-test based on the nonlinear hypothesis  $r(\boldsymbol{\beta}) = \theta$ . The trouble is that in Section 9.17 we learned that there is no unique t-statistic for tests of nonlinear hypotheses and that the choice of parameterization matters greatly.

For example, if  $\theta = \beta_1 / \beta_2$  then the coverage probability of the standard interval (9.16) is 1 minus the probability of the Type I error, which as shown in Table 8.2 can be far from the nominal 5%.

In this example a good solution is the same as discussed in Section 9.17 – to rewrite the hypothesis as a linear restriction. The hypothesis  $\theta = \beta_1 / \beta_2$  is the same as  $\theta\beta_2 = \beta_1$ . The t-statistic for this restriction is

$$T(\theta) = \frac{\hat{\beta}_1 - \hat{\beta}_2\theta}{\left(\mathbf{R}' \hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}} \mathbf{R}\right)^{1/2}}$$

where

$$\mathbf{R} = \begin{pmatrix} 1 \\ -\theta \end{pmatrix}$$

and  $\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$  is the covariance matrix for  $(\hat{\beta}_1 \ \hat{\beta}_2)$ . A 95% confidence interval for  $\theta = \beta_1 / \beta_2$  is the set of values of  $\theta$  such that  $|T(\theta)| \leq 1.96$ . Since  $T(\theta)$  is a non-linear function of  $\theta$  one method to find the confidence set is by grid search over  $\theta$ .

For example, in the wage equation

$$\log(Wage) = \beta_1 Experience + \beta_2 Experience^2 / 100 + \dots$$

the highest expected wage occurs at  $Experience = -50\beta_1/\beta_2$ . From Table 4.1 we have the point estimate  $\hat{\theta} = 29.8$  and we can calculate the standard error  $s(\hat{\theta}) = 0.022$  for a 95% confidence interval [29.8, 29.9]. However, if we instead invert the linear form of the test we can numerically find the interval [29.1, 30.6] which is much larger. From the evidence presented in Section 9.17 we know the first interval can be quite inaccurate and the second interval is greatly preferred.

## 9.20 Multiple Tests and Bonferroni Corrections

In most applications, economists examine a large number of estimates, test statistics, and p-values. What does it mean (or does it mean anything) if one statistic appears to be “significant” after examining a large number of statistics? This is known as the problem of **multiple testing** or **multiple comparisons**.

To be specific, suppose we examine a set of  $k$  coefficients, standard errors and t-ratios, and consider the “significance” of each statistic. Based on conventional reasoning, for each coefficient we would reject the hypothesis that the coefficient is zero with asymptotic size  $\alpha$  if the absolute t-statistic exceeds the  $1 - \alpha$  critical value of the normal distribution, or equivalently if the p-value for the t-statistic is smaller than  $\alpha$ . If we observe that one of the  $k$  statistics is “significant” based on this criteria, that means that one of the p-values is smaller than  $\alpha$ , or equivalently, that the smallest p-value is smaller than  $\alpha$ . We can then rephrase the question: Under the joint hypothesis that a set of  $k$  hypotheses are all true, what is the probability that the smallest p-value is smaller than  $\alpha$ ? In general, we cannot provide a precise answer to this question, but the Bonferroni correction bounds this probability by  $\alpha k$ . The Bonferroni method furthermore suggests that if we want the familywise error probability (the probability that one of the tests falsely rejects) to be bounded below  $\alpha$ , then an appropriate rule is to reject only if the smallest p-value is smaller than  $\alpha/k$ . Equivalently, the Bonferroni familywise p-value is  $k \min_{j \leq k} p_j$ .

Formally, suppose we have  $k$  hypotheses  $H_j$ ,  $j = 1, \dots, k$ . For each we have a test and associated p-value  $p_j$  with the property that when  $H_j$  is true  $\lim_{n \rightarrow \infty} \mathbb{P}(p_j < \alpha) = \alpha$ . We then observe that among the  $k$  tests, one of the  $k$  will appear “significant” if  $\min_{j \leq k} p_j < \alpha$ . This event can be written as

$$\left\{ \min_{j \leq k} p_j < \alpha \right\} = \bigcup_{j=1}^k \{p_j < \alpha\}.$$

Boole's inequality states that for any  $k$  events  $A_j$ ,  $\mathbb{P}\left(\bigcup_{j=1}^k A_j\right) \leq \sum_{j=1}^k \mathbb{P}(A_j)$ . Thus

$$\mathbb{P}\left(\min_{j \leq k} p_j < \alpha\right) \leq \sum_{j=1}^k \mathbb{P}(p_j < \alpha) \longrightarrow k\alpha$$

as stated. This demonstrates that the familywise rejection probability is at most  $k$  times the individual rejection probability.

Furthermore,

$$\mathbb{P}\left(\min_{j \leq k} p_j < \frac{\alpha}{k}\right) \leq \sum_{j=1}^k \mathbb{P}\left(p_j < \frac{\alpha}{k}\right) \longrightarrow \alpha.$$

This demonstrates that the family rejection probability can be controlled (bounded below  $\alpha$ ) if each individual test is subjected to the stricter standard that a p-value must be smaller than  $\alpha/k$  to be labeled as “significant.”

To illustrate, suppose we have two coefficient estimates, with individual p-values 0.04 and 0.15. Based on a conventional 5% level, the standard individual tests would suggest that the first coefficient

estimate is “significant” but not the second. A Bonferroni 5% test, however, does not reject as it would require that the smallest p-value be smaller than 0.025, which is not the case in this example. Alternatively, the Bonferroni familywise p-value is 0.08, which is not significant at the 5% level.

In contrast, if the two p-values are 0.01 and 0.15, then the Bonferroni familywise p-value is 0.02, which is significant at the 5% level.

## 9.21 Power and Test Consistency

The **power** of a test is the probability of rejecting  $\mathbb{H}_0$  when  $\mathbb{H}_1$  is true.

For simplicity suppose that  $y_i$  is i.i.d.  $N(\theta, \sigma^2)$  with  $\sigma^2$  known, consider the t-statistic  $T(\theta) = \sqrt{n}(\bar{y} - \theta)/\sigma$ , and tests of  $\mathbb{H}_0 : \theta = 0$  against  $\mathbb{H}_1 : \theta > 0$ . We reject  $\mathbb{H}_0$  if  $T = T(0) > c$ . Note that

$$T = T(\theta) + \sqrt{n}\theta/\sigma$$

and  $T(\theta)$  has an exact  $N(0, 1)$  distribution. This is because  $T(\theta)$  is centered at the true mean  $\theta$ , while the test statistic  $T(0)$  is centered at the (false) hypothesized mean of 0.

The power of the test is

$$\mathbb{P}(T > c | \theta) = \mathbb{P}(Z + \sqrt{n}\theta/\sigma > c) = 1 - \Phi(c - \sqrt{n}\theta/\sigma).$$

This function is monotonically increasing in  $\mu$  and  $n$ , and decreasing in  $\sigma$  and  $c$ .

Notice that for any  $c$  and  $\theta \neq 0$ , the power increases to 1 as  $n \rightarrow \infty$ . This means that for  $\theta \in \mathbb{H}_1$ , the test will reject  $\mathbb{H}_0$  with probability approaching 1 as the sample size gets large. We call this property **test consistency**.

**Definition 9.3** A test of  $\mathbb{H}_0 : \theta \in \Theta_0$  is **consistent against fixed alternatives** if for all  $\theta \in \Theta_1$ ,  $\mathbb{P}(\text{Reject } \mathbb{H}_0 | \theta) \rightarrow 1$  as  $n \rightarrow \infty$ .

For tests of the form “Reject  $\mathbb{H}_0$  if  $T > c$ ”, a sufficient condition for test consistency is that the  $T$  diverges to positive infinity with probability one for all  $\theta \in \Theta_1$ .

**Definition 9.4** We say that  $T \xrightarrow{p} \infty$  as  $n \rightarrow \infty$  if for all  $M < \infty$ ,  $\mathbb{P}(T \leq M) \rightarrow 0$  as  $n \rightarrow \infty$ . Similarly, we say that  $T \xrightarrow{p} -\infty$  as  $n \rightarrow \infty$  if for all  $M < \infty$ ,  $\mathbb{P}(T \geq -M) \rightarrow 0$  as  $n \rightarrow \infty$ .

In general, t-tests and Wald tests are consistent against fixed alternatives. Take a t-statistic for a test of  $\mathbb{H}_0 : \theta = \theta_0$

$$T = \frac{\hat{\theta} - \theta_0}{s(\hat{\theta})}$$

where  $\theta_0$  is a known value and  $s(\hat{\theta}) = \sqrt{n^{-1}\hat{V}_{\theta}}$ . Note that

$$T = \frac{\hat{\theta} - \theta}{s(\hat{\theta})} + \frac{\sqrt{n}(\theta - \theta_0)}{\sqrt{\hat{V}_{\theta}}}.$$

The first term on the right-hand-side converges in distribution to  $N(0, 1)$ . The second term on the right-hand-side equals zero if  $\theta = \theta_0$ , converges in probability to  $+\infty$  if  $\theta > \theta_0$ , and converges in probability to  $-\infty$  if  $\theta < \theta_0$ . Thus the two-sided t-test is consistent against  $\mathbb{H}_1 : \theta \neq \theta_0$ , and one-sided t-tests are consistent against the alternatives for which they are designed.

**Theorem 9.8** Under Assumptions 7.2, 7.3, and 7.4, for  $\boldsymbol{\theta} = \mathbf{r}(\boldsymbol{\beta}) \neq \boldsymbol{\theta}_0$  and  $q = 1$ , then  $|T| \xrightarrow{P} \infty$ , so for any  $c < \infty$  the test “Reject  $\mathbb{H}_0$  if  $|T| > c$ ” is consistent against fixed alternatives.

The Wald statistic for  $\mathbb{H}_0 : \boldsymbol{\theta} = \mathbf{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$  against  $\mathbb{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$  is

$$W = n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \hat{V}_{\boldsymbol{\theta}}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

Under  $\mathbb{H}_1$ ,  $\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ . Thus  $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \hat{V}_{\boldsymbol{\theta}}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{P} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)' V_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) > 0$ . Hence under  $\mathbb{H}_1$ ,  $W \xrightarrow{P} \infty$ . Again, this implies that Wald tests are consistent tests.

**Theorem 9.9** Under Assumptions 7.2, 7.3, and 7.4, for  $\boldsymbol{\theta} = \mathbf{r}(\boldsymbol{\beta}) \neq \boldsymbol{\theta}_0$ , then  $W \xrightarrow{P} \infty$ , so for any  $c < \infty$  the test “Reject  $\mathbb{H}_0$  if  $W > c$ ” is consistent against fixed alternatives.

## 9.22 Asymptotic Local Power

Consistency is a good property for a test, but does not give a useful approximation to the power of a test. To approximate the power function we need a distributional approximation.

The standard asymptotic method for power analysis uses what are called **local alternatives**. This is similar to our analysis of restriction estimation under misspecification (Section 8.13). The technique is to index the parameter by sample size so that the asymptotic distribution of the statistic is continuous in a localizing parameter. In this section we consider t-tests on real-valued parameters and in the next section consider Wald tests. Specifically, we consider parameter vectors  $\boldsymbol{\beta}_n$  which are indexed by sample size  $n$  and satisfy the real-valued relationship

$$\theta_n = \mathbf{r}(\boldsymbol{\beta}_n) = \theta_0 + n^{-1/2} h \quad (9.17)$$

where the scalar  $h$  is called a **localizing parameter**. We index  $\boldsymbol{\beta}_n$  and  $\theta_n$  by sample size to indicate their dependence on  $n$ . The way to think of (9.17) is that the true value of the parameters are  $\boldsymbol{\beta}_n$  and  $\theta_n$ . The parameter  $\theta_n$  is close to the hypothesized value  $\theta_0$ , with deviation  $n^{-1/2} h$ .

The specification (9.17) states that for any fixed  $h$ ,  $\theta_n$  approaches  $\theta_0$  as  $n$  gets large. Thus  $\theta_n$  is “close” or “local” to  $\theta_0$ . The concept of a localizing sequence (9.17) might seem odd since in the actual world the sample size cannot mechanically affect the value of the parameter. Thus (9.17) should not be interpreted literally. Instead, it should be interpreted as a technical device which allows the asymptotic distribution to be continuous in the alternative hypothesis.

To evaluate the asymptotic distribution of the test statistic we start by examining the scaled estimate centered at the hypothesized value  $\theta_0$ . Breaking it into a term centered at the true value  $\theta_n$  and a remainder we find

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_n) + \sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) \\ &= \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_n) + h \end{aligned}$$

where the second equality is (9.17). The first term is asymptotically normal:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_n) \xrightarrow{d} \sqrt{V_{\boldsymbol{\theta}}} Z$$

where  $Z \sim N(0, 1)$ . Therefore

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \sqrt{V_\theta}Z + h \sim N(h, V_\theta).$$

This asymptotic distribution depends continuously on the localizing parameter  $h$ .

Applied to the t statistic we find

$$\begin{aligned} T &= \frac{\hat{\theta} - \theta_0}{s(\hat{\theta})} \\ &\xrightarrow{d} \frac{\sqrt{V_\theta}Z + h}{\sqrt{V_\theta}} \\ &\sim Z + \delta \end{aligned} \tag{9.18}$$

where  $\delta = h/\sqrt{V_\theta}$ . This generalizes Theorem 9.1 (which assumes  $H_0$  is true) to allow for local alternatives of the form (9.17).

Consider a t-test of  $H_0$  against the one-sided alternative  $H_1 : \theta > \theta_0$  which rejects  $H_0$  for  $T > c$  where  $\Phi(c) = 1 - \alpha$ . The **asymptotic local power** of this test is the limit (as the sample size diverges) of the rejection probability under the local alternative (9.17)

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P}(\text{Reject } H_0) &= \lim_{n \rightarrow \infty} \mathbb{P}(T > c) \\ &= \mathbb{P}(Z + \delta > c) \\ &= 1 - \Phi(c - \delta) \\ &= \Phi(\delta - c) \\ &\stackrel{\text{def}}{=} \pi(\delta). \end{aligned}$$

We call  $\pi(\delta)$  the **asymptotic local power function**.

In Figure 9.4 we plot the local power function  $\pi(\delta)$  as a function of  $\delta \in [-1, 4]$  for tests of asymptotic size  $\alpha = 0.10$ ,  $\alpha = 0.05$ , and  $\alpha = 0.01$ .  $\delta = 0$  corresponds to the null hypothesis so  $\pi(\delta) = \alpha$ . The power functions are monotonically increasing in  $\delta$ . Note that the power is lower than  $\alpha$  for  $\delta < 0$  due to the one-sided nature of the test.

We can see that the three power functions are ranked by  $\alpha$  so that the test with  $\alpha = 0.10$  has higher power than the test with  $\alpha = 0.01$ . This is the inherent trade-off between size and power. Decreasing size induces a decrease in power, and conversely.

The coefficient  $\delta$  can be interpreted as the parameter deviation measured as a multiple of the standard error  $s(\hat{\theta})$ . To see this, recall that  $s(\hat{\theta}) = n^{-1/2}\sqrt{\hat{V}_\theta} \approx n^{-1/2}\sqrt{V_\theta}$  and then note that

$$\delta = \frac{h}{\sqrt{V_\theta}} \approx \frac{n^{-1/2}h}{s(\hat{\theta})} = \frac{\theta_n - \theta_0}{s(\hat{\theta})}.$$

Thus  $\delta$  approximately equals the deviation  $\theta_n - \theta_0$  expressed as multiples of the standard error  $s(\hat{\theta})$ . Thus as we examine Figure 9.4, we can interpret the power function at  $\delta = 1$  (e.g. 26% for a 5% size test) as the power when the parameter  $\theta_n$  is one standard error above the hypothesized value. For example, from Table 4.1 the standard error for the coefficient on “Married Female” is 0.010. Thus in this example,  $\delta = 1$  corresponds to  $\theta_n = 0.010$  or an 1.0% wage premium for married females. Our calculations show that the asymptotic power of a one-sided 5% test against this alternative is about 26%.

The difference between power functions can be measured either vertically or horizontally. For example, in Figure 9.4 there is a vertical dotted line at  $\delta = 1$ , showing that the asymptotic local power function  $\pi(\delta)$  equals 39% for  $\alpha = 0.10$ , equals 26% for  $\alpha = 0.05$  and equals 9% for  $\alpha = 0.01$ . This is the difference in power across tests of differing size, holding fixed the parameter in the alternative.

A horizontal comparison can also be illuminating. To illustrate, in Figure 9.4 there is a horizontal dotted line at 50% power. 50% power is a useful benchmark, as it is the point where the test has equal

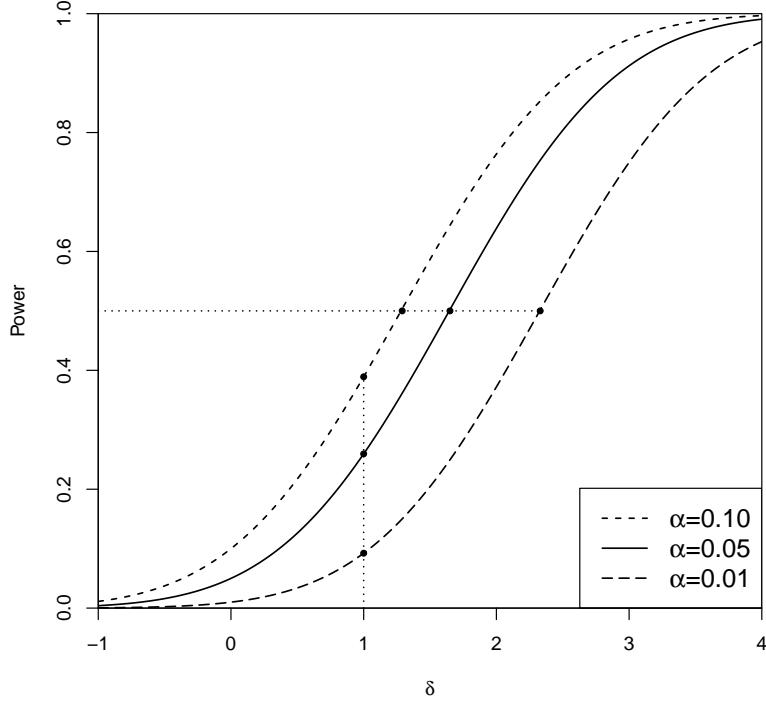


Figure 9.4: Asymptotic Local Power Function of One-Sided t Test

odds of rejection and acceptance. The dotted line crosses the three power curves at  $\delta = 1.29$  ( $\alpha = 0.10$ ),  $\delta = 1.65$  ( $\alpha = 0.05$ ), and  $\delta = 2.33$  ( $\alpha = 0.01$ ). This means that the parameter  $\theta$  must be at least 1.65 standard errors above the hypothesized value for a one-sided 5% test to have 50% (approximate) power.

The ratio of these values (e.g.  $1.65/1.29 = 1.28$  for the asymptotic 5% versus 10% tests) measures the relative parameter magnitude needed to achieve the same power. (Thus, for a 5% size test to achieve 50% power, the parameter must be 28% larger than for a 10% size test.) Even more interesting, the square of this ratio (e.g.  $(1.65/1.29)^2 = 1.64$ ) can be interpreted as the increase in sample size needed to achieve the same power under fixed parameters. That is, to achieve 50% power, a 5% size test needs 64% more observations than a 10% size test. This interpretation follows by the following informal argument. By definition and (9.17)  $\delta = h/\sqrt{V_\theta} = \sqrt{n}(\theta_n - \theta_0)/\sqrt{V_\theta}$ . Thus holding  $\theta$  and  $V_\theta$  fixed,  $\delta^2$  is proportional to  $n$ .

The analysis of a two-sided t test is similar. (9.18) implies that

$$T = \left| \frac{\hat{\theta} - \theta_0}{s(\hat{\theta})} \right| \xrightarrow{d} |Z + \delta|$$

and thus the local power of a two-sided t test is

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P}(\text{Reject } H_0) &= \lim_{n \rightarrow \infty} \mathbb{P}(T > c) \\ &= \mathbb{P}(|Z + \delta| > c) \\ &= \Phi(\delta - c) + \Phi(-\delta - c) \end{aligned}$$

which is monotonically increasing in  $|\delta|$ .

**Theorem 9.10** Under Assumptions 7.2, 7.3, 7.4, and  $\theta_n = r(\beta_n) = \theta_0 + n^{-1/2}h$ , then

$$T(\theta_0) = \frac{\hat{\theta} - \theta_0}{s(\hat{\theta})} \xrightarrow{d} Z + \delta$$

where  $Z \sim N(0, 1)$  and  $\delta = h/\sqrt{V_\theta}$ . For  $c$  such that  $\Phi(c) = 1 - \alpha$ ,

$$\mathbb{P}(T(\theta_0) > c) \longrightarrow \Phi(\delta - c).$$

Furthermore, for  $c$  such that  $\Phi(c) = 1 - \alpha/2$ ,

$$\mathbb{P}(|T(\theta_0)| > c) \longrightarrow \Phi(\delta - c) + \Phi(-\delta - c).$$

## 9.23 Asymptotic Local Power, Vector Case

In this section we extend the local power analysis of the previous section to the case of vector-valued alternatives. We generalize (9.17) to allow  $\theta_n$  to be vector-valued. The local parameterization takes the form

$$\theta_n = r(\beta_n) = \theta_0 + n^{-1/2}h \quad (9.19)$$

where  $h$  is  $q \times 1$ .

Under (9.19),

$$\begin{aligned} \sqrt{n}(\hat{\theta} - \theta_0) &= \sqrt{n}(\hat{\theta} - \theta_n) + h \\ &\xrightarrow{d} Z_h \sim N(h, V_\theta), \end{aligned}$$

a normal random vector with mean  $h$  and variance matrix  $V_\theta$ .

Applied to the Wald statistic we find

$$\begin{aligned} W &= n(\hat{\theta} - \theta_0)' \hat{V}_\theta^{-1} (\hat{\theta} - \theta_0) \\ &\xrightarrow{d} Z_h' V_\theta^{-1} Z_h \sim \chi_q^2(\lambda) \end{aligned} \quad (9.20)$$

where  $\lambda = h' V_\theta^{-1} h$ .  $\chi_q^2(\lambda)$  is a non-central chi-square random variable with non-centrality parameter  $\lambda$ . (See Section 5.3 and Theorem 5.12.)

The convergence (9.20) shows that under the local alternatives (9.19),  $W \xrightarrow{d} \chi_q^2(\lambda)$ . This generalizes the null asymptotic distribution which obtains as the special case  $\lambda = 0$ . We can use this result to obtain a continuous asymptotic approximation to the power function. For any significance level  $\alpha > 0$  set the asymptotic critical value  $c$  so that  $\mathbb{P}(\chi_q^2 > c) = \alpha$ . Then as  $n \rightarrow \infty$ ,

$$\mathbb{P}(W > c) \longrightarrow \mathbb{P}(\chi_q^2(\lambda) > c) \stackrel{\text{def}}{=} \pi(\lambda).$$

The asymptotic local power function  $\pi(\lambda)$  depends only on  $\alpha$ ,  $q$ , and  $\lambda$ .

**Theorem 9.11** Under Assumptions 7.2, 7.3, 7.4, and  $\theta_n = r(\beta_n) = \theta_0 + n^{-1/2}h$ , then

$$W \xrightarrow{d} \chi_q^2(\lambda)$$

where  $\lambda = h' V_\theta^{-1} h$ . Furthermore, for  $c$  such that  $\mathbb{P}(\chi_q^2 > c) = \alpha$ ,

$$\mathbb{P}(W > c) \longrightarrow \mathbb{P}(\chi_q^2(\lambda) > c).$$

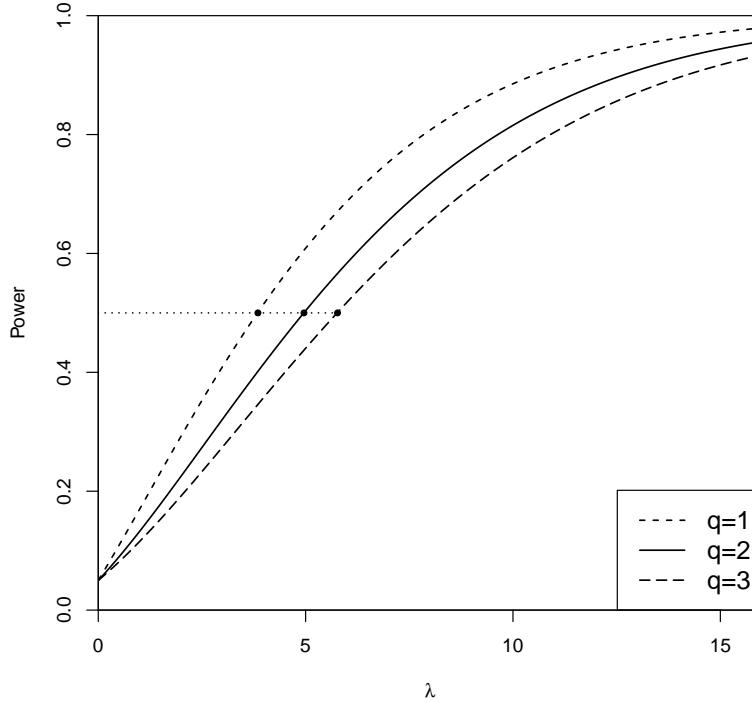


Figure 9.5: Asymptotic Local Power Function, Varying  $q$

Figure 9.5 plots  $\pi(\lambda)$  as a function of  $\lambda$  for  $q = 1$ ,  $q = 2$ , and  $q = 3$ , and  $\alpha = 0.05$ . The asymptotic power functions are monotonically increasing in  $\lambda$  and asymptote to one.

Figure 9.5 also shows the power loss for fixed non-centrality parameter  $\lambda$  as the dimensionality of the test increases. The power curves shift to the right as  $q$  increases, resulting in a decrease in power. This is illustrated by the dotted line at 50% power. The dotted line crosses the three power curves at  $\lambda = 3.85$  ( $q = 1$ ),  $\lambda = 4.96$  ( $q = 2$ ), and  $\lambda = 5.77$  ( $q = 3$ ). The ratio of these  $\lambda$  values correspond to the relative sample sizes needed to obtain the same power. Thus increasing the dimension of the test from  $q = 1$  to  $q = 2$  requires a 28% increase in sample size, or an increase from  $q = 1$  to  $q = 3$  requires a 50% increase in sample size, to obtain a test with 50% power.

## Exercises

**Exercise 9.1** Prove that if an additional regressor  $\mathbf{X}_{k+1}$  is added to  $\mathbf{X}$ , Theil's adjusted  $\bar{R}^2$  increases if and only if  $|T_{k+1}| > 1$ , where  $T_{k+1} = \hat{\beta}_{k+1}/s(\hat{\beta}_{k+1})$  is the t-ratio for  $\hat{\beta}_{k+1}$  and

$$s(\hat{\beta}_{k+1}) = (s^2[(\mathbf{X}'\mathbf{X})^{-1}]_{k+1,k+1})^{1/2}$$

is the homoskedasticity-formula standard error.

**Exercise 9.2** You have two independent samples  $(\mathbf{y}_1, \mathbf{X}_1)$  and  $(\mathbf{y}_2, \mathbf{X}_2)$  which satisfy  $\mathbf{y}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{e}_1$  and  $\mathbf{y}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}_2$ , where  $\mathbb{E}(\mathbf{x}_{1i}\mathbf{e}_{1i}) = \mathbf{0}$  and  $\mathbb{E}(\mathbf{x}_{2i}\mathbf{e}_{2i}) = \mathbf{0}$ , and both  $\mathbf{X}_1$  and  $\mathbf{X}_2$  have  $k$  columns. Let  $\hat{\boldsymbol{\beta}}_1$  and  $\hat{\boldsymbol{\beta}}_2$  be the OLS estimates of  $\boldsymbol{\beta}_1$  and  $\boldsymbol{\beta}_2$ . For simplicity, you may assume that both samples have the same number of observations  $n$ .

- (a) Find the asymptotic distribution of  $\sqrt{n}((\hat{\boldsymbol{\beta}}_2 - \hat{\boldsymbol{\beta}}_1) - (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1))$  as  $n \rightarrow \infty$ .
- (b) Find an appropriate test statistic for  $\mathbb{H}_0 : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_1$ .
- (c) Find the asymptotic distribution of this statistic under  $\mathbb{H}_0$ .

**Exercise 9.3** Let  $T$  be a t-statistic for  $\mathbb{H}_0 : \theta = 0$  versus  $\mathbb{H}_1 : \theta \neq 0$ . Since  $|T| \rightarrow_d |Z|$  under  $\mathbb{H}_0$ , someone suggests the test "Reject  $\mathbb{H}_0$  if  $|T| < c_1$  or  $|T| > c_2$ , where  $c_1$  is the  $\alpha/2$  quantile of  $|Z|$  and  $c_2$  is the  $1 - \alpha/2$  quantile of  $|Z|$ ".

- (a) Show that the asymptotic size of the test is  $\alpha$ .
- (b) Is this a good test of  $\mathbb{H}_0$  versus  $\mathbb{H}_1$ ? Why or why not?

**Exercise 9.4** Let  $W$  be a Wald statistic for  $\mathbb{H}_0 : \boldsymbol{\theta} = \mathbf{0}$  versus  $\mathbb{H}_1 : \boldsymbol{\theta} \neq \mathbf{0}$ , where  $\boldsymbol{\theta}$  is  $q \times 1$ . Since  $W \rightarrow_d \chi_q^2$  under  $\mathbb{H}_0$ , someone suggests the test "Reject  $\mathbb{H}_0$  if  $W < c_1$  or  $W > c_2$ , where  $c_1$  is the  $\alpha/2$  quantile of  $\chi_q^2$  and  $c_2$  is the  $1 - \alpha/2$  quantile of  $\chi_q^2$ ".

- (a) Show that the asymptotic size of the test is  $\alpha$ .
- (b) Is this a good test of  $\mathbb{H}_0$  versus  $\mathbb{H}_1$ ? Why or why not?

**Exercise 9.5** Take the linear model

$$\begin{aligned} y_i &= \mathbf{x}'_{1i}\boldsymbol{\beta}_1 + \mathbf{x}'_{2i}\boldsymbol{\beta}_2 + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) &= \mathbf{0} \end{aligned}$$

where both  $\mathbf{x}_{1i}$  and  $\mathbf{x}_{2i}$  are  $q \times 1$ . Show how to test the hypotheses  $\mathbb{H}_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$  against  $\mathbb{H}_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$ .

**Exercise 9.6** Suppose a researcher wants to know which of a set of 20 regressors has an effect on a variable *testscore*. He regresses *testscore* on the 20 regressors and reports the results. One of the 20 regressors (*studytime*) has a large t-ratio (about 2.5), while other t-ratios are insignificant (smaller than 2 in absolute value). He argues that the data show that *studytime* is the key predictor for *testscore*. Do you agree with this conclusion? Is there a deficiency in his reasoning?

**Exercise 9.7** Take the model

$$\begin{aligned} y_i &= x_i\beta_1 + x_i^2\beta_2 + e_i \\ \mathbb{E}(e_i | x_i) &= 0 \end{aligned}$$

where  $y_i$  is wages (dollars per hour) and  $x_i$  is age. Describe how you would test the hypothesis that the expected wage for a 40-year-old worker is \$20 an hour.

**Exercise 9.8** You want to test  $\mathbb{H}_0 : \beta_2 = 0$  against  $\mathbb{H}_1 : \beta_2 \neq 0$  in the model

$$\begin{aligned} y_i &= \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) &= 0 \end{aligned}$$

You read a paper which estimates model

$$y_i = \mathbf{x}'_{1i} \hat{\boldsymbol{\gamma}}_1 + (\mathbf{x}_{2i} - \mathbf{x}_{1i})' \hat{\boldsymbol{\gamma}}_2 + \hat{e}_i$$

and reports a test of  $\mathbb{H}_0 : \boldsymbol{\gamma}_2 = 0$  against  $\mathbb{H}_1 : \boldsymbol{\gamma}_2 \neq 0$ . Is this related to the test you wanted to conduct?

**Exercise 9.9** Suppose a researcher uses one dataset to test a specific hypothesis  $\mathbb{H}_0$  against  $\mathbb{H}_1$ , and finds that he can reject  $\mathbb{H}_0$ . A second researcher gathers a similar but independent dataset, uses similar methods and finds that she cannot reject  $\mathbb{H}_0$ . How should we (as interested professionals) interpret these mixed results?

**Exercise 9.10** In Exercise 7.8, you showed that  $\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \rightarrow_d N(0, V)$  as  $n \rightarrow \infty$  for some  $V$ . Let  $\hat{V}$  be an estimator of  $V$ .

- (a) Using this result, construct a t-statistic for  $\mathbb{H}_0 : \sigma^2 = 1$  against  $\mathbb{H}_1 : \sigma^2 \neq 1$ .
- (b) Using the Delta Method, find the asymptotic distribution of  $\sqrt{n}(\hat{\sigma} - \sigma)$ .
- (c) Use the previous result to construct a t-statistic for  $\mathbb{H}_0 : \sigma = 1$  against  $\mathbb{H}_1 : \sigma \neq 1$ .
- (d) Are the null hypotheses in (a) and (c) the same or are they different? Are the tests in (a) and (c) the same or are they different? If they are different, describe a context in which the two tests would give contradictory results.

**Exercise 9.11** Consider a regression such as Table 4.1 where both *experience* and its square are included. A researcher wants to test the hypothesis that *experience* does not affect mean wages, and does this by computing the t-statistic for *experience*. Is this the correct approach? If not, what is the appropriate testing method?

**Exercise 9.12** A researcher estimates a regression and computes a test of  $\mathbb{H}_0$  against  $\mathbb{H}_1$  and finds a p-value of  $p = 0.08$ , or “not significant”. She says “I need more data. If I had a larger sample the test will have more power and then the test will reject.” Is this interpretation correct?

**Exercise 9.13** A common view is that “If the sample size is large enough, any hypothesis will be rejected.” What does this mean? Interpret and comment.

**Exercise 9.14** Take the model

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) &= 0 \end{aligned}$$

with parameter of interest  $\theta = \mathbf{R}' \boldsymbol{\beta}$  with  $\mathbf{R}$   $k \times 1$ . Let  $\hat{\boldsymbol{\beta}}$  be the least-squares estimator and  $\hat{V}_{\hat{\boldsymbol{\beta}}}$  its variance estimator.

- (a) Write down  $\hat{C}$ , the 95% asymptotic confidence interval for  $\theta$ , in terms of  $\hat{\boldsymbol{\beta}}$ ,  $\hat{V}_{\hat{\boldsymbol{\beta}}}$ ,  $\mathbf{R}$ , and  $z = 1.96$  (the 97.5% quantile of  $N(0, 1)$ ).
- (b) Show that the decision “Reject  $\mathbb{H}_0$  if  $\theta_0 \notin \hat{C}$ ” is an asymptotic 5% test of  $\mathbb{H}_0 : \theta = \theta_0$ .

**Exercise 9.15** You are at a seminar where a colleague presents a simulation study of a test of a hypothesis  $H_0$  with nominal size 5%. Based on  $B = 100$  simulation replications under  $H_0$  the estimated size is 7%. Your colleague says: “Unfortunately the test over-rejects.”

- (a) Do you agree or disagree with your colleague? Explain. Hint: Use an asymptotic (large  $B$ ) approximation.
- (b) Suppose the number of simulation replications were  $B = 1000$  yet the estimated size is still 7%. Does your answer change?

**Exercise 9.16** You have  $n$  i.i.d. observations  $(y_i, x_{1i}, x_{2i})$ , and consider two alternative regression models

$$\begin{aligned} y_i &= \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + e_{1i} \\ \mathbb{E}(\mathbf{x}_{1i} e_{1i}) &= 0 \end{aligned} \tag{9.21}$$

$$\begin{aligned} y_i &= \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + e_{2i} \\ \mathbb{E}(\mathbf{x}_{2i} e_{2i}) &= 0 \end{aligned} \tag{9.22}$$

where  $\mathbf{x}_{1i}$  and  $\mathbf{x}_{2i}$  have at least some different regressors. (For example, (9.21) is a wage regression on geographic variables and (2) is a wage regression on personal appearance measurements.) You want to know if model (9.21) or model (9.22) fits the data better. Define  $\sigma_1^2 = E(e_{1i}^2)$  and  $\sigma_2^2 = E(e_{2i}^2)$ . You decide that the model with the smaller variance fit (e.g., model (9.21) fits better if  $\sigma_1^2 < \sigma_2^2$ .) You decide to test for this by testing the hypothesis of equal fit  $H_0 : \sigma_1^2 = \sigma_2^2$  against the alternative of unequal fit  $H_1 : \sigma_1^2 \neq \sigma_2^2$ . For simplicity, suppose that  $e_{1i}$  and  $e_{2i}$  are observed.

- (a) Construct an estimator  $\hat{\theta}$  of  $\theta = \sigma_1^2 - \sigma_2^2$ .
- (b) Find the asymptotic distribution of  $\sqrt{n}(\hat{\theta} - \theta)$  as  $n \rightarrow \infty$ .
- (c) Find an estimator of the asymptotic variance of  $\hat{\theta}$ .
- (d) Propose a test of asymptotic size  $\alpha$  of  $H_0$  against  $H_1$ .
- (e) Suppose the test accepts  $H_0$ . Briefly, what is your interpretation?

**Exercise 9.17** You have two regressors  $x_1$  and  $x_2$ , and estimate a regression with all quadratic terms

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \beta_4 x_{2i}^2 + \beta_5 x_{1i} x_{2i} + e_i$$

One of your advisors asks: Can we exclude the variable  $x_{2i}$  from this regression?

How do you translate this question into a statistical test? When answering these questions, be specific, not general.

- (a) What is the relevant null and alternative hypotheses?
- (b) What is an appropriate test statistic? Be specific.
- (c) What is the appropriate asymptotic distribution for the statistic? Be specific.
- (d) What is the rule for acceptance/rejection of the null hypothesis?

**Exercise 9.18** The observed data is  $\{y_i, \mathbf{x}_i, \mathbf{z}_i\} \in \mathbb{R} \times \mathbb{R}^k \times \mathbb{R}^\ell$ ,  $k > 1$  and  $\ell > 1$ ,  $i = 1, \dots, n$ . An econometrician first estimates

$$y_i = \mathbf{x}'_i \hat{\boldsymbol{\beta}} + \hat{e}_i$$

by least squares. The econometrician next regresses the residual  $\hat{e}_i$  on  $\mathbf{z}_i$ , which can be written as

$$\hat{e}_i = \mathbf{z}'_i \tilde{\boldsymbol{\gamma}} + \tilde{u}_i.$$

- (a) Define the population parameter  $\gamma$  being estimated in this second regression.
- (b) Find the probability limit for  $\tilde{\gamma}$ .
- (c) Suppose the econometrician constructs a Wald statistic  $W_n$  for  $H_0 : \gamma = \mathbf{0}$  from the second regression, ignoring the regression. Write down the formula for  $W_n$ .
- (d) Assuming  $\mathbb{E}(\mathbf{z}_i \mathbf{x}'_i) = \mathbf{0}$ , find the asymptotic distribution for  $W_n$  under  $H_0 : \gamma = \mathbf{0}$ .
- (e) If  $\mathbb{E}(\mathbf{z}_i \mathbf{x}'_i) \neq \mathbf{0}$  will your answer to (d) change?

**Exercise 9.19** An economist estimates  $y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + e_i$  by least-squares and tests the hypothesis  $H_0 : \beta_2 = 0$  against  $H_1 : \beta_2 \neq 0$ . She obtains a Wald statistic  $W_n = 0.34$ . The sample size is  $n = 500$ .

- (a) What is the correct degrees of freedom for the  $\chi^2$  distribution to evaluate the significance of the Wald statistic?
- (b) The Wald statistic  $W_n$  is very small. Indeed, is it less than the 1% quantile of the appropriate  $\chi^2$  distribution? If so, should you reject  $H_0$ ? Explain your reasoning.

**Exercise 9.20** You are reading a paper, and it reports the results from two nested OLS regressions:

$$\begin{aligned} y_i &= \mathbf{x}'_{1i} \tilde{\boldsymbol{\beta}}_1 + \tilde{e}_i \\ y_i &= \mathbf{x}'_{1i} \hat{\boldsymbol{\beta}}_1 + \mathbf{x}'_{2i} \hat{\boldsymbol{\beta}}_2 + \hat{e}_i \end{aligned}$$

Some summary statistics are reported:

Short Regression	Long Regression
$R^2 = .20$	$R^2 = .26$
$\sum_{i=1}^n \tilde{e}_i^2 = 106$	$\sum_{i=1}^n \hat{e}_i^2 = 100$
# of coefficients=5	# of coefficients=8
$n = 50$	$n = 50$

You are curious if the estimate  $\hat{\boldsymbol{\beta}}_2$  is statistically different from the zero vector. Is there a way to determine an answer from this information? Do you have to make any assumptions (beyond the standard regularity conditions) to justify your answer?

**Exercise 9.21** Take the model

$$\begin{aligned} y_i &= x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{4i}\beta_4 + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) &= 0 \end{aligned}$$

Describe how you would test

$$H_0 : \frac{\beta_1}{\beta_2} = \frac{\beta_3}{\beta_4}$$

against

$$H_1 : \frac{\beta_1}{\beta_2} \neq \frac{\beta_3}{\beta_4}.$$

**Exercise 9.22** You have a random sample from the model

$$\begin{aligned} y_i &= x_i \beta_1 + x_i^2 \beta_2 + e_i \\ \mathbb{E}(e_i | x_i) &= 0 \end{aligned}$$

where  $y_i$  is wages (dollars per hour) and  $x_i$  is age. Describe how you would test the hypothesis that the expected wage for a 40-year-old worker is \$20 an hour.

**Exercise 9.23** Let  $T_n$  be a test statistic such that under  $H_0$ ,  $T_n \rightarrow_d \chi^2_3$ . Since  $P(\chi^2_3 > 7.815) = 0.05$ , an asymptotic 5% test of  $H_0$  rejects when  $T_n > 7.815$ . An econometrician is interested in the Type I error of this test when  $n = 100$  and the data structure is well specified. She performs the following Monte Carlo experiment.

- $B = 200$  samples of size  $n = 100$  are generated from a distribution satisfying  $H_0$ .
- On each sample, the test statistic  $T_{nb}$  is calculated.
- She calculates  $\hat{p} = \frac{1}{B} \sum_{b=1}^B 1(T_{nb} > 7.815) = 0.070$
- The econometrician concludes that the test  $T_n$  is oversized in this context – it rejects too frequently under  $H_0$ .

Is her conclusion correct, incorrect, or incomplete? Be specific in your answer.

**Exercise 9.24** Do a Monte Carlo simulation. Take the model

$$\begin{aligned} y_i &= \alpha + x_i \beta + e_i \\ \mathbb{E}(x_i e_i) &= 0 \end{aligned}$$

where the parameter of interest is  $\theta = \exp(\beta)$ . Your data generating process (DGP) for the simulation is:  $x_i$  is  $U[0, 1]$ ,  $e_i$  is independent of  $x_i$  and  $N(0, 1)$ ,  $n = 50$ . Set  $\alpha = 0$  and  $\beta = 1$ . Generate  $B = 1000$  independent samples with  $\alpha$ . On each, estimate the regression by least-squares, calculate the covariance matrix using a standard (heteroskedasticity-robust) formula, and similarly estimate  $\theta$  and its standard error. For each replication, store  $\hat{\beta}$ ,  $\hat{\theta}$ ,  $t_\beta = (\hat{\beta} - \beta) / s(\hat{\beta})$ , and  $t_\theta = (\hat{\theta} - \theta) / s(\hat{\theta})$

- (a) Does the value of  $\alpha$  matter? Explain why the described statistics are **invariant** to  $\alpha$  and thus setting  $\alpha = 0$  is irrelevant.
- (b) From the 1000 replications estimate  $\mathbb{E}(\hat{\beta})$  and  $\mathbb{E}(\hat{\theta})$ . Discuss if you see evidence if either estimator is biased or unbiased.
- (c) From the 1000 replications estimate  $\mathbb{P}(t_\beta > 1.645)$  and  $\mathbb{P}(t_\theta > 1.645)$ . What does asymptotic theory predict these probabilities should be in large samples? What do your simulation results indicate?

**Exercise 9.25** The data set `Invest1993` on the textbook website contains data on 1962 U.S. firms extracted from Compustat and assembled by Bronwyn Hall. This particular dataset was used in Hall and Hall (1993).

The variables we use in this exercise are

- year year of the observation
- inva Investment to Capital Ratio
- vala Total Market Value to Asset Ratio (Tobin's Q)
- cfa Cash Flow to Asset Ratio
- debta Long Term Debt to Asset Ratio

The flow variables are annual sums. The stock variables are beginning of year.

- (a) Extract the sub-sample of observations for 1987. There should be 1028 observations. Estimate a linear regression of  $I$  (investment to capital ratio) on the other variables. Calculate appropriate standard errors.

- (b) Calculate asymptotic confidence intervals for the coefficients.
- (c) This regression is related to Tobin's  $q$  theory of investment, which suggests that investment should be predicted solely by  $Q$  (Tobin's  $Q$ ). This theory predicts that the coefficient on  $Q$  should be positive and the others should be zero. Test the joint hypothesis that the coefficients on cash flow ( $C$ ) and debt ( $D$ ) are zero. Test the hypothesis that the coefficient on  $Q$  is zero. Are the results consistent with the predictions of the theory?
- (d) Now try a non-linear (quadratic) specification. Regress  $I$  on  $Q, C, D, Q^2, C^2, D^2, QC, QD, CD$ . Test the joint hypothesis that the six interaction and quadratic coefficients are zero.

**Exercise 9.26** In a paper in 1963, Marc Nerlove analyzed a cost function for 145 American electric companies. His data set `Nerlove1963` is on the textbook website. The variables are

- $C$  Total cost
- $Q$  Output
- $PL$  Unit price of labor
- $PK$  Unit price of capital
- $PF$  Unit price of fuel

Nerlov was interested in estimating a *cost function*:  $C = f(Q, PL, PF, PK)$ .

- (a) First estimate an unrestricted Cobb-Douglas specification

$$\log C_i = \beta_1 + \beta_2 \log Q_i + \beta_3 \log PL_i + \beta_4 \log PK_i + \beta_5 \log PF_i + e_i. \quad (9.23)$$

Report parameter estimates and standard errors.

- (b) What is the economic meaning of the restriction  $H_0 : \beta_3 + \beta_4 + \beta_5 = 1$ ?
- (c) Estimate (9.23) by constrained least-squares imposing  $\beta_3 + \beta_4 + \beta_5 = 1$ . Report your parameter estimates and standard errors.
- (d) Estimate (9.23) by efficient minimum distance imposing  $\beta_3 + \beta_4 + \beta_5 = 1$ . Report your parameter estimates and standard errors.
- (e) Test  $H_0 : \beta_3 + \beta_4 + \beta_5 = 1$  using a Wald statistic.
- (f) Test  $H_0 : \beta_3 + \beta_4 + \beta_5 = 1$  using a minimum distance statistic.

**Exercise 9.27** In Section 8.12 we report estimates from Mankiw, Romer and Weil (1992). We reported estimation both by unrestricted least-squares and by constrained estimation, imposing the constraint that three coefficients ( $2^{nd}$ ,  $3^{rd}$  and  $4^{th}$  coefficients) sum to zero, as implied by the Solow growth theory. Using the same dataset `MRW1992` estimate the unrestricted model and test the hypothesis that the three coefficients sum to zero.

**Exercise 9.28** Using the CPS dataset and the subsample of non-hispanic blacks (race code = 2), test the hypothesis that marriage status does not affect mean wages.

- (a) Take the regression reported in Table 4.1. Which variables will need to be omitted to estimate a regression for the subsample of blacks?
- (b) Express the hypothesis "marriage status does not affect mean wages" as a restriction on the coefficients. How many restrictions is this?

- (c) Find the Wald (or F) statistic for this hypothesis. What is the appropriate distribution for the test statistic? Calculate the p-value of the test.
- (d) What do you conclude?

**Exercise 9.29** Using the CPS dataset and the subsample of non-hispanic blacks (race code = 2) and whites (race code = 1), test the hypothesis that the returns to education is common across groups.

- (a) Allow the return to education to vary across the four groups (white male, white female, black male, black female) by interacting dummy variables with *education*. Estimate an appropriate version of the regression reported in Table 4.1.
- (b) Find the Wald (or F) statistic for this hypothesis. What is the appropriate distribution for the test statistic? Calculate the p-value of the test.
- (c) What do you conclude?

# Chapter 10

## Resampling Methods

### 10.1 Introduction

So far in this textbook we have discussed two approaches to inference: exact and asymptotic. Both have their strengths and weaknesses. In this chapter we introduce a set of alternative approximation methods which are based around the concept of resampling – which means using sampling information extracted from the empirical distribution of the data. These are powerful methods, widely applicable, and often more accurate than exact or asymptotic approximations. Two disadvantages, however, are (1) resampling methods typically require more computation power; and (2) the theory is considerably more challenging. A consequence of the computation requirement is that most empirical researchers use asymptotic approximations for routine calculations, while resampling approximations are more typically used for final reporting.

We will discuss two categories of resampling methods used in statistical and econometric practice: jackknife and bootstrap. Most of our attention will be given to the bootstrap as it is the most commonly used resampling method in econometric practice.

The **jackknife** is the distribution obtained from the  $n$  leave-one-out estimators (see Section 3.20). The jackknife is most commonly used for variance estimation.

The **bootstrap** is the distribution obtained by estimation on samples created by i.i.d. sampling with replacement from the dataset. (There are other variants of bootstrap sampling, including parametric sampling and residual sampling.) The bootstrap is commonly used for variance estimation, confidence interval construction, and hypothesis testing.

There is a third category of resampling methods known as **sub-sampling** which we will not cover in this textbook. Sub-sampling is the distribution obtained by estimation on sub-samples (sampling without replacement) of the dataset. Sub-sampling can be used for most of same purposes as the bootstrap. See the excellent monograph by Politis, Romano and Wolf (1999).

### 10.2 Example

To motivate our discussion we focus on the application presented in Section 3.7, which is a bivariate regression applied to the CPS subsample of married black female wage earners with 12 years potential work experience and displayed in Table 3.1. The regression equation is

$$\log(Wage) = \beta_1 education + \beta_2 + e.$$

The estimates as reported in (4.39) are

$$\log(Wage) = \begin{array}{ll} 0.155 & education + 0.698 + \hat{\epsilon} \\ (0.031) & (0.493) \end{array}$$

$$\hat{\sigma}^2 = \begin{array}{l} 0.144 \\ (0.043) \end{array}$$

$$n = 20.$$

We focus on four estimates constructed from this regression. The first two are the coefficient estimates  $\hat{\beta}_1$  and  $\hat{\beta}_2$ . The third is the variance estimate  $\hat{\sigma}^2$ . The fourth is an estimate of the expected level of wages for an individual with 16 years of education (a college graduate), which turns out to be a nonlinear function of the parameters. Under the simplifying assumption that the error  $e$  is independent of the level of education we find that the expected level of wages is

$$\begin{aligned} \mu &= \mathbb{E}(Wage | Education = 16) \\ &= \mathbb{E}\exp(16\beta_1 + \beta_2 + e) \\ &= \exp(16\beta_1 + \beta_2)\mathbb{E}(\exp(e)) \\ &= \exp(16\beta_1 + \beta_2 + \sigma^2/2). \end{aligned}$$

The final equality holds under the further simplifying assumption that  $e \sim N(0, \sigma^2)$ . (In this case,  $\mathbb{E}(\exp(e)) = \exp(\sigma^2/2)$  can be obtained from the moment generating function.) The parameter  $\mu$  is a nonlinear function of the coefficients. The natural estimate of  $\mu$  replaces the unknowns by the point estimates. Thus

$$\hat{\mu} = \exp(16\hat{\beta}_1 + \hat{\beta}_2 + \hat{\sigma}^2/2) = 25.80 \quad (2.29)$$

The standard error for  $\hat{\mu}$  can be found by extending Exercise 7.8 to find the joint asymptotic distribution of  $\hat{\sigma}^2$  and the slope estimates, and then applying the delta method.

We are interested in calculating standard errors for the four estimates described above and constructing confidence intervals for the parameters. We are interested in going beyond exact and asymptotic approximations, especially given the small sample, the use of robust covariance matrix estimates, and the non-linear transformations. One of the challenges is that standard packages, such as Stata, provide standard errors for the coefficient estimates  $\hat{\beta}_1$  and  $\hat{\beta}_2$  and smooth nonlinear functions of the coefficient estimates, but not for the variance estimate  $\hat{\sigma}^2$  nor functionals of it such as  $\hat{\mu}$ .

### 10.3 Jackknife Estimation of Variance

The jackknife estimates moments of estimators using the distribution of the leave-one-out estimators. The jackknife estimator of bias was introduced by Quenouille (1949) and extended by Tukey (1958) to the jackknife estimator of variance. The idea was expanded further in the monographs of Efron (1982) and Shao and Tu (1995).

Let  $\hat{\theta}$  be any estimator of a vector-valued parameter  $\theta$  which is a function of a random sample of size  $n$ . Let  $V_{\hat{\theta}} = \text{var}(\hat{\theta})$  be the variance of  $\hat{\theta}$ . Define the leave-one-out estimators  $\hat{\theta}_{(-i)}$  which are computed using the formula for  $\hat{\theta}$  except that observation  $i$  is deleted. Tukey's jackknife estimator for  $V_{\hat{\theta}}$  is defined as a scale of the sample variance of the leave-one-out estimators:

$$\hat{V}_{\hat{\theta}}^{\text{jack}} = \frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}_{(-i)} - \bar{\theta})(\hat{\theta}_{(-i)} - \bar{\theta})' \quad (10.1)$$

where  $\bar{\theta}$  is the sample mean of the leave-one-out estimators

$$\bar{\theta} = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{(-i)}.$$

For scalar estimators  $\hat{\theta}$  the jackknife standard error is the square root of (10.1).

$$s_{\hat{\theta}}^{\text{jack}} = \sqrt{\hat{V}_{\hat{\theta}}^{\text{jack}}}$$

A convenient feature of the jackknife estimator  $\hat{V}_{\hat{\theta}}^{\text{jack}}$  is that the formula (10.1) is quite general and does not require any technical (exact or asymptotic) calculations. A downside is that it can require  $n$  separate estimations, which in some cases can be computationally costly.

In most cases  $\hat{V}_{\hat{\theta}}^{\text{jack}}$  will be similar to a robust asymptotic variance matrix estimator. Thus the main attractions of the jackknife estimator are that it can be used when an explicit asymptotic variance formula is not available, and that it can be used as a check on the reliability of an asymptotic formula.

The formula (10.1) is not immediately intuitive, so may benefit from some motivation. We start by examining the case of the sample mean  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ . The leave-one-out estimator is

$$\bar{y}_{(-i)} = \frac{1}{n-1} \sum_{j \neq i} y_j = \frac{n}{n-1} \bar{y} - \frac{1}{n-1} y_i. \quad (10.2)$$

The sample mean of the leave-one-out estimators is

$$\frac{1}{n} \sum_{i=1}^n \bar{y}_{(-i)} = \frac{n}{n-1} \bar{y} - \frac{1}{n-1} \bar{y} = \bar{y}.$$

The difference is

$$\bar{y}_{(-i)} - \bar{y} = \frac{1}{n-1} (\bar{y} - y_i).$$

The jackknife estimate of variance (10.1) is then

$$\begin{aligned} \hat{V}_{\bar{y}}^{\text{jack}} &= \frac{n-1}{n} \sum_{i=1}^n \left( \frac{1}{n-1} \right)^2 (\bar{y} - y_i) (\bar{y} - y_i)' \\ &= \frac{1}{n} \left( \frac{1}{n-1} \right) \sum_{i=1}^n (\bar{y} - y_i) (\bar{y} - y_i)' . \end{aligned} \quad (10.3)$$

This is identical to the conventional estimator for the variance of  $\bar{y}$ . Indeed, Tukey proposed the  $(n-1)/n$  scaling in (10.1) so that  $\hat{V}_{\bar{y}}^{\text{jack}}$  precisely equals the conventional estimator. This calculation shows that for the sample mean, the jackknife estimate of variance is identical to the conventional estimator.

We next examine the case of least-squares regression coefficient estimates. Recall from (3.44) that the leave-one-out OLS estimator equals

$$\hat{\beta}_{(-i)} = \hat{\beta} - (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i \tilde{e}_i \quad (10.4)$$

where  $\tilde{e}_i = (1 - h_{ii})^{-1} \hat{e}_i$  and  $h_{ii} = \mathbf{x}_i' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i$ . The sample mean of the leave-one-out estimators is

$$\bar{\beta} = \hat{\beta} - (\mathbf{X}' \mathbf{X})^{-1} \tilde{\mu}$$

where  $\tilde{\mu} = n^{-1} \sum_{i=1}^n \mathbf{x}_i \tilde{e}_i$ . Thus

$$\hat{\beta}_{(-i)} - \bar{\beta} = -(\mathbf{X}' \mathbf{X})^{-1} (\mathbf{x}_i \tilde{e}_i - \tilde{\mu}).$$

The jackknife estimate of variance for  $\hat{\beta}$  is

$$\begin{aligned}\hat{V}_{\hat{\beta}}^{\text{jack}} &= \frac{n-1}{n} \sum_{i=1}^n (\hat{\beta}_{(-i)} - \bar{\beta}) (\hat{\beta}_{(-i)} - \bar{\beta})' \\ &= \frac{n-1}{n} (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \tilde{e}_i^2 - n \tilde{\mu} \tilde{\mu}' \right) (\mathbf{X}' \mathbf{X})^{-1} \\ &= \frac{n-1}{n} \hat{V}_{\hat{\beta}}^{\text{HC3}} - (n-1) (\mathbf{X}' \mathbf{X})^{-1} \tilde{\mu} \tilde{\mu}' (\mathbf{X}' \mathbf{X})^{-1}\end{aligned}\quad (10.5)$$

where  $\hat{V}_{\hat{\beta}}^{\text{HC3}}$  is the HC3 covariance estimator (4.34) based on prediction errors. The second term in (10.5) is typically quite small since  $\tilde{\mu}$  is typically small in magnitude. Thus  $\hat{V}_{\hat{\beta}}^{\text{jack}} \approx \tilde{V}_{\hat{\beta}}$ . Indeed (4.34) was originally motivated as a simplification of the jackknife estimator. This shows that for regression coefficients the jackknife estimator of variance is similar to a conventional robust estimator. This is accomplished without the user “knowing” the form of the asymptotic covariance matrix. This is further confirmation that the jackknife is making a reasonable calculation.

Third, we examine the jackknife estimator for a function  $\hat{\theta} = \mathbf{r}(\hat{\beta})$  of a least-squares estimator. The leave-one-out estimator of  $\theta$  is

$$\begin{aligned}\hat{\theta}_{(-i)} &= \mathbf{r}(\hat{\beta}_{(-i)}) \\ &= \mathbf{r}(\hat{\beta} - (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i \tilde{e}_i) \\ &\approx \hat{\theta} - \hat{\mathbf{R}}' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i \tilde{e}_i.\end{aligned}$$

The second equality is (10.4). The final approximation is obtained by a mean-value expansion, using  $\mathbf{r}(\hat{\beta}) = \hat{\theta}$  and setting  $\hat{\mathbf{R}} = (\partial/\partial \beta) \mathbf{r}(\hat{\beta})'$ . This approximation holds in large samples since  $\hat{\beta}_{(-i)}$  are uniformly consistent for  $\beta$ . The jackknife variance estimator for  $\hat{\theta}$  thus equals

$$\begin{aligned}\hat{V}_{\hat{\theta}}^{\text{jack}} &= \frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}_{(-i)} - \bar{\theta}) (\hat{\theta}_{(-i)} - \bar{\theta})' \\ &\approx \frac{n-1}{n} \hat{\mathbf{R}}' (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \tilde{e}_i^2 - n \tilde{\mu} \tilde{\mu}' \right) (\mathbf{X}' \mathbf{X})^{-1} \hat{\mathbf{R}} \\ &= \hat{\mathbf{R}}' \hat{V}_{\hat{\beta}}^{\text{jack}} \hat{\mathbf{R}} \\ &\approx \hat{\mathbf{R}}' \tilde{V}_{\hat{\beta}} \hat{\mathbf{R}}.\end{aligned}$$

The final line equals a delta-method estimator for the variance of  $\hat{\theta}$  constructed with the covariance estimator (4.34). This shows that the jackknife estimator of variance for  $\hat{\theta}$  is approximately an asymptotic delta-method estimator. While this is an asymptotic approximation, it again shows that the jackknife produces an estimator which is asymptotically similar to one produced by asymptotic methods. This is despite the fact that the jackknife estimator is calculated without reference to asymptotic theory and does not require calculation of the derivatives of  $\mathbf{r}(\beta)$ .

This argument extends directly to any “smooth function” estimator. Most of the estimators discussed so far in this textbook take the form  $\hat{\theta} = \mathbf{g}(\bar{\mathbf{w}})$  where  $\bar{\mathbf{w}} = n^{-1} \sum_{i=1}^n \mathbf{w}_i$  and  $\mathbf{w}_i$  is some vector-valued function of the data. For any such estimator  $\hat{\theta}$ , the leave-one-out estimator equals  $\hat{\theta}_{(-i)} = \mathbf{g}(\bar{\mathbf{w}}_{(-i)})$  and its jackknife estimator of variance is (10.1). Using (10.2) and a mean-value expansion, we have the large-sample approximation

$$\begin{aligned}\hat{\theta}_{(-i)} &= \mathbf{g}(\bar{\mathbf{w}}_{(-i)}) \\ &= \mathbf{g} \left( \frac{n}{n-1} \bar{\mathbf{w}} - \frac{1}{n-1} \mathbf{w}_i \right) \\ &\approx \mathbf{g}(\bar{\mathbf{w}}) - \frac{1}{n-1} \mathbf{G}(\bar{\mathbf{w}})' \mathbf{w}_i\end{aligned}$$

where  $\mathbf{G}(\mathbf{w}) = (\partial/\partial \mathbf{w}) \mathbf{g}(\mathbf{w})'$ . Thus

$$\widehat{\boldsymbol{\theta}}_{(-i)} - \bar{\boldsymbol{\theta}} \approx -\frac{1}{n-1} \mathbf{G}(\bar{\mathbf{w}})' (\mathbf{w}_i - \bar{\mathbf{w}})$$

and the jackknife estimator of the variance of  $\widehat{\boldsymbol{\theta}}$  approximately equals

$$\begin{aligned}\widehat{V}_{\widehat{\boldsymbol{\theta}}}^{\text{jack}} &= \frac{n-1}{n} \sum_{i=1}^n (\widehat{\boldsymbol{\theta}}_{(-i)} - \widehat{\boldsymbol{\theta}}_{(.)}) (\widehat{\boldsymbol{\theta}}_{(-i)} - \widehat{\boldsymbol{\theta}}_{(.)})' \\ &\approx \frac{n-1}{n} \mathbf{G}(\bar{\mathbf{w}})' \left( \frac{1}{(n-1)^2} \sum_{i=1}^n (\mathbf{w}_i - \bar{\mathbf{w}}) (\mathbf{w}_i - \bar{\mathbf{w}})' \right) \mathbf{G}(\bar{\mathbf{w}}) \\ &= \mathbf{G}(\bar{\mathbf{w}})' \widehat{V}_{\bar{\mathbf{w}}}^{\text{jack}} \mathbf{G}(\bar{\mathbf{w}})\end{aligned}$$

where  $\widehat{V}_{\bar{\mathbf{w}}}^{\text{jack}}$  as defined in (10.3) is the conventional (and jackknife) estimator for the variance of  $\bar{\mathbf{w}}$ . Thus  $\widehat{V}_{\widehat{\boldsymbol{\theta}}}^{\text{jack}}$  is approximately the delta-method estimator. Once again, we see that the jackknife estimator automatically calculates what is effectively the delta-method variance estimator, but without requiring the user to explicitly calculate the derivative of  $\mathbf{g}(\mathbf{w})$ .

## 10.4 Example

We illustrate by reporting the asymptotic and jackknife standard errors for the four parameters given earlier. In Table 10.1 we report the actual values of the leave-one-out estimates for each of the twenty observations in the sample. The jackknife standard errors are calculated as the scaled square roots of the sample variances of these leave-one-out estimates, and are reported in the second-to-last row. For comparison the asymptotic standard errors are reported in the final row.

For all estimators the jackknife and asymptotic standard errors are quite similar. This reinforces the credibility of both standard error estimates. The largest differences arise for  $\widehat{\beta}_2$  and  $\widehat{\mu}$ , whose jackknife standard errors are about 5% larger than the asymptotic standard errors.

The take-away from our presentation is that the jackknife is a simple and flexible method for variance and standard error calculation. Circumventing technical asymptotic and exact calculations, the jackknife produces estimates which in many cases are very similar to asymptotic delta-method counterparts. The jackknife is especially appealing in cases where asymptotic standard errors are not available or are difficult to calculate. They can also be used as a double-check on the reasonability of asymptotic delta-method calculations.

In Stata, jackknife standard errors for coefficient estimates in many models are simply obtained by the `vce(jackknife)` option. For nonlinear functions of the coefficients or other estimators, the `jackknife` command can be combined with any other command to obtain jackknife standard errors.

To illustrate, below we list the Stata commands which will calculate the jackknife standard errors listed above. The first line is least squares estimation with standard errors calculated by the jackknife. The second line calculates the error variance estimate  $\widehat{\sigma}^2$  with a jackknife standard error. The third line does the same for the estimate  $\widehat{\mu}$ .

### Stata Commands

```
reg wage education if mbf12 == 1, vce(jackknife)
jackknife (e(rss)/e(N)): reg wage education if mbf12 == 1
jackknife exp(16*_b[education]+_b[_cons]+e(rss)/e(N)/2): ///
reg wage education if mbf12 == 1
```

Table 10.1: Leave-one-out Estimators and Jackknife Standard Errors

Observation	$\hat{\beta}_{1(-i)}$	$\hat{\beta}_{2(-i)}$	$\hat{\sigma}_{(-i)}^2$	$\hat{\mu}_{(-i)}$
1	0.150	0.764	0.150	25.63
2	0.148	0.798	0.149	25.48
3	0.153	0.739	0.151	25.97
4	0.156	0.695	0.144	26.31
5	0.154	0.701	0.146	25.38
6	0.158	0.655	0.151	26.05
7	0.152	0.705	0.114	24.32
8	0.146	0.822	0.147	25.37
9	0.162	0.588	0.151	25.75
10	0.157	0.693	0.139	26.40
11	0.168	0.510	0.141	26.40
12	0.158	0.691	0.118	26.48
13	0.139	0.974	0.141	26.56
14	0.169	0.451	0.131	26.26
15	0.146	0.852	0.150	24.93
16	0.156	0.696	0.148	26.06
17	0.165	0.513	0.140	25.22
18	0.155	0.698	0.151	25.90
19	0.152	0.742	0.151	25.73
20	0.155	0.697	0.151	25.95
$s^{\text{jack}}$	0.032	0.514	0.046	2.39
$s^{\text{asy}}$	0.031	0.493	0.043	2.29

## 10.5 Jackknife for Clustered Observations

In Section 4.21 we introduced the clustered regression model, cluster-robust variance estimators, and cluster-robust standard errors. Jackknife variance estimation can also be used for clustered samples, but with some natural modifications. Recall that the least-squares estimator in the clustered sample context can be written as

$$\hat{\boldsymbol{\beta}} = \left( \sum_{g=1}^G \mathbf{X}_g' \mathbf{X}_g \right)^{-1} \left( \sum_{g=1}^G \mathbf{X}_g' \mathbf{y}_g \right)$$

where  $g = 1, \dots, G$  indexes the cluster. Instead of leave-one-out estimators, it is natural to use delete-cluster estimators, which delete one cluster at a time. They take the form (4.52):

$$\hat{\boldsymbol{\beta}}_{(-g)} = \hat{\boldsymbol{\beta}} - (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_g' \tilde{\mathbf{e}}_g$$

where

$$\begin{aligned} \tilde{\mathbf{e}}_g &= \left( \mathbf{I}_{n_g} - \mathbf{X}_g (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_g' \right)^{-1} \hat{\mathbf{e}}_g \\ \hat{\mathbf{e}}_g &= \mathbf{y}_g - \mathbf{X}_g \hat{\boldsymbol{\beta}}. \end{aligned}$$

The delete-cluster jackknife estimator of the variance of  $\hat{\boldsymbol{\beta}}$  is

$$\begin{aligned} \hat{V}_{\hat{\boldsymbol{\beta}}}^{\text{jack}} &= \frac{G-1}{G} \sum_{g=1}^G \left( \hat{\boldsymbol{\beta}}_{(-g)} - \bar{\boldsymbol{\beta}} \right) \left( \hat{\boldsymbol{\beta}}_{(-g)} - \bar{\boldsymbol{\beta}} \right)' \\ \bar{\boldsymbol{\beta}} &= \frac{1}{G} \sum_{g=1}^G \hat{\boldsymbol{\beta}}_{(-g)}. \end{aligned}$$

We can also call  $\widehat{V}_{\widehat{\beta}}^{\text{jack}}$  a cluster-robust jackknife estimator of variance.

Using the same approximations as the previous section, we can show that the delete-cluster jackknife estimator is asymptotically equivalent to the cluster-robust covariance matrix estimator (4.53) calculated with the delete-cluster prediction errors. This verifies that the delete-cluster jackknife is the appropriate jackknife approach for clustered dependence.

For parameters which are functions  $\widehat{\boldsymbol{\theta}} = \mathbf{r}(\widehat{\boldsymbol{\beta}})$  of the least-squares estimator, the delete-cluster jackknife estimator of the variance of  $\widehat{\boldsymbol{\theta}}$  is

$$\begin{aligned}\widehat{V}_{\widehat{\boldsymbol{\theta}}}^{\text{jack}} &= \frac{G-1}{G} \sum_{g=1}^G (\widehat{\boldsymbol{\theta}}_{(-g)} - \bar{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_{(-g)} - \bar{\boldsymbol{\theta}})' \\ \widehat{\boldsymbol{\theta}}_{(-i)} &= \mathbf{r}(\widehat{\boldsymbol{\beta}}_{(-g)}) \\ \bar{\boldsymbol{\theta}} &= \frac{1}{G} \sum_{g=1}^G \widehat{\boldsymbol{\theta}}_{(-g)}.\end{aligned}$$

Using a mean-value expansion, we can show that this estimator is asymptotically equivalent to the delta-method cluster-robust covariance matrix estimator for  $\widehat{\boldsymbol{\theta}}$ . This shows that the jackknife estimator is appropriate for covariance matrix estimation.

As in the context of i.i.d. samples, one advantage of the jackknife covariance matrix estimators is that they do not require the user to make a technical calculation of the asymptotic distribution. A downside is an increase in computation cost, as  $G$  separate regressions are effectively estimated.

In Stata, jackknife standard errors for coefficient estimates with clustered observations are obtained by using the options `cluster(id)` `vce(jackknife)` where `id` denotes the cluster variable.

## 10.6 Empirical Distribution Function

Recall that the distribution function of a random variable  $y$  is  $F(u) = \mathbb{P}(y \leq u) = \mathbb{E}(\mathbf{1}(y \leq u))$ . Given a sample  $\{y_1, \dots, y_n\}$  of observations from  $F$ , the method of moments estimator for  $F(u)$  is the fraction of observations less than or equal to  $u$ .

$$F_n(u) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(y_i \leq u).$$

The function  $F_n(u)$  is called the **empirical distribution function** (EDF).

For any sample, the EDF is a valid distribution function. (It is non-decreasing, right-continuous, and limits to 0 and 1.) It is the discrete distribution which puts probability mass  $1/n$  on each observation. It is a nonparametric estimator, as it uses no prior information about the distribution function  $F(u)$ . Note that while  $F(u)$  may be either discrete or continuous,  $F_n(u)$  is by construction a step function.

The distribution function of a random vector  $\mathbf{y}$  is  $F(\mathbf{u}) = \mathbb{P}(\mathbf{y} \leq \mathbf{u}) = \mathbb{E}(\mathbf{1}(\mathbf{y} \leq \mathbf{u}))$ , where the inequalities apply to all elements of the vector. The EDF for a sample  $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$  is

$$F_n(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(\mathbf{y}_i \leq \mathbf{u}).$$

As for scalar variables, the multivariate EDF is a valid distribution function, and is the probability distribution which puts probability mass  $1/n$  at each observation.

The EDF  $F_n(\mathbf{u})$  is a consistent estimator of the distribution function  $F(\mathbf{u})$ . To see this, note that for any  $\mathbf{u}$ ,  $\mathbf{1}(\mathbf{y}_i \leq \mathbf{u})$  is a bounded i.i.d. random variable with expectation  $F(\mathbf{u})$ . Thus by the WLLN (Theorem 6.2),  $F_n(\mathbf{u}) \xrightarrow{P} F(\mathbf{u})$ . Furthermore, it is consistent uniformly over  $\mathbf{u}$ .

**Theorem 10.1** (Glivenko-Cantelli) If  $y_i$  are i.i.d., as  $n \rightarrow \infty$

$$\sup_{\mathbf{u}} |F_n(\mathbf{u}) - F(\mathbf{u})| \xrightarrow{P} 0.$$

The proof is presented in Section 10.33.

Theorem 10.1 is a famous example of **functional convergence**. You can view  $F_n(\cdot)$  as a functional estimate of  $F(\cdot)$ , and then ask in which sense does  $F_n$  converge to  $F$ . Theorem 10.1 shows that it converges in the uniform metric. Specifically, the uniform metric is the largest discrepancy between two functions:

$$\rho(f, g) = \sup_{\mathbf{u}} |f(\mathbf{u}) - g(\mathbf{u})|.$$

Theorem 10.1 shows that  $\rho(F_n, F) \xrightarrow{P} 0$ . The Glivenko-Cantelli Theorem was the first case of functional convergence established in the statistics literature. This is the foundation for an important class of convergence concepts known as empirical process theory and Donsker classes.

## 10.7 Quantiles

Quantiles are a useful representation of a distribution.

**Definition 10.1** For any  $\alpha \in (0, 1]$  the  $\alpha^{th}$  quantile of a distribution  $F(u)$  is  $q_\alpha = \inf\{u : F(u) \geq \alpha\}$ .

When  $F(u)$  is strictly increasing then  $q_\alpha$  satisfies  $F(q_\alpha) = \alpha$ , and is thus the “inverse” of the distribution function. In this case we can write  $q_\alpha = F^{-1}(\alpha)$ .

One way to think about a quantile is that it is the point which splits the probability mass so that  $100\alpha\%$  of the distribution is to the left of  $q_\alpha$  and  $100(1 - \alpha)\%$  is to the right of  $q_\alpha$ .

Only univariate quantiles are defined; there is not a multivariate version.

A related concept are percentiles, which are expressed in terms of percentages. For any  $\alpha$ , the  $\alpha^{th}$  quantile and  $100\alpha^{th}$  percentile are identical.

The empirical analog of  $q_\alpha$  given a univariate sample  $\{y_1, \dots, y_n\}$  is the **empirical quantile**, which is obtained by replacing  $F(u)$  with the empirical distribution function  $F_n(u)$ . Thus

$$\hat{q}_\alpha = \inf\{u : F_n(u) \geq \alpha\}.$$

It turns out that this can be written as a simple order statistic. (The **order statistics** of the sample are the observations arranged in increasing order  $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$ .) Note that  $F_n(y_{(j)}) \geq j/n$  (with equality if the sample values are unique.) Set  $j = \lceil n\alpha \rceil$ , the value  $n\alpha$  rounded up to the nearest integer (also known as the ceiling function). Thus  $F_n(y_{(j)}) \geq j/n \geq \alpha$ . For any  $u < y_{(j)}$ ,  $F_n(u) \leq (j-1)/n < \alpha$ . Thus  $y_{(j)}$  is the  $\alpha^{th}$  empirical quantile.

**Theorem 10.2**  $\hat{q}_\alpha = y_{(j)}$ , where  $j = \lceil n\alpha \rceil$ .

To illustrate, consider estimation of the median wage from the dataset reported in Table 3.1. In this example,  $n = 20$  and  $\alpha = 0.5$ . Thus  $n\alpha = 10$  is an integer. The 10<sup>th</sup> order statistic for the wage (the 10<sup>th</sup> smallest observed wage) is  $wage_{(10)} = 23.08$ . This is the empirical median  $\hat{q}_{0.5} = 23.08$ . To estimate the 0.66 quantile of this distribution,  $n\alpha = 13.2$ , so we round up to 14. The 14<sup>th</sup> order statistic (and empirical 0.66 quantile) is  $\hat{q}_{0.66} = 31.73$ .

A useful property of quantiles and sample quantiles is that they are **equivariant to monotone transformations**. Specifically, let  $h(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$  be nondecreasing and set  $w = h(y)$ . Let  $q_\alpha^y$  and  $q_\alpha^w$  be the quantile functions of  $y$  and  $w$ . The equivariance property is  $q_\alpha^w = h(q_\alpha^y)$ . That is, the quantiles of  $w$  are the transformations of the quantiles of  $y$ . For example, the  $\alpha^{th}$  quantile of  $\log(y)$  is the log of the  $\alpha^{th}$  quantile of  $y$ .

To illustrate with our empirical example, the log of the median wage is  $\log(\hat{q}_{0.5}) = \log(23.08) = 3.14$ . This equals the 10<sup>th</sup> order statistic of the log(Wage) observations. The two are identical because of the equivariance property.

The quantile estimator is consistent for  $q_\alpha$  when  $F(u)$  is strictly increasing.

**Theorem 10.3** If  $y_i$  are i.i.d. and  $F(u)$  is strictly increasing at  $q_\alpha$  then  $\hat{q}_\alpha \xrightarrow{P} q_\alpha$  as  $n \rightarrow \infty$ .

Theorem 10.3 is a special case of Theorem 10.4 presented below, so its proof is omitted. The assumption that  $F(u)$  is strictly increasing at  $q_\alpha$  excludes discrete distributions and those with flat sections.

For most users, the above information is sufficient to understand and work with quantiles. However, for completeness we now give a few more details.

While Definition 10.1 is convenient because it defines quantiles uniquely, it may be more insightful to define the **quantile interval** as the set of solutions to  $\alpha = F(q_\alpha)$ . To handle this rigorously it is useful to define the left limit version of the probability function,  $F^+(q) = \mathbb{P}(y < u)$ . We can then define the  $\alpha^{th}$  quantile interval as the set of numbers  $q$  which satisfy  $F^+(q) \leq \alpha \leq F(q)$ . This equals  $[q_\alpha, q_\alpha^+]$  where  $q_\alpha$  is from Definition 10.1 and  $q_\alpha^+ = \sup \{u : F^+(u) \leq \alpha\}$ . We have the equality  $q_\alpha^+ = q_\alpha$  when  $F(u)$  is strictly increasing (in both directions) at  $q_\alpha$ .

We can similarly extend the definition of the empirical quantile. The empirical analog of the interval  $[q_\alpha, q_\alpha^+]$  is the empirical quantile interval  $[\hat{q}_\alpha, \hat{q}_\alpha^+]$  where  $\hat{q}_\alpha$  is the empirical quantile defined earlier and  $\hat{q}_\alpha^+ = \sup \{u : F_n^+(u) \leq \alpha\}$  where  $F_n^+(u) = \frac{1}{n} \sum_{i=1}^n 1(y_i < u)$ . We can calculate that when  $n\alpha$  is an integer then  $\hat{q}_\alpha^+ = y_{(j+1)}$  where  $j = \lceil n\alpha \rceil$  but otherwise  $\hat{q}_\alpha^+ = y_{(j)}$ . Thus when  $n\alpha$  is an integer the empirical quantile interval is  $[y_{(j)}, y_{(j+1)}]$ , and otherwise is the unique value  $y_{(j)}$ .

A number of estimators for  $q_\alpha$  have been proposed and been implemented in standard software. We will describe four of these estimators, using the labeling system expressed in the R documentation.

The Type 1 estimator is the empirical quantile,  $\hat{q}_\alpha^1 = \hat{q}_\alpha$ .

The Type 2 estimator takes the midpoint of the empirical quantile interval  $[\hat{q}_\alpha, \hat{q}_\alpha^+]$ . Thus the estimator is  $\hat{q}_\alpha^2 = (y_{(j)} + y_{(j+1)})/2$  when  $n\alpha$  is an integer, and  $y_{(j)}$  otherwise. This is the method implemented in Stata. Quantiles can be obtained by the `summarize`, `detail`, `xtile`, `pctile`, and `_pctile` commands.

The Type 5 estimator defines  $m = n\alpha + 0.5$ ,  $\ell = \text{int}(m)$  (integer part), and  $r = m - \ell$  (remainder). It then sets  $\hat{q}_\alpha^5$  as a weighted average of  $y_{(\ell)}$  and  $y_{(\ell+1)}$ , using the interpolating weights  $1 - r$  and  $r$ , respectively. This is the method implemented in Matlab, and can be obtained by the `quantile` command.

The Type 7 estimator defines  $m = n\alpha + 1 - \alpha$ ,  $\ell = \text{int}(m)$ , and  $r = m - \ell$ . It then sets  $\hat{q}_\alpha^7$  as a weighted average of  $y_{(\ell)}$  and  $y_{(\ell+1)}$ , using the interpolating weights  $1 - r$  and  $r$ , respectively. This is the default method implemented in R, and can be obtained by the `quantile` command. The other methods (including Types 1, 2, and 5) can be obtained in R by specifying the Type as an option.

The Type 5 and 7 estimators may not be immediately intuitive. What they implement is to first smooth the empirical distribution function by interpolation, thus creating a strictly increasing estimator, and then inverting the interpolated EDF to obtain the corresponding quantile. The two methods differ

in terms of how they implement interpolation. The estimates lie in the interval  $[y_{(j-1)}, y_{(j+1)}]$  where  $j = \lceil n\alpha \rceil$ , but do not necessarily lie in the empirical quantile interval  $[\hat{q}_\alpha, \hat{q}_\alpha^+]$ .

To illustrate, consider again estimation of the median wage from Table 3.1. The 10<sup>th</sup> and 11<sup>th</sup> order statistics are 23.08 and 24.04, respectively, and  $n\alpha = 10$  is an integer, so the empirical quantile interval for the median is [23.08, 24.04]. The point estimates are  $\hat{q}_\alpha^1 = 23.08$  and  $\hat{q}_\alpha^2 = \hat{q}_\alpha^5 = \hat{q}_\alpha^7 = 23.56$ .

Consider the 0.66 quantile. The point estimates are  $\hat{q}_\alpha^1 = \hat{q}_\alpha^2 = 31.73$ ,  $\hat{q}_\alpha^5 = 31.15$ , and  $\hat{q}_\alpha^7 = 30.85$ . Note that the latter two are smaller than the empirical quantile 31.73.

The differences can be greatest at the extreme quantiles. Consider the 0.95 quantile. The empirical quantile is the 19<sup>th</sup> order statistic  $\hat{q}_\alpha^1 = 43.08$ .  $\hat{q}_\alpha^2 = \hat{q}_\alpha^5 = 48.85$  is average of the 19<sup>th</sup> and 20<sup>th</sup> order statistics, and  $\hat{q}_\alpha^7 = 43.65$ .

The differences between the methods diminish in large samples. However, it is useful to know that the packages implement distinct estimates when comparing results across packages.

Theorem 10.3 can be generalized to allow for interval-valued quantiles. To do so we need a convergence concept for interval-valued parameters.

**Definition 10.2** We say that a random variable  $z_n$  **converges in probability** to the interval  $[a, b]$  with  $a \leq b$ , as  $n \rightarrow \infty$ , denoted  $z_n \xrightarrow{P} [a, b]$ , if for all  $\varepsilon > 0$

$$\mathbb{P}(a - \varepsilon \leq z_n \leq b + \varepsilon) \xrightarrow{P} 1.$$

This is the natural extension of the concept of convergence in probability to interval-valued parameters. It says that the variable  $z_n$  lies within  $\varepsilon$  of the interval  $[a, b]$  with probability approaching one.

The following result includes the quantile estimators described above (Types 1, 2, 5, and 7.)

**Theorem 10.4** Let  $\hat{q}_\alpha$  be any estimator satisfying  $y_{(j-1)} \leq \hat{q}_\alpha \leq y_{(j+1)}$  where  $j = \lceil n\alpha \rceil$ . If  $y_i$  are i.i.d. and  $0 < \alpha < 1$ , then  $\hat{q}_\alpha \xrightarrow{P} [q_\alpha, q_\alpha^+]$  as  $n \rightarrow \infty$ .

The proof is presented in Section 10.33. Theorem 10.4 applies to all distribution functions, including continuous and discrete.

## 10.8 The Bootstrap Algorithm

The bootstrap is a powerful approach to inference, and is due to the pioneering work of Efron (1979). There are many textbook and monograph treatments of the bootstrap, including Efron (1982), Hall (1992), Efron and Tibshirani (1993), Shao and Tu (1995), and Davison and Hinkley (1997). Reviews for econometricians are provided by Hall (1994) and Horowitz (2001).

There are several ways to describe or define the bootstrap, and there are several forms of the bootstrap. We start in this section by describing the basic nonparametric bootstrap algorithm. In subsequent sections we give more formal definitions of the bootstrap as well as theoretical justifications.

Briefly, the bootstrap distribution is obtained by estimation on independent samples created by i.i.d. sampling (sampling with replacement) from the original dataset.

To understand this, it is useful to start with the concept of sampling with replacement from the dataset. To continue the empirical example used earlier in the chapter, we focus on the dataset displayed in Table 3.1, which has  $n = 20$  observations. Sampling from this distribution means randomly selecting one row from this table. Mathematically this is the same as randomly selecting an integer from the set  $\{1, 2, \dots, 20\}$ . To illustrate, Matlab has a random integer generator (the function `randi`), and using

the random number seed of 13 (an arbitrary choice) we obtain the random draw 16. This means that we draw observation number 16 from Table 3.1. Examining the table, we can see that this is an individual with wage \$18.75 and education of 16 years. We repeat by drawing another random integer on the set  $\{1, 2, \dots, 20\}$  and this time obtain 5. This means we take observation 5 from Table 3.1, which is an individual with wage \$33.17 and education of 16 years. We continue until we have  $n = 20$  such draws. This random set of observations are  $\{16, 5, 17, 20, 20, 10, 13, 16, 13, 15, 1, 6, 2, 18, 8, 14, 6, 7, 1, 8\}$ . We call this the **bootstrap sample**.

Notice that the observations 1, 6, 8, 13, 16, 20 each appear twice in the bootstrap sample, and the observations 3, 4, 9, 11, 12, 19 do not appear at all. That is okay. In fact, it is necessary for the bootstrap to work. This is because we are *drawing with replacement*. (If we instead made draws without replacement, then the constructed dataset would have exactly the same observations as in Table 3.1, only in different order.) We can also ask the question “What is the probability that an individual observation will appear at least once in the bootstrap sample? The answer is

$$\begin{aligned}\mathbb{P}(\text{Observation in Bootstrap Sample}) &= 1 - \left(1 - \frac{1}{n}\right)^n \\ &\longrightarrow 1 - e^{-1} \\ &\simeq 0.632.\end{aligned}\tag{10.6}$$

The limit holds as  $n \rightarrow \infty$ . The approximation 0.632 is excellent even for small  $n$ . Indeed, for our example with  $n = 20$  the probability (10.6) is 0.641. These calculations show that an individual observation is in the bootstrap sample with probability near 2/3, and is not in the bootstrap sample with probability near 1/3.

Once again, the bootstrap sample is the constructed dataset with the 20 observations drawn randomly from the original sample. Notationally, we write the  $i^{\text{th}}$  bootstrap observation as  $(y_i^*, \mathbf{x}_i^*)$  and the bootstrap sample as  $\{(y_1^*, \mathbf{x}_1^*), \dots, (y_n^*, \mathbf{x}_n^*)\}$ . In our present example with  $y$  denoting the log wage, the bootstrap sample is

$$\{(y_1^*, \mathbf{x}_1^*), \dots, (y_n^*, \mathbf{x}_n^*)\} = \{(2.93, 16), (3.50, 16), \dots, (3.76, 18)\}.$$

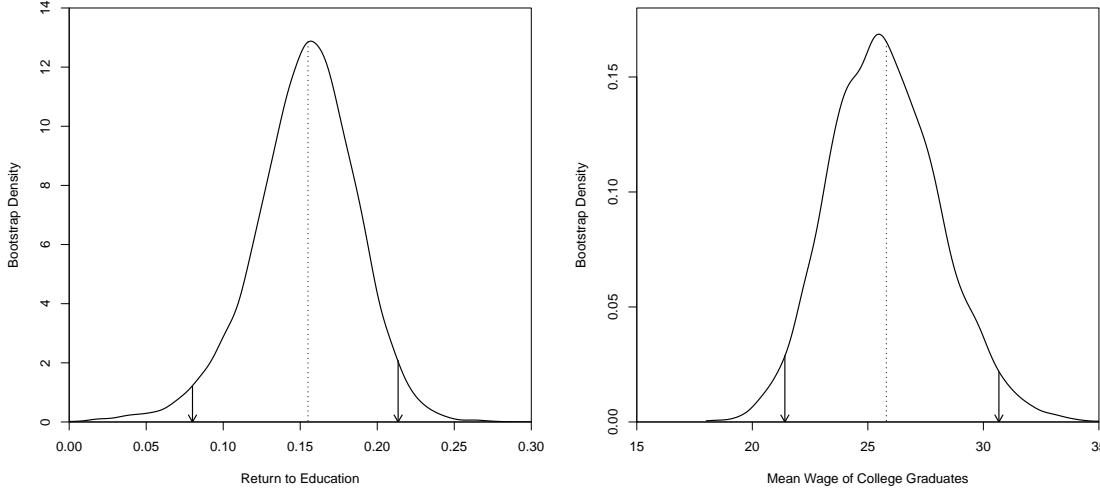
The bootstrap estimate  $\hat{\boldsymbol{\beta}}^*$  is then obtained applying the least-squares estimation formula to the bootstrap sample. Thus we regress  $y_i^*$  on  $\mathbf{x}_i^*$ . The other bootstrap estimates, in our example  $\hat{\sigma}^{2*}$  and  $\hat{\mu}^*$ , are obtained by applying the estimation formula to the bootstrap sample as well. Writing  $\hat{\boldsymbol{\theta}}^* = (\hat{\beta}_1^*, \hat{\beta}_2^*, \hat{\sigma}^{2*}, \hat{\mu}^*)'$  we have the bootstrap estimate of the parameter vector  $\boldsymbol{\theta} = (\beta_1, \beta_2, \sigma^2, \mu)'$ . In our example (the bootstrap sample described above)  $\hat{\boldsymbol{\theta}}^* = (0.195, 0.113, 0.107, 26.7)'$ . This is one draw from the bootstrap distribution of the estimates.

The estimate  $\hat{\boldsymbol{\theta}}^*$  as described is one random draw from the distribution of estimates obtained by i.i.d. sampling from the original data. With one draw we can say relatively little. But we can repeat this exercise to obtain multiple draws from this bootstrap distribution. To distinguish between these draws we index the bootstrap samples by  $b = 1, \dots, B$ , and write the bootstrap estimates as  $\hat{\boldsymbol{\theta}}_b^*$  or  $\hat{\boldsymbol{\theta}}^*(b)$ .

To continue our illustration, we draw 20 more random integers  $\{19, 5, 7, 19, 1, 2, 13, 18, 1, 15, 17, 2, 14, 11, 10, 20, 1, 5, 15, 7\}$  and construct a second bootstrap sample. On this sample we again estimate the parameters, and obtain  $\hat{\boldsymbol{\theta}}^*(2) = (0.175, 0.52, 0.124, 29.3)'$ . This is a second random draw from the distribution of  $\hat{\boldsymbol{\theta}}^*$ . We repeat this  $B$  times, storing the parameter estimates  $\hat{\boldsymbol{\theta}}^*(b)$ . We have thus created a new dataset of bootstrap draws  $\{\hat{\boldsymbol{\theta}}^*(b) : b = 1, \dots, B\}$ . By construction, the draws are independent across  $b$  and identically distributed.

The number of bootstrap draws,  $B$ , is often called the “number of bootstrap replications”. Typical choices for  $B$  are 1000, 5000, and 10,000. We discuss selecting  $B$  later, but roughly speaking, larger  $B$  results in a more precise estimate at an increased computation cost. For our application we set  $B = 10,000$ .

To illustrate, Figure 13.1 displays the densities of the distributions of the bootstrap estimates  $\hat{\beta}_1^*$  and  $\hat{\mu}^*$  across 10,000 draws. The dotted lines show the point estimate. You can notice that the density for  $\hat{\beta}_1^*$  is slightly skewed to the left.

Figure 10.1: Bootstrap Distributions of  $\hat{\beta}_1^*$  and  $\hat{\mu}^*$ 

## 10.9 Bootstrap Variance and Standard Errors

Given the bootstrap draws we can estimate features of the bootstrap distribution. The **bootstrap estimator of variance** of an estimator  $\hat{\theta}$  is the sample variance across the bootstrap draws  $\hat{\theta}^*(b)$ . It equals

$$\begin{aligned}\hat{V}_{\hat{\theta}}^{\text{boot}} &= \frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}^*(b) - \bar{\theta}^*) (\hat{\theta}^*(b) - \bar{\theta}^*)' \\ \bar{\theta}^* &= \frac{1}{B} \sum_{b=1}^B \hat{\theta}^*(b).\end{aligned}\quad (10.7)$$

For a scalar estimator  $\hat{\theta}$  the **bootstrap standard error** is the square root of the bootstrap estimator of variance:

$$s_{\hat{\theta}}^{\text{boot}} = \sqrt{\hat{V}_{\hat{\theta}}^{\text{boot}}}.$$

This is a very simple statistic to calculate, and is the most common use of the bootstrap in applied econometric practice. A caveat (discussed in more detail in Section 10.17) is that in many cases it is better to use a trimmed estimator.

Standard errors are conventionally reported to convey the precision of the estimator. They are also commonly used to construct confidence intervals. Bootstrap standard errors can be used for this purpose. The **normal-approximation bootstrap confidence interval** is

$$C^{\text{nb}} = [\hat{\theta} - z_{1-\alpha/2} s_{\hat{\theta}}^{\text{boot}}, \quad \hat{\theta} + z_{1-\alpha/2} s_{\hat{\theta}}^{\text{boot}}]$$

where  $z_{1-\alpha/2}$  is the  $1 - \alpha/2$  quantile of the  $N(0, 1)$  distribution. This interval  $C^{\text{nb}}$  is identical in format to an asymptotic confidence interval, but with the bootstrap standard error replacing the asymptotic standard error.  $C^{\text{nb}}$  is the default confidence interval reported by Stata when the bootstrap has been used to calculate standard errors. However, the normal-approximation interval is in general a poor choice for confidence interval construction as it relies on the normal approximation to the t-ratio which can be inaccurate in finite samples. There are other methods – such as the bias-corrected percentile method to be discussed in Section 10.19 – which are just as simple to compute but have better performance. In general, bootstrap standard errors should be used as estimates of precision rather than as tools to construct confidence intervals.

Since  $B$  is finite, all bootstrap statistics, such as  $\widehat{V}_{\widehat{\theta}}^{\text{boot}}$ , are estimates and hence random. Their values will vary across different choices for  $B$  and simulation runs (depending on how the simulation seed is set). Thus you should not expect to obtain the exact same bootstrap standard errors as other researchers when replicating their results. They should be similar (up to simulation sampling error) but not precisely the same.

In Table 10.2 we report the four parameter estimates introduced in Section 10.2, along with asymptotic, jackknife and bootstrap standard errors. We also report four bootstrap confidence intervals which will be introduced in subsequent sections.

For these four estimators, we can see that the bootstrap standard errors are quite similar to the asymptotic and jackknife standard errors. The most noticeable difference arises for  $\widehat{\beta}_2$ , where the bootstrap standard error is about 10% larger than the asymptotic standard error.

Table 10.2: Comparison of Methods

	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\sigma}^2$	$\widehat{\mu}$
Estimate	0.155	0.698	0.144	25.80
Asymptotic s.e.	(0.031)	(0.493)	(0.043)	(2.29)
Jackknife s.e.	(0.032)	(0.514)	(0.046)	(2.39)
Bootstrap s.e.	(0.034)	(0.548)	(0.041)	(2.38)
95% Percentile Interval	[0.08, 0.21]	[-0.27, 1.91]	[0.06, 0.22]	[21.4, 30.7]
95% BC Percentile Interval	[0.08, 0.21]	[-0.25, 1.93]	[0.09, 0.28]	[22.0, 31.5]
95% BC <sub>a</sub> Percentile Interval	[0.08, 0.21]	[-0.25, 1.93]	[0.09, 0.28]	[22.0, 31.5]
95% Percentile-t Interval	[0.09, 0.21]	[-0.20, 1.81]	[0.08, 0.34]	[21.6, 32.2]

In Stata, bootstrap standard errors for coefficient estimates in many models are simply obtained by the `vce(bootstrap, reps(#))` option, where  $#$  is the number of bootstrap replications. For nonlinear functions of the coefficients or other estimators, the `bootstrap` command can be combined with any other command to obtain bootstrap standard errors. Synonyms for `bootstrap` are `bstrap` and `bs`.

To illustrate, below we list the Stata commands which will calculate<sup>1</sup> the bootstrap standard errors listed above.

#### Stata Commands

```
reg wage education if mbf12 == 1, vce(bootstrap, reps(10000))
bs (e(rss)/e(N)), reps(10000): reg wage education if mbf12 == 1
bs (exp(16*_b[education]+_b[_cons]+e(rss)/e(N)/2)), reps(10000): ///
    reg wage education if mbf12 == 1
```

## 10.10 Percentile Interval

The second most common use of bootstrap methods is for confidence intervals. There are multiple bootstrap methods to form confidence intervals. A popular and simple method is called the **percentile interval**. It is based on the quantiles of the bootstrap distribution.

In Section 10.8 we described the bootstrap algorithm, which creates an i.i.d. sample of bootstrap estimates  $\{\widehat{\theta}_1^*, \widehat{\theta}_2^*, \dots, \widehat{\theta}_B^*\}$  corresponding to an estimator  $\widehat{\theta}$  of a parameter  $\theta$ . We focus on the case of a scalar parameter  $\theta$ .

<sup>1</sup>They will not *precisely* replicate the standard errors, since those in Table 10.2 were produced in Matlab, which uses a different random number sequence.

For any  $0 < \alpha < 1$  we can calculate the empirical quantile  $q_\alpha^*$  of these bootstrap estimates. This is the number such that  $n\alpha$  bootstrap estimates are smaller than  $q_\alpha^*$ , and typically calculated by taking the  $n\alpha^{th}$  order statistic of the  $\hat{\theta}_b^*$ . See Section 10.7 for a precise discussion of empirical quantiles and common quantile estimators.

The percentile bootstrap  $100(1 - \alpha)\%$  confidence interval is

$$C^{\text{pc}} = [q_{\alpha/2}^*, q_{1-\alpha/2}^*]. \quad (10.8)$$

For example, if  $B = 1000$ ,  $\alpha = 0.05$ , and the empirical quantile estimator is used, then  $C^{\text{pc}} = [\hat{\theta}_{(25)}^*, \hat{\theta}_{(975)}^*]$ .

To illustrate, the 0.025 and 0.975 quantiles of the bootstrap distributions of  $\hat{\beta}_1^*$  and  $\hat{\mu}^*$  are indicated in Figure 13.1 by the arrows. The intervals between the arrows are the 95% percentile interval.

The percentile interval has the convenience that it does not require calculation of a standard error. This is particularly convenient in contexts where asymptotic standard error calculation is complicated, burdensome, or unknown.  $C^{\text{pc}}$  is a simple by-product of the bootstrap algorithm and does not require meaningful computational cost above that required to calculate the bootstrap standard error.

The percentile interval has the useful property that it is **transformation-respecting**. The percentile interval for any monotone parameter transformation  $\phi = m(\theta)$  is simply the percentile interval for  $\theta$  mapped by  $m(\theta)$ . That is, if  $[q_{\alpha/2}^*, q_{1-\alpha/2}^*]$  is the percentile interval for  $\theta$ , then  $[m(q_{\alpha/2}^*), m(q_{1-\alpha/2}^*)]$  is the percentile interval for  $\phi$ . This property follows directly from the equivariance property of sample quantiles. Many confidence-interval methods, such as the delta-method asymptotic interval and the normal-approximation interval  $C^{\text{nb}}$ , do not share this property.

To illustrate the usefulness of the transformation-respecting property, consider the variance  $\sigma^2$ . In some cases it is useful to report the variance  $\sigma^2$ , and in other cases it is useful to report the standard deviation  $\sigma$ . Thus we may be interested in confidence intervals for  $\sigma^2$  or  $\sigma$ . To illustrate, the asymptotic 95% normal confidence interval for  $\sigma^2$  which we calculate from Table 13.2 is  $[0.060, 0.228]$ . Taking square roots we obtain an interval for  $\sigma$  of  $[0.244, 0.477]$ . Alternatively, the delta method standard error for  $\hat{\sigma} = 0.379$  is 0.057, leading to an asymptotic 95% confidence interval for  $\sigma$  of  $[0.265, 0.493]$  which is different. This shows that the delta method is not transformation-respecting. In contrast, the 95% percentile interval for  $\sigma^2$  is  $[0.062, 0.220]$  and that for  $\sigma$  is  $[0.249, 0.469]$  which is identical to the square roots of the interval for  $\sigma^2$ .

The bootstrap percentile intervals for the four estimators are reported in Table 13.2.

In Stata, percentile confidence intervals can be obtained by using the command `estat bootstrap, percentile` or the command `estat bootstrap, all` after an estimation command which calculates standard errors via the bootstrap.

## 10.11 The Bootstrap Distribution

For applications, it is often sufficient if one understands the bootstrap as an algorithm. However, for theory it is more useful to view the bootstrap as a specific estimator of the sampling distribution. For this, it is useful to introduce some additional notation.

The key is that the distribution of any estimator or statistic is determined by the distribution of the data. While the latter is unknown it can be estimated by the empirical distribution of the data. This is what the bootstrap does.

To fix notation, let  $F$  denote the distribution of an individual observation  $w$ . (In regression,  $w$  is the pair  $(y, \mathbf{x})$ .) Let  $G_n(u, F)$  denote the distribution of an estimator  $\hat{\theta}$ . That is,

$$G_n(u, F) = \mathbb{P}(\hat{\theta} \leq u | F).$$

We write the distribution  $G_n$  as a function of  $n$  and  $F$  since they (generally) affect the distribution of  $\hat{\theta}$ . We are interested in the distribution  $G_n$ . For example, we want to know its variance to calculate a standard error, or its quantiles to calculate a percentile interval.

In principle, if we knew the distribution  $F$  we should be able to determine the distribution  $G_n$ . In practice there are two barriers to implementation. The first barrier is that the calculation of  $G_n(u, F)$  is generally infeasible except in certain special cases such as the normal regression model. The second barrier is that in general we do not know  $F$ .

The bootstrap simultaneously circumvents these two barriers by two clever ideas. First, the bootstrap proposes estimation of  $F$  by the empirical distribution  $F_n$ , which is the simplest nonparametric estimator of the joint distribution of the observations. Replacing  $F$  with  $F_n$  we obtain the ideal bootstrap estimator of the distribution of  $\hat{\theta}$

$$G_n^*(u) = G_n(u, F_n). \quad (10.9)$$

$G_n^*$  is an idealized estimator of  $G_n$ . It is unknown in practice. The bootstrap proposes estimation of  $G_n^*$  by simulation. This is the bootstrap algorithm described in the previous sections. The essential idea is that simulation from  $F_n$  is sampling with replacement from the original data, and this is computationally very simple. Applying the estimation formula for  $\hat{\theta}$ , we obtain i.i.d. draws from the distribution  $G_n^*(u)$ . By making a large number  $B$  of such draws, we can estimate any feature of  $G_n^*$  of interest. The bootstrap combines these two ideas: (1) estimate  $G_n(u, F)$  by  $G_n(u, F_n)$ ; (2) estimate  $G_n(u, F_n)$  by simulation. These ideas are intertwined. Only by considering these steps together do we obtain a feasible inference method.

The way to think about the connection between  $G_n$  and  $G_n^*$  is as follows.  $G_n$  is the distribution of the estimator  $\hat{\theta}$  obtained when the observations are sampled i.i.d. from the population distribution  $F$ .  $G_n^*$  is the distribution of the same statistic, denoted  $\hat{\theta}^*$ , obtained when the observations are sampled i.i.d. from the empirical distribution  $F_n$ . It is useful to conceptualize the “universe” which separately generates the dataset and the bootstrap sample. The “sampling universe” is the population distribution  $F$ . In this universe the true parameter is  $\theta$ . The “bootstrap universe” is the empirical distribution  $F_n$ . When drawing from the bootstrap universe we are treating  $F_n$  as if it is the true distribution. Thus anything which is true about  $F_n$  should be treated as true in the bootstrap universe. In the bootstrap universe, the “true” value of the parameter  $\theta$  is the value determined by the EDF  $F_n$ . In most cases this is the estimate  $\hat{\theta}$ . It is the true value of the coefficient when the true distribution is  $F_n$ .

We now carefully explain the connection with the bootstrap algorithm as previously described.

First, observe that sampling with replacement from the sample  $\{y_1, \dots, y_n\}$  is identical to sampling from the EDF  $F_n$ . This is because the EDF is the probability distribution which puts probability mass  $1/n$  on each observation. Thus sampling from  $F_n$  means sampling an observation with probability  $1/n$ , which is sampling with replacement.

Second, observe that the bootstrap estimator  $\hat{\theta}^*$  described here is identical to the bootstrap algorithm described in Section 10.8. That is,  $\hat{\theta}^*$  is the random vector generated by applying the estimator formula  $\hat{\theta}$  to samples obtained by random sampling from  $F_n$ .

Third, observe that the distribution of these bootstrap estimators is the bootstrap distribution (10.9). This is a precise equality. That is, the bootstrap algorithm generates i.i.d. samples from  $F_n$ , and when the estimators are applied we obtain random variables  $\hat{\theta}^*$  with the distribution  $G_n^*$ .

Fourth, observe that the bootstrap statistics described earlier – bootstrap variance, standard error, and quantiles – are estimators of the corresponding features of the bootstrap distribution  $G_n^*$ .

This discussion is meant to carefully describe why the notation  $G_n^*(u)$  is useful to help understand the properties of the bootstrap algorithm. Since  $F_n$  is the natural nonparametric estimator of the unknown distribution  $F$ ,  $G_n^*(u) = G_n(u, F_n)$  is the natural plug-in estimator of the unknown  $G_n(u, F)$ . Furthermore, since  $F_n$  is uniformly consistent for  $F$  by Theorem 10.1, we also can expect  $G_n^*(u)$  to be consistent for  $G_n(u)$ . Making this precise it a bit challenging since  $F_n$  and  $G_n$  are functions. In the next several sections we develop an asymptotic distribution theory for the bootstrap distribution based on extending classical asymptotic theory to the case of conditional distributions.

## 10.12 The Distribution of the Bootstrap Observations

Let  $y^*$  be a random draw from the sample  $\{y_1, \dots, y_n\}$ . What is the distribution of  $y^*$ ?

Since we are fixing the observations, the correct question is: What is the *conditional* distribution of  $\mathbf{y}^*$ , conditional on the observed data? The empirical distribution function  $F_n$  summarizes the information in the sample, so equivalently we are talking about the distribution conditional on  $F_n$ . Consequently we will write the bootstrap probability function and expectation as

$$\begin{aligned}\mathbb{P}^*(\mathbf{y}^* \leq x) &= \mathbb{P}(\mathbf{y}^* \leq x | F_n) \\ \mathbb{E}^*(\mathbf{y}^*) &= \mathbb{E}(\mathbf{y}^* | F_n).\end{aligned}$$

Notationally, the starred distribution and expectation are conditional given the data.

The (conditional) distribution of  $\mathbf{y}^*$  is the empirical distribution function  $F_n$ , which is a discrete distribution with mass points  $1/n$  on each observation  $\mathbf{y}_i$ . Thus even if the original data come from a continuous distribution, the bootstrap data distribution is necessarily discrete.

The (conditional) mean and variance of  $\mathbf{y}^*$  are calculated from the EDF, and equal the sample mean and variance of the data. The mean is

$$\begin{aligned}\mathbb{E}^*(\mathbf{y}^*) &= \sum_{i=1}^n \mathbf{y}_i \mathbb{P}^*(\mathbf{y}^* = \mathbf{y}_i) \\ &= \sum_{i=1}^n \mathbf{y}_i \frac{1}{n} \\ &= \bar{\mathbf{y}}\end{aligned}\tag{10.10}$$

and the variance is

$$\begin{aligned}\text{var}^*(\mathbf{y}^*) &= \mathbb{E}^*(\mathbf{y}^* \mathbf{y}^{*\prime}) - (\mathbb{E}^*(\mathbf{y}^*)) (\mathbb{E}^*(\mathbf{y}^*))' \\ &= \sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i' \mathbb{P}^*(\mathbf{y}^* = \mathbf{y}_i) - \bar{\mathbf{y}} \bar{\mathbf{y}}' \\ &= \sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i' \frac{1}{n} - \bar{\mathbf{y}} \bar{\mathbf{y}}' \\ &= \widehat{\Sigma}.\end{aligned}\tag{10.11}$$

To summarize, the conditional distribution of  $\mathbf{y}^*$ , given  $F_n$ , is the discrete distribution on  $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ , with mean  $\bar{\mathbf{y}}$  and variance matrix  $\widehat{\Sigma}$ .

We can extend this analysis to any integer moment  $r$ . Assume  $y_i$  is scalar. The  $r^{th}$  moment of  $y^*$  is

$$\mu_r^* = \mathbb{E}^*(y^*)^r = \sum_{i=1}^n y_i^r \mathbb{P}^*(y^* = y_i) = \frac{1}{n} \sum_{i=1}^n y_i^r = \widehat{\mu}_r,$$

the  $r^{th}$  sample moment. The  $r^{th}$  central moment of  $y^*$  is

$$\mu_r^* = \mathbb{E}^*(y^* - \bar{y})^r = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^r = \widehat{\mu}_r,$$

the  $r^{th}$  central sample moment. Similarly, the  $r^{th}$  cumulant of  $y^*$  is  $\kappa_r^* = \widehat{\kappa}_r$ , the  $r^{th}$  sample cumulant.

### 10.13 The Distribution of the Bootstrap Sample Mean

The bootstrap sample mean is

$$\bar{\mathbf{y}}^* = \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i^*.$$

We can calculate its (conditional) mean and variance. The mean is

$$\mathbb{E}^*(\bar{\mathbf{y}}^*) = \mathbb{E}^*\left(\frac{1}{n} \sum_{i=1}^n \mathbf{y}_i^*\right) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}^*(\mathbf{y}_i^*) = \frac{1}{n} \sum_{i=1}^n \bar{\mathbf{y}} = \bar{\mathbf{y}}.\tag{10.12}$$

using (10.10). Thus the bootstrap sample mean  $\bar{y}^*$  has a distribution centered at the sample mean  $\bar{y}$ . This is because the bootstrap observations  $y_i^*$  are drawn from the bootstrap universe, which treats the EDF as the truth, and the mean of the latter distribution is  $\bar{y}$ .

The (conditional) variance of the bootstrap sample mean is

$$\text{var}^*(\bar{y}^*) = \text{var}^*\left(\frac{1}{n} \sum_{i=1}^n y_i^*\right) = \frac{1}{n^2} \sum_{i=1}^n \text{var}^*(y_i^*) = \frac{1}{n^2} \sum_{i=1}^n \hat{\Sigma} = \frac{1}{n} \hat{\Sigma} \quad (10.13)$$

using (10.11). In the scalar case,  $\text{var}^*(\bar{y}^*) = \hat{\sigma}^2/n$ . This shows that the bootstrap variance of  $\bar{y}^*$  is precisely described by the sample variance of the original observations. Again, this is because the bootstrap observations  $y_i^*$  are drawn from the bootstrap universe.

We can extend this to any integer moment  $r$ . Assume  $y_i$  is scalar. Define the normalized bootstrap sample mean  $z_n^* = \sqrt{n}(\bar{y}^* - \bar{y})$ . Using expressions (6.12)-(6.13), the 3<sup>rd</sup> through 6<sup>th</sup> conditional moments of  $z_n^*$  are

$$\begin{aligned} \mathbb{E}^*(z_n^*)^3 &= \hat{\kappa}_3/n^{1/2} \\ \mathbb{E}^*(z_n^*)^4 &= \hat{\kappa}_4/n + 3\hat{\kappa}_2^2 \\ \mathbb{E}^*(z_n^*)^5 &= \hat{\kappa}_5/n^{3/2} + 10\hat{\kappa}_3\hat{\kappa}_2/n^{1/2} \\ \mathbb{E}^*(z_n^*)^6 &= \hat{\kappa}_6/n^2 + (15\hat{\kappa}_4\kappa_2 + 10\hat{\kappa}_3^2)/n + 15\hat{\kappa}_2^3 \end{aligned} \quad (10.14)$$

where  $\hat{\kappa}_r$  is the  $r^{th}$  sample cumulant. Similar expressions can be derived for higher moments.

The moments (10.14) are exact, not approximations.

## 10.14 Bootstrap Asymptotics

The bootstrap mean  $\bar{y}^*$  is a sample average over  $n$  i.i.d. random variables, so we might expect it to converge in probability to its expectation. Indeed, this is the case, but we have to be a bit careful since the bootstrap mean has a conditional distribution (given the data) so we need to define convergence in probability for conditional distributions.

**Definition 10.3** We say that a random vector  $z_n^*$  **converges in bootstrap probability** to  $z$  as  $n \rightarrow \infty$ , denoted  $z_n^* \xrightarrow{p^*} z$ , if for all  $\varepsilon > 0$

$$\mathbb{P}^*(\|z_n^* - z\| > \varepsilon) \xrightarrow{p} 0.$$

To understand this definition recall that conventional convergence in probability  $z_n \xrightarrow{p} z$  means that for a sufficiently large sample size  $n$ , the probability is high that  $z_n$  is arbitrarily close to its limit  $z$ . In contrast, Definition 10.3 says  $z_n^* \xrightarrow{p^*} z$  means that for a sufficiently large  $n$ , the probability is high that the conditional probability that  $z_n^*$  is close to its limit  $z$  is high. Note that there are two uses of probability – both unconditional and conditional.

Our label “convergence in bootstrap probability” is a bit unusual. The label used in much of the statistical literature is “convergence in probability, in probability” but that seems like a mouthful. That literature more often focuses on the related concept of “convergence in probability, almost surely” which holds if we replace the “ $\xrightarrow{p}$ ” convergence with almost sure convergence. We do not use this concept in this chapter as it is an unnecessary complication.

While we have stated Definition 10.3 for the specific conditional probability distribution  $\mathbb{P}^*$ , the idea is more general and can be used for any conditional distribution and any sequence of random vectors.

The following may seem obvious, but it is useful to state for clarity, and its proof is given in Section 10.33.

**Theorem 10.5** If  $z_n \xrightarrow{p} z$  as  $n \rightarrow \infty$  then  $z_n \xrightarrow{p^*} z$ .

Given Definition 10.3, we can establish a law of large numbers for the bootstrap sample mean.

**Theorem 10.6 Bootstrap WLLN.** If  $y_i$  are independent and uniformly integrable then  $\bar{y}^* - \bar{y} \xrightarrow{p^*} \mathbf{0}$  and  $\bar{y}^* \xrightarrow{p^*} \boldsymbol{\mu} = \mathbb{E}(y_i)$  as  $n \rightarrow \infty$ .

The proof (presented in Section 10.33) is somewhat different from the classical case, as it is based on the Marcinkiewicz WLLN (Theorem 6.39).

Notice that the conditions for the bootstrap WLLN are the same for the conventional WLLN. Notice as well that we state two related but slightly different results. The first is that the difference between the bootstrap sample mean  $\bar{y}^*$  and the sample mean  $\bar{y}$  diminishes as the sample size diverges. The second result is that the bootstrap sample mean converges to the population mean  $\boldsymbol{\mu}$ . The latter is not surprising (since the sample mean  $\bar{y}$  converges in probability to  $\boldsymbol{\mu}$ ) but it is constructive to be precise since we are dealing with a new convergence concept.

**Theorem 10.7 Bootstrap Continuous Mapping Theorem.** If  $z_n^* \xrightarrow{p^*} c$  as  $n \rightarrow \infty$  and  $g(\cdot)$  is continuous at  $c$ , then  $g(z_n^*) \xrightarrow{p^*} g(c)$  as  $n \rightarrow \infty$ .

The proof is essentially identical to that of Theorem 6.19, so is omitted.

We next would like to show that the bootstrap sample mean is asymptotically normally distributed, but for that we need a definition of convergence for conditional distributions.

**Definition 10.4** Let  $z_n^*$  be a random vector with conditional distribution  $G_n^*(\mathbf{u}) = \mathbb{P}^*(z_n^* \leq \mathbf{u})$ . We say that  $z_n^*$  **converges in bootstrap distribution** to  $z$  as  $n \rightarrow \infty$ , denoted  $z_n^* \xrightarrow{d^*} z$ , if for all  $\mathbf{u}$  at which  $G(\mathbf{u}) = \mathbb{P}(z \leq \mathbf{u})$  is continuous,  $G_n^*(\mathbf{u}) \xrightarrow{p} G(\mathbf{u})$  as  $n \rightarrow \infty$ .

The difference with the conventional definition is that Definition 10.4 treats the conditional distribution as random. An alternative label for Definition 10.4 is “convergence in distribution, in probability”.

We now state a CLT for the bootstrap sample mean, with a proof given in Section 10.33.

**Theorem 10.8 Bootstrap CLT.** If  $y_i$  are independent,  $\|y_i\|^2$  is uniformly integrable, and  $\Sigma = \text{var}(\mathbf{y}) > 0$  then

$$\sqrt{n}(\bar{y}^* - \bar{y}) \xrightarrow{d^*} N(\mathbf{0}, \Sigma)$$

as  $n \rightarrow \infty$ .

Theorem 10.8 shows that the normalized bootstrap sample mean has the same asymptotic distribution as the sample mean. Thus the bootstrap distribution is asymptotically the same as the sampling distribution. A notable difference, however, is that the bootstrap sample mean is normalized by centering at the sample mean, not at the population mean. This is because  $\bar{\mathbf{y}}$  is the true mean in the bootstrap universe.

We next state the distributional form of the continuous mapping theorem for bootstrap distributions and the Bootstrap Delta Method.

**Theorem 10.9 Bootstrap Continuous Mapping Theorem**

If  $\mathbf{z}_n^* \xrightarrow{d^*} \mathbf{z}$  as  $n \rightarrow \infty$  and  $\mathbf{g} : \mathbb{R}^m \rightarrow \mathbb{R}^k$  has the set of discontinuity points  $D_g$  such that  $\mathbb{P}^*(\mathbf{z}^* \in D_g) = 0$ , then  $\mathbf{g}(\mathbf{z}_n^*) \xrightarrow{d^*} \mathbf{g}(\mathbf{z})$  as  $n \rightarrow \infty$ .

**Theorem 10.10 Bootstrap Delta Method:**

If  $\hat{\boldsymbol{\mu}} \xrightarrow{p} \boldsymbol{\mu}$ ,  $\sqrt{n}(\hat{\boldsymbol{\mu}}^* - \hat{\boldsymbol{\mu}}) \xrightarrow{d^*} \boldsymbol{\xi}$ , and  $\mathbf{g}(\mathbf{u})$  is continuously differentiable in a neighborhood of  $\boldsymbol{\mu}$ , then as  $n \rightarrow \infty$

$$\sqrt{n}(\mathbf{g}(\hat{\boldsymbol{\mu}}^*) - \mathbf{g}(\hat{\boldsymbol{\mu}})) \xrightarrow{d^*} \mathbf{G}' \boldsymbol{\xi}$$

where  $\mathbf{G}(\mathbf{u}) = \frac{\partial}{\partial \mathbf{u}} \mathbf{g}(\mathbf{u})'$  and  $\mathbf{G} = \mathbf{G}(\boldsymbol{\mu})$ . In particular, if  $\boldsymbol{\xi} \sim N(\mathbf{0}, \mathbf{V})$  then as  $n \rightarrow \infty$

$$\sqrt{n}(\mathbf{g}(\hat{\boldsymbol{\mu}}^*) - \mathbf{g}(\hat{\boldsymbol{\mu}})) \xrightarrow{d^*} N(0, \mathbf{G}' \mathbf{V} \mathbf{G}).$$

For a proof, see Exercise 10.8.

We state an analog of Theorem 6.24, which presented the asymptotic distribution for general smooth functions of sample means, which covers most econometric estimators.

**Theorem 10.11** Under the assumptions of Theorem 6.24, that is, if  $\mathbf{y}_i$  is i.i.d.,  $\boldsymbol{\mu} = \mathbb{E}(\mathbf{h}(\mathbf{y}))$ ,  $\boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\mu})$ ,  $\mathbb{E}\|\mathbf{h}(\mathbf{y})\|^2 < \infty$ , and  $\mathbf{G}(\mathbf{u}) = \frac{\partial}{\partial \mathbf{u}} \mathbf{g}(\mathbf{u})'$  is continuous in a neighborhood of  $\boldsymbol{\mu}$ , for  $\hat{\boldsymbol{\theta}} = \mathbf{g}(\hat{\boldsymbol{\mu}})$  with  $\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^n \mathbf{h}(\mathbf{y}_i)$  and  $\hat{\boldsymbol{\theta}}^* = \mathbf{g}(\hat{\boldsymbol{\mu}}^*)$  with  $\hat{\boldsymbol{\mu}}^* = \frac{1}{n} \sum_{i=1}^n \mathbf{h}(\mathbf{y}_i^*)$ , as  $n \rightarrow \infty$

$$\sqrt{n}(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}) \xrightarrow{d^*} N(\mathbf{0}, \mathbf{V}_{\boldsymbol{\theta}})$$

where  $\mathbf{V}_{\boldsymbol{\theta}} = \mathbf{G}' \mathbf{V} \mathbf{G}$ ,  $\mathbf{V} = \mathbb{E}((\mathbf{h}(\mathbf{y}) - \boldsymbol{\mu})(\mathbf{h}(\mathbf{y}) - \boldsymbol{\mu})')$  and  $\mathbf{G} = \mathbf{G}(\boldsymbol{\mu})$ .

For a proof, see Exercise 10.9.

Theorem 10.11 shows that the asymptotic distribution of the bootstrap estimator  $\hat{\boldsymbol{\theta}}^*$  is identical to that of the sample estimator  $\hat{\boldsymbol{\theta}}$ . This means that we can learn the distribution of  $\hat{\boldsymbol{\theta}}$  from the bootstrap distribution, and hence perform asymptotically correct inference.

For some bootstrap applications we use bootstrap estimates of variance. From Section 6.16 we know that the plug-in estimator of  $V_{\theta}$  is  $\widehat{V}_{\theta} = \widehat{\mathbf{G}}' \widehat{\mathbf{V}} \widehat{\mathbf{G}}$  where  $\widehat{\mathbf{G}} = \mathbf{G}(\widehat{\boldsymbol{\mu}})$  and

$$\widehat{V} = \frac{1}{n} \sum_{i=1}^n (\mathbf{h}(y_i) - \widehat{\boldsymbol{\mu}})(\mathbf{h}(y_i) - \widehat{\boldsymbol{\mu}})'$$

The bootstrap version is

$$\begin{aligned}\widehat{V}_{\theta}^* &= \widehat{\mathbf{G}}^{*'} \widehat{\mathbf{V}}^* \widehat{\mathbf{G}}^* \\ \widehat{\mathbf{G}}^* &= \mathbf{G}(\widehat{\boldsymbol{\mu}}^*) \\ \widehat{\mathbf{V}}^* &= \frac{1}{n} \sum_{i=1}^n (\mathbf{h}(y_i^*) - \widehat{\boldsymbol{\mu}}^*)(\mathbf{h}(y_i^*) - \widehat{\boldsymbol{\mu}}^*)'\end{aligned}$$

Application of the bootstrap WLLN and bootstrap CMT show that  $\widehat{V}_{\theta}^*$  is consistent for  $V_{\theta}$ .

**Theorem 10.12** Under the assumptions of Theorem 10.11,  $\widehat{V}_{\theta}^* \xrightarrow{P^*} V_{\theta}$  as  $n \rightarrow \infty$ .

For a proof, see Exercise 10.10.

## 10.15 Consistency of the Bootstrap Estimate of Variance

Recall the definition (10.7) of the bootstrap estimator of variance  $\widehat{V}_{\widehat{\theta}}^{\text{boot}}$  of an estimator  $\widehat{\theta}$ . In this section we explore conditions under which  $\widehat{V}_{\widehat{\theta}}^{\text{boot}}$  is consistent for the asymptotic variance of  $\widehat{\theta}$ .

To do so, it is useful to focus on a normalized version of the estimator so that the asymptotic variance is not degenerate. Suppose that for some sequence  $a_n$  we have

$$z_n = a_n (\widehat{\theta} - \theta) \xrightarrow{d} \xi \quad (10.15)$$

and

$$z_n^* = a_n (\widehat{\theta}^* - \widehat{\theta}) \xrightarrow{d^*} \xi \quad (10.16)$$

for some limit distribution  $\xi$ . That is, for some normalization, both  $\widehat{\theta}$  and  $\widehat{\theta}^*$  have the same asymptotic distribution. This is quite general as it includes the smooth function model. The conventional bootstrap estimator of the variance of  $z_n$  is the sample variance of the bootstrap draws  $\{z_n^*(b) : b = 1, \dots, B\}$ . This equals the estimator (10.7) multiplied by  $a_n^2$ . Thus it is equivalent (up to scale) whether we discuss estimating the variance of  $\widehat{\theta}$  or  $z_n$ .

The bootstrap estimator of variance of  $z_n$  is

$$\begin{aligned}\widehat{V}_{\theta}^{\text{boot},B} &= \frac{1}{B-1} \sum_{b=1}^B (z_n^*(b) - \bar{z}_n^*)(z_n^*(b) - \bar{z}_n^*)' \\ \bar{z}_n^* &= \frac{1}{B} \sum_{b=1}^B z_n^*(b).\end{aligned}$$

Notice that we index the estimator by the number of bootstrap replications  $B$ .

Since  $z_n^*$  converges in bootstrap distribution to the same asymptotic distribution as  $z_n$ , it seems reasonable to guess that the variance of  $z_n^*$  will converge to that of  $\xi$ . However, we learned in Section 6.21 that convergence in distribution is not sufficient for convergence in moments. For the variance to converge it is also necessary for the sequence  $z_n^*$  to be uniformly square integrable.

**Theorem 10.13** If (10.15) and (10.16) hold for some sequence  $a_n$ , and  $\|z_n^*\|^2$  is uniformly integrable, then as  $B \rightarrow \infty$

$$\hat{V}_{\theta}^{\text{boot},B} \xrightarrow{P^*} \hat{V}_{\theta}^{\text{boot}} = \text{var}(z_n^*),$$

and as  $n \rightarrow \infty$

$$\hat{V}_{\theta}^{\text{boot}} \xrightarrow{P^*} V_{\theta} = \text{var}(\xi).$$

This raises the question: Is the normalized sequence  $z_n$  uniformly integrable? We spend the remainder of this section exploring this question, and then turn in the next section to trimmed variance estimators which do not require uniform integrability.

This condition is reasonably straightforward to verify for the case of a scalar sample mean with a finite variance. That is, suppose  $z_n^* = \sqrt{n}(\bar{y}^* - \bar{y})$  and assume  $\mathbb{E}(y^2) < \infty$ . In (10.14) we calculated the exact fourth central moment of  $z_n^*$ :

$$\mathbb{E}^*(z_n^*)^4 = \frac{\hat{\kappa}_4}{n} + 3\hat{\sigma}^4 = \frac{\hat{\mu}_4 - 3\hat{\sigma}^4}{n} + 3\hat{\sigma}^4$$

where  $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (y_i - \bar{y})^2$  and  $\hat{\mu}_4 = n^{-1} \sum_{i=1}^n (y_i - \bar{y})^4$ . The assumption  $\mathbb{E}(y^2) < \infty$  implies that  $\mathbb{E}(\hat{\sigma}^2) = O(1)$  so  $\hat{\sigma}^2 = O_p(1)$ . Furthermore,  $n^{-1}\hat{\mu}_4 = n^{-2} \sum_{i=1}^n (y_i - \bar{y})^4 = o_p(1)$  by the Marcinkiewicz WLLN (Theorem 6.39). It follows that

$$\mathbb{E}^*(z_n^*)^4 = n^2 \mathbb{E}^*(\bar{y}^* - \bar{y})^4 = O_p(1). \quad (10.17)$$

Theorem 6.31 shows that this implies that  $z_n^{*2}$  is uniformly integrable. Thus if  $y_i$  has a finite variance, the normalized bootstrap sample mean is uniformly square integrable, and the bootstrap estimate of variance is consistent by Theorem 10.13.

Now consider the smooth function model of Theorem 10.11. We can establish the following result.

**Theorem 10.14** In the smooth function model of Theorem 10.11, if for some  $p \geq 1$  the  $p^{\text{th}}$ -order derivatives of  $\mathbf{g}(\mathbf{u})$  are bounded, then  $z_n^* = \sqrt{n}(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})$  is uniformly square integrable and the bootstrap estimator of variance is consistent as in Theorem 10.13.

For a proof see Section 10.33.

This shows that the bootstrap estimate of variance is consistent for a reasonably broad class of estimators. The class of functions  $\mathbf{g}(\mathbf{u})$  covered by this result includes all  $p^{\text{th}}$ -order polynomials.

## 10.16 Trimmed Estimator of Bootstrap Variance

Theorem 10.14 showed that the bootstrap estimate of variance is consistent for smooth functions with a bounded  $p^{\text{th}}$  order derivative. This is a fairly broad class, but excludes many important applications. As a leading example, consider  $\theta = \mu_1/\mu_2$  where  $\mu_1 = \mathbb{E}(y_1)$  and  $\mu_2 = \mathbb{E}(y_2)$ . This function does not have a bounded derivative (unless  $\mu_2$  is bounded away from zero) so is not covered by Theorem 10.14.

This is more than a technical issue. When  $(y_{1i}, y_{2i})$  are jointly normally distributed, then it is known that the estimator  $\hat{\theta} = \bar{y}_1/\bar{y}_2$  does not possess a finite variance. Consequently we cannot expect the bootstrap estimator of variance to perform well. (It is attempting to estimate the variance of  $\hat{\theta}$ , which is infinity.)

In these cases it is preferred to use a trimmed estimator of bootstrap variance. Let  $\tau_n \rightarrow \infty$  be a sequence of positive trimming numbers satisfying  $\tau_n = O(e^{n/8})$ . Define the trimmed statistic

$$z_n^{**} = z_n^* \mathbf{1}(\|z_n^*\| \leq \tau_n).$$

The trimmed bootstrap estimator of variance is

$$\begin{aligned}\hat{V}_{\theta}^{\text{boot},B,\tau} &= \frac{1}{B-1} \sum_{b=1}^B (z_n^{**}(b) - \bar{z}_n^{**})(z_n^{**}(b) - \bar{z}_n^{**})' \\ \bar{z}_n^{**} &= \frac{1}{B} \sum_{b=1}^B z_n^{**}(b).\end{aligned}$$

We first examine the behavior of  $\hat{V}_{\theta}^{\text{boot},B}$  as the number of bootstrap replications  $B$  grows to infinity. It is a sample variance of independent bounded random vectors. Thus by the bootstrap WLLN (Theorem 10.6)  $\hat{V}_{\theta}^{\text{boot},B,\tau}$  converges in bootstrap probability to the variance of  $z_n^{**}$ .

**Theorem 10.15** As  $B \rightarrow \infty$ ,  $\hat{V}_{\theta}^{\text{boot},B,\tau} \xrightarrow{p^*} \hat{V}_{\theta}^{\text{boot},\tau} = \text{var}(z_n^{**})$ .

We next examine the behavior of the bootstrap estimator  $\hat{V}_{\theta}^{\text{boot},\tau}$  as  $n$  grows to infinity. We focus on the smooth function model of Theorem 10.11, which showed that  $z_n^* = \sqrt{n}(\hat{\theta}^* - \hat{\theta}) \xrightarrow{d^*} Z \sim N(\mathbf{0}, V_{\theta})$ . Since the trimming is asymptotically negligible, it follows that  $z_n^{**} \xrightarrow{d^*} Z$ . If we can show that  $z_n^{**}$  is uniformly square integrable, Theorem 10.13 will show that  $\text{var}(z_n^{**}) \rightarrow \text{var}(Z) = V_{\theta}$  as  $n \rightarrow \infty$ . This is shown in the following result, whose proof is presented in Section 10.33.

**Theorem 10.16** Under the assumptions of Theorem 10.11,  $\hat{V}_{\theta}^{\text{boot},\tau} \xrightarrow{p^*} V_{\theta}$ .

Theorems 10.15 and 10.16 show that the trimmed bootstrap estimator of variance is consistent for the asymptotic variance in the smooth function model, which includes most econometric estimators. This justifies bootstrap standard errors as consistent estimators for the asymptotic distribution.

An important caveat is that these results critically rely on the use of the trimmed variance estimator rather than the standard untrimmed version. This is a critical caveat as conventional statistical packages (e.g. Stata) calculate bootstrap standard errors using the untrimmed estimator (10.7). Thus there is no guarantee that the reported standard errors are consistent. The untrimmed variance estimator works in the context of Theorem 10.14 and whenever the bootstrap statistic is uniformly square integrable, but not necessarily in general applications.

In practice, it may be difficult to know how to select the trimming sequence  $\tau_n$ . The rule  $\tau_n = O(e^{n/8})$  does not provide practical guidance. Instead, it may be useful to think about trimming in terms of percentages of the bootstrap draws. Thus we can set  $\tau_n$  so that a given small percentage  $\gamma_n$  is trimmed. For theoretical interpretation we would set  $\gamma_n \rightarrow 0$  as  $n \rightarrow \infty$ . In practice we might set  $\gamma_n = 1\%$ .

## 10.17 Unreliability of Untrimmed Bootstrap Standard Errors

In the previous section we presented a trimmed bootstrap variance estimator which should be used to form bootstrap standard errors for nonlinear estimators. Otherwise, the standard untrimmed estimator is potentially unreliable.

This is an unfortunate situation, because reporting of bootstrap standard errors is very commonplace in contemporary applied econometric practice, and standard applications (including Stata) use the untrimmed estimator.

To illustrate the seriousness of the problem, we use the simple wage regression (7.31) which we repeat here. This is the subsample of married black women with 982 observations. The point estimates and standard errors are

$$\widehat{\log(Wage)} = 0.118 \text{ education} + 0.016 \text{ experience} - 0.022 \text{ experience}^2/100 + 0.947 . \\ (0.008) \qquad \qquad \qquad (0.006) \qquad \qquad \qquad (0.012) \qquad \qquad \qquad (0.157)$$

We are interested in the experience level which maximizes expected log wages  $\theta_3 = -50\beta_2/\beta_3$ . The point estimate and standard errors calculated with different methods are reported in Table 10.3.3 below.

The point estimate of the experience level with maximum earnings is  $\hat{\theta}_3 = 35$ . The asymptotic and jackknife standard errors are about 7. The bootstrap standard error, however, is 825! Confused by this unusual value we rerun the bootstrap again and obtain a standard error of 544. Both were computed with 10,000 bootstrap replications. The fact that the two bootstrap standard errors are considerably different when recomputed (with different starting seeds) is indicative of moment failure. When there is an enormous discrepancy like this between the asymptotic and bootstrap standard error, and between bootstrap runs, it is a signal that there may be moment failure and consequently bootstrap standard errors are unreliable.

A trimmed bootstrap with  $\tau = 25$  (set to slightly exceed three asymptotic standard errors) produces a more reasonable standard error of 10.

One message from this application is that when different methods produce very different standard errors we should be cautious about trusting any single method. The large discrepancies indicate poor asymptotic approximations, rendering all methods inaccurate. Another message is to be cautious about reporting conventional bootstrap standard errors. Trimmed versions are preferred, especially for non-linear functions of estimated coefficients.

Table 10.3: Experience Level Which Maximizes Expected log Wages

Estimate	35.2
Asymptotic s.e.	(7.0)
Jackknife s.e.	(7.0)
Bootstrap s.e. (standard)	(825)
Bootstrap s.e. (repeat)	(544)
Bootstrap s.e. (trimmed)	(10.1)

## 10.18 Consistency of the Percentile Interval

Recall the percentile interval (10.8). We now provide conditions under which it has asymptotically correct coverage.

**Theorem 10.17** Assume that for some sequence  $a_n$

$$a_n(\hat{\theta} - \theta) \xrightarrow{d} \xi \quad (10.18)$$

and

$$a_n(\hat{\theta}^* - \hat{\theta}) \xrightarrow{d^*} \xi \quad (10.19)$$

where  $\xi$  is continuously distributed and symmetric about zero. Then

$$\mathbb{P}(\theta \in C^{pc}) \longrightarrow 1 - \alpha$$

as  $n \rightarrow \infty$ .

The assumptions (10.18)-(10.19) hold for the smooth function model of Theorem 10.11, so this result incorporates many applications. The beauty of Theorem 10.17 is that the very simple confidence interval  $C^{pc}$  – which does not require technical calculation of asymptotic standard errors – has asymptotically valid coverage for any estimator which falls in the smooth function class, as well as any other estimator satisfying the convergence results (10.18)-(10.19) with  $\xi$  symmetrically distributed. The conditions are weaker than those required for consistent bootstrap variance estimation (and normal-approximation confidence intervals) because it is not necessary to verify that  $\hat{\theta}^*$  is uniformly integrable, nor necessary to employ trimming.

The proof of Theorem 10.11 is not difficult. The convergence assumption (10.19) implies that the  $\alpha^{th}$  quantile of  $a_n(\hat{\theta}^* - \hat{\theta})$ , which is  $a_n(q_\alpha^* - \hat{\theta})$  by quantile equivariance, converges in probability to the  $\alpha^{th}$  quantile of  $\xi$ , which we can denote as  $\bar{q}_\alpha$ . Thus

$$a_n(q_\alpha^* - \hat{\theta}) \xrightarrow{p} \bar{q}_\alpha. \quad (10.20)$$

Let  $H(x) = \mathbb{P}(\xi \leq x)$  be the distribution function of  $\xi$ . The assumption of symmetry implies  $H(-x) = 1 - H(x)$ . Then the percentile interval has coverage

$$\begin{aligned} \mathbb{P}(\theta \in C^{pc}) &= \mathbb{P}(q_{\alpha/2}^* \leq \theta \leq q_{1-\alpha/2}^*) \\ &= \mathbb{P}(-a_n(q_{\alpha/2}^* - \hat{\theta}) \geq a_n(\hat{\theta} - \theta) \geq -a_n(q_{1-\alpha/2}^* - \hat{\theta})) \\ &\longrightarrow \mathbb{P}(-\bar{q}_{\alpha/2} \geq \xi \geq -\bar{q}_{1-\alpha/2}) \\ &= H(-\bar{q}_{\alpha/2}) - H(-\bar{q}_{1-\alpha/2}) \\ &= H(\bar{q}_{1-\alpha/2}) - H(\bar{q}_{\alpha/2}) \\ &= 1 - \alpha. \end{aligned}$$

The convergence holds by (10.18) and (10.20). The following equality uses the definition of  $H$ , the next-to-last is the symmetry of  $H$ , and the final equality is the definition of  $\bar{q}_\alpha$ . This establishes Theorem 10.17.

Theorem 10.17 seems quite general, but it critically rests on the assumption that the asymptotic distribution  $\xi$  is symmetrically distributed about zero. This may seem innocuous, since conventional asymptotic distributions are normal and hence symmetric, but it bears further scrutiny. It is not merely a technical assumption – an examination of the steps in the preceding argument isolate quite clearly that if the symmetry assumption is violated, then the asymptotic coverage will not be  $1 - \alpha$ . While Theorem 10.17 does show that the percentile interval is asymptotically valid for a conventional asymptotically normal estimator, the reliance on symmetry in the argument suggests that the percentile method will work poorly when the finite sample distribution is asymmetric. This turns out to be the case, and will lead us to consider alternative methods in the following sections.

It is also worthwhile to investigate a finite sample justification for the percentile interval, based on a heuristic analogy due to Efron.

Assume that there exists an unknown but strictly increasing transformation  $\psi(\theta)$  such that  $\psi(\hat{\theta}) - \psi(\theta)$  has a pivotal distribution  $H(u)$  (does not vary with  $\theta$ ) which is symmetric about zero. For example, if  $\hat{\theta} \sim N(\theta, \sigma^2)$  we can set  $\psi(\theta) = \theta/\sigma$ . Alternatively, if  $\hat{\theta} = \exp(\hat{\mu})$  and  $\hat{\mu} \sim N(\mu, \sigma^2)$  then we can set  $\psi(\theta) = \log(\theta)/\sigma$ .

To assess the coverage of the percentile interval, observe that since the distribution  $H$  is pivotal the bootstrap distribution  $\psi(\hat{\theta}^*) - \psi(\hat{\theta})$  also has distribution  $H(u)$ . Let  $\bar{q}_\alpha$  be the  $\alpha^{th}$  quantile of the distribution  $H$ . Since  $q_\alpha^*$  is the  $\alpha^{th}$  quantile of the distribution of  $\hat{\theta}^*$ , and  $\psi(\hat{\theta}^*) - \psi(\hat{\theta})$  is a monotonic transformation of  $\hat{\theta}^*$ , by the quantile equivariance property we deduce that  $\bar{q}_\alpha + \psi(\hat{\theta}) = \psi(q_\alpha^*)$ . The percentile interval has coverage

$$\begin{aligned}\mathbb{P}(\theta \in C^{\text{pc}}) &= \mathbb{P}(q_{\alpha/2}^* \leq \theta \leq q_{1-\alpha/2}^*) \\ &= \mathbb{P}(\psi(q_{\alpha/2}^*) \leq \psi(\theta) \leq \psi(q_{1-\alpha/2}^*)) \\ &= \mathbb{P}(\psi(\hat{\theta}) - \psi(q_{\alpha/2}^*) \geq \psi(\hat{\theta}) - \psi(\theta) \geq \psi(\hat{\theta}) - \psi(q_{1-\alpha/2}^*)) \\ &= \mathbb{P}(-\bar{q}_{\alpha/2} \geq \psi(\hat{\theta}) - \psi(\theta) \geq -\bar{q}_{1-\alpha/2}) \\ &= H(-\bar{q}_{\alpha/2}) - H(-\bar{q}_{1-\alpha/2}) \\ &= H(\bar{q}_{1-\alpha/2}) - H(\bar{q}_{\alpha/2}) \\ &= 1 - \alpha.\end{aligned}$$

The second equality applies the monotonic transformation  $\psi(u)$  to all elements. The fourth uses the relationship  $\bar{q}_\alpha + \psi(\hat{\theta}) = \psi(q_\alpha^*)$ . The fifth uses the defintion of  $H$ . The sixth uses the symmetry property of  $H$ , and the final is by the definition of  $\bar{q}_\alpha$  as the  $\alpha^{th}$  quantile of  $H$ .

This calculation shows that under these assumptions the percentile interval has exact coverage  $1 - \alpha$ . The nice thing about this argument is the introduction of the unknown transformation  $\psi(u)$  for which the percentile interval automatically adapts. The unpleasant feature is the assumption of symmetry. Similar to the asymptotic argument, the calculation strongly relies on the symmetry of the distribution  $H(x)$ . Without symmetry the coverage will be incorrect.

Intuitively, we expect that when the assumptions are approximately true, then the percentile interval will have approximately correct coverage. Thus so long as there is a transformation  $\psi(u)$  such that  $\psi(\hat{\theta}) - \psi(\theta)$  is approximately pivotal and symmetric about zero, then the percentile interval should work well.

This argument has the following application. Suppose that the parameter of interest is  $\theta = \exp(\mu)$  where  $\mu = \mathbb{E}(y)$  and suppose  $y$  has a pivotal symmetric distribution about  $\mu$ . Then even though  $\hat{\theta} = \exp(\bar{y})$  does not have a symmetric distribution, the percentile interval applied to  $\hat{\theta}$  will have the correct coverage, because the monotonic transformation  $\log(\hat{\theta})$  has a pivotal symmetric distribution.

## 10.19 Bias-Corrected Percentile Interval

The accuracy of the percentile interval depends critically upon the assumption that the sampling distribution is approximately symmetrically distributed. This excludes finite sample bias, for an estimator which is biased cannot be symmetrically distributed. Many contexts in which we want to apply bootstrap methods (rather than asymptotic) are when the parameter of interest is a nonlinear function of the original estimates, and nonlinearity typically induces estimation bias. Consequently it is difficult to expect the percentile method to generally have accurate coverage.

To remove the bias problem, Efron (1982) introduced the **bias-corrected (BC) percentile interval**. The justification is heuristic, but there is considerable evidence that the bias-corrected method is an important improvement on the percentile interval.

The construction is based on the assumption is that there is a an unknown but strictly increasing transformation  $\psi(\theta)$  and unknown constant  $z_0$  such that

$$Z = \psi(\hat{\theta}) - \psi(\theta) + z_0 \sim N(0, 1). \quad (10.21)$$

(The assumption that  $Z$  is normal is not critical. It could be replaced by any known symmetric and invertible distribution.) Let  $\Phi(x)$  denote the normal distribution function,  $\Phi^{-1}(p)$  its quantile function,

and  $z_\alpha = \Phi^{-1}(\alpha)$  the normal critical values. Then the BC interval can be constructed from the bootstrap estimators  $\hat{\theta}_b^*$  and bootstrap quantiles  $q_\alpha^*$  as follows. Set

$$p^* = \frac{1}{B} \sum_{b=1}^B \mathbf{1}(\hat{\theta}_b^* \leq \hat{\theta}) \quad (10.22)$$

and

$$z_0^* = \Phi^{-1}(p^*). \quad (10.23)$$

$p^*$  is a measure of median bias, and  $z_0$  is  $p^*$  transformed into normal units. If the bias of  $\hat{\theta}$  is zero then  $p^* = 0.5$  and  $z_0^* = 0$ . If  $\hat{\theta}$  is upwards biased then  $p^* < 0.5$  and  $z_0^* < 0$ . Conversely if  $\hat{\theta}$  is downward biased then  $p^* > 0.5$  and  $z_0^* > 0$ . Define for any  $\alpha$  an adjusted version

$$x(\alpha) = \Phi(z_\alpha + 2z_0). \quad (10.24)$$

If  $z_0 = 0$  then  $x(\alpha) = \alpha$ . If  $z_0 > 0$  then  $x(\alpha) > \alpha$ , and conversely when  $x(\alpha) < 0$ . The BC percentile interval is

$$C^{\text{bc}} = [q_{x(\alpha/2)}^*, q_{x(1-\alpha/2)}^*]. \quad (10.25)$$

Essentially, rather than going from the 2.5% to 97.5% quantile, the BC interval uses adjusted quantiles, with the degree of adjustment depending on the extent of the bias.

The construction of the BC interval is not intuitive. We now show that assumption (10.21) implies that the BC interval has exact coverage. (10.21) implies that

$$\mathbb{P}(\psi(\hat{\theta}) - \psi(\theta) + z_0 \leq x) = \Phi(x).$$

Since the distribution is pivotal the result carries over to the bootstrap distribution

$$\mathbb{P}^*(\psi(\hat{\theta}^*) - \psi(\hat{\theta}) + z_0 \leq x) = \Phi(x). \quad (10.26)$$

Evaluating (10.26) at  $x = z_0$  we find  $\mathbb{P}^*(\psi(\hat{\theta}^*) - \psi(\hat{\theta}) \leq 0) = \Phi(z_0)$  which implies  $\mathbb{P}^*(\hat{\theta}^* \leq \hat{\theta}) = \Phi(z_0)$ . Inverting, we obtain

$$z_0 = \Phi^{-1}(\mathbb{P}^*(\hat{\theta}^* \leq \hat{\theta})) \quad (10.27)$$

which is the probability limit of (10.23) as  $B \rightarrow \infty$ . Thus the unknown  $z_0$  is recovered by (10.23), and we can treat  $z_0$  as if it were known.

From (10.26) we deduce that

$$\begin{aligned} x(\alpha) &= \Phi(z_\alpha + 2z_0) \\ &= \mathbb{P}^*(\psi(\hat{\theta}^*) - \psi(\hat{\theta}) \leq z_\alpha + z_0) \\ &= \mathbb{P}^*(\hat{\theta}^* \leq \psi^{-1}(\psi(\hat{\theta}) + z_0 + z_\alpha)). \end{aligned}$$

This equation shows that  $\psi^{-1}(\psi(\hat{\theta}) + z_0 + z_\alpha)$  equals the  $x(\alpha)^{\text{th}}$  bootstrap quantile. That is,  $q_{x(\alpha)}^* = \psi^{-1}(\psi(\hat{\theta}) + z_0 + z_\alpha)$ . Hence we can write (10.25) as

$$C^{\text{bc}} = [\psi^{-1}(\psi(\hat{\theta}) + z_0 + z_{\alpha/2}), \psi^{-1}(\psi(\hat{\theta}) + z_0 + z_{1-\alpha/2})].$$

It has coverage probability

$$\begin{aligned} \mathbb{P}(\theta \in C^{\text{bc}}) &= \mathbb{P}(\psi^{-1}(\psi(\hat{\theta}) + z_0 + z_{\alpha/2}) \leq \theta \leq \psi^{-1}(\psi(\hat{\theta}) + z_0 + z_{1-\alpha/2})) \\ &= \mathbb{P}(\psi(\hat{\theta}) + z_0 + z_{\alpha/2} \leq \psi(\theta) \leq \psi(\hat{\theta}) + z_0 + z_{1-\alpha/2}) \\ &= \mathbb{P}(-z_{\alpha/2} \geq \psi(\hat{\theta}) - \psi(\theta) + z_0 \geq -z_{1-\alpha/2}) \\ &= \mathbb{P}(z_{1-\alpha/2} \geq Z \geq z_{\alpha/2}) \\ &= \Phi(z_{1-\alpha/2}) - \Phi(z_{\alpha/2}) \\ &= 1 - \alpha. \end{aligned}$$

The second equality applies the transformation  $\psi(\theta)$ . The fourth equality uses the model (10.21) and the fact  $z_\alpha = -z_{1-\alpha}$ . This shows that the BC interval (10.25) has exact coverage under the assumption (10.21).

Furthermore, under the assumptions of Theorem 10.17, the BC interval has asymptotic coverage probability  $1 - \alpha$ , since the bias correction is asymptotically negligible.

An important property of the BC percentile interval is that it is transformation-respecting (like the percentile interval). To see this, observe that  $p^*$  is invariant to transformations since it is a probability, and thus  $z_0^*$  and  $x(\alpha)$  are invariant. Since the interval is constructed from the  $x(\alpha/2)$  and  $x(1 - \alpha/2)$  quantiles, the quantile equivariance property shows that the interval is transformation-respecting.

The bootstrap BC percentile intervals for the four estimators are reported in Table 13.2. They are generally similar to the percentile intervals, though the intervals for  $\sigma^2$  and  $\mu$  are somewhat shifted to the right.

In Stata, BC percentile confidence intervals can be obtained by using the command `estat bootstrap` after an estimation command which calculates standard errors via the bootstrap.

## 10.20 BC <sub>$\alpha$</sub> Percentile Interval

A further improvement on the BC interval was made by Efron (1987) to account for the skewness in the sampling distribution, which can be modeled by specifying that the variance of the estimator depends on the parameter. The resulting **bootstrap accelerated bias-corrected percentile interval** (BC <sub>$\alpha$</sub> ) has improved performance on the BC interval, but requires a bit more computation and is less intuitive to understand.

The construction is a generalization of that for the BC intervals. The assumption is that there is an unknown but strictly increasing transformation  $\psi(\theta)$ , and unknown constants  $a$  and  $z_0$  such that

$$Z = \frac{\psi(\hat{\theta}) - \psi(\theta)}{1 + a\psi(\theta)} + z_0 \sim N(0, 1). \quad (10.28)$$

(As before, the assumption that  $Z$  is normal could be replaced by any known symmetric and invertible distribution.)

The constant  $z_0$  is estimated by (10.23) just as for the BC interval. There are several possible estimators of  $a$ . Efron's suggestion is a scaled jackknife estimator of the skewness of  $\hat{\theta}$ :

$$\begin{aligned} \hat{a} &= \frac{\sum_{i=1}^n (\bar{\theta} - \hat{\theta}_{(-i)})^3}{6 \left( \sum_{i=1}^n (\bar{\theta} - \hat{\theta}_{(-i)})^2 \right)^{3/2}} \\ \bar{\theta} &= \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{(-i)}. \end{aligned}$$

The jackknife estimator of  $\hat{a}$  makes the BC <sub>$\alpha$</sub>  interval more computationally costly than the other intervals.

Define for any  $\alpha$  the adjusted version

$$x(\alpha) = \Phi \left( z_0 + \frac{z_\alpha + z_0}{1 - a(z_\alpha + z_0)} \right).$$

The BC <sub>$\alpha$</sub>  percentile interval is

$$C^{\text{bca}} = [q_{x(\alpha/2)}^*, q_{x(1-\alpha/2)}^*].$$

Note that  $x(\alpha)$  simplifies to (10.24) and  $C^{\text{bca}}$  simplifies to  $C^{\text{bc}}$  when  $a = 0$ . While  $C^{\text{bc}}$  improves on  $C^{\text{pc}}$  by correcting the median bias,  $C^{\text{bca}}$  makes a further correction for skewness.

The BC <sub>$\alpha$</sub>  interval is only well-defined for values of  $\alpha$  such that  $a(z_\alpha + z_0) < 1$ . (Or equivalently, if  $\alpha < \Phi(a^{-1} - z_0)$  for  $a > 0$  and  $\alpha > \Phi(a^{-1} - z_0)$  for  $a < 0$ .)

The  $\text{BC}_a$  interval, like the BC and percentile intervals, is transformation-respecting. Thus if  $[q_{x(\alpha/2)}^*, q_{x(1-\alpha/2)}^*]$  is the  $\text{BC}_a$  interval for  $\theta$ , then  $[m(q_{x(\alpha/2)}^*), m(q_{x(1-\alpha/2)}^*)]$  is the  $\text{BC}_a$  interval for  $\phi = m(\theta)$  when  $m(\theta)$  is monotone.

We now give a justification for the  $\text{BC}_a$  interval. The most difficult feature to understand is the estimator  $\hat{a}$  for  $a$ . This involves higher-order approximations which are too advanced for our treatment, so we instead refer readers to Chapter 4.1.4 of Shao and Tu (1995), and simply assume that  $a$  is known.

We now show that assumption (10.28) with  $a$  known implies that  $C^{\text{bca}}$  has exact coverage. The argument is essentially the same as that given in the previous section. Assumption (10.28) implies that the bootstrap distribution satisfies

$$\mathbb{P}^* \left( \frac{\psi(\hat{\theta}^*) - \psi(\hat{\theta})}{1 + a\psi(\hat{\theta})} + z_0 \leq x \right) = \Phi(x). \quad (10.29)$$

Evaluating at  $x = z_0$  and inverting we obtain (10.27) which is the same as for the BC interval. Thus the estimator (10.23) is consistent as  $B \rightarrow \infty$ , and we can treat  $z_0$  as if it were known.

From (10.29) we deduce that

$$\begin{aligned} x(a) &= \mathbb{P}^* \left( \frac{\psi(\hat{\theta}^*) - \psi(\hat{\theta})}{1 + a\psi(\hat{\theta})} \leq \frac{z_\alpha + z_0}{1 - a(z_\alpha + z_0)} \right) \\ &= \mathbb{P}^* \left( \hat{\theta}^* \leq \psi^{-1} \left( \frac{\psi(\hat{\theta}) + z_\alpha + z_0}{1 - a(z_\alpha + z_0)} \right) \right). \end{aligned}$$

This shows that  $\psi^{-1} \left( \frac{\psi(\hat{\theta}) + z_\alpha + z_0}{1 - a(z_\alpha + z_0)} \right)$  equals the  $x(a)^{\text{th}}$  bootstrap quantile. Hence we can write  $C^{\text{bca}}$  as

$$C^{\text{bca}} = \left[ \psi^{-1} \left( \frac{\psi(\hat{\theta}) + z_{\alpha/2} + z_0}{1 - a(z_{\alpha/2} + z_0)} \right), \quad \psi^{-1} \left( \frac{\psi(\hat{\theta}) + z_{1-\alpha/2} + z_0}{1 - a(z_{1-\alpha/2} + z_0)} \right) \right].$$

It has coverage probability

$$\begin{aligned} \mathbb{P}(\theta \in C^{\text{bca}}) &= \mathbb{P} \left( \psi^{-1} \left( \frac{\psi(\hat{\theta}) + z_{\alpha/2} + z_0}{1 - a(z_{\alpha/2} + z_0)} \right) \leq \theta \leq \psi^{-1} \left( \frac{\psi(\hat{\theta}) + z_{1-\alpha/2} + z_0}{1 - a(z_{1-\alpha/2} + z_0)} \right) \right) \\ &= \mathbb{P} \left( \frac{\psi(\hat{\theta}) + z_{\alpha/2} + z_0}{1 - a(z_{\alpha/2} + z_0)} \leq \psi(\theta) \leq \frac{\psi(\hat{\theta}) + z_{1-\alpha/2} + z_0}{1 - a(z_{1-\alpha/2} + z_0)} \right) \\ &= \mathbb{P} \left( -z_{\alpha/2} \geq \frac{\psi(\hat{\theta}) - \psi(\theta)}{1 + a\psi(\theta)} + z_0 \geq -z_{1-\alpha/2} \right) \\ &= \mathbb{P}(z_{1-\alpha/2} \geq Z \geq z_{\alpha/2}) \\ &= 1 - \alpha. \end{aligned}$$

The second equality applies the transformation  $\psi(\theta)$ . The fourth equality uses the model (10.28) and the fact  $z_\alpha = -z_{1-\alpha}$ . This shows that the  $\text{BC}_a$  interval  $C^{\text{bca}}$  has exact coverage under the assumption (10.28) with  $a$  known.

The bootstrap  $\text{BC}_a$  percentile intervals for the four estimators are reported in Table 13.2. They are generally similar to the BC intervals, though the intervals for  $\sigma^2$  and  $\mu$  are slightly shifted to the right.

In Stata,  $\text{BC}_a$  intervals can be obtained by using the command `estat bootstrap, bca` or the command `estat bootstrap, all` after an estimation command which calculates standard errors via the bootstrap using the `bca` option.

## 10.21 Percentile-t Interval

In many cases we can obtain improvement in accuracy by bootstrapping a studentized statistic such as a t-ratio. Let  $\hat{\theta}$  be an estimator of a scalar parameter  $\theta$  and  $s(\hat{\theta})$  a standard error. The sample t-ratio is

$$T = \frac{\hat{\theta} - \theta}{s(\hat{\theta})}.$$

The bootstrap t-ratio is

$$T^* = \frac{\hat{\theta}^* - \hat{\theta}}{s(\hat{\theta}^*)}$$

where  $s(\hat{\theta}^*)$  is the standard error calculated on the bootstrap sample. Notice that the bootstrap t-ratio is centered at the parameter estimate  $\hat{\theta}$ . This is because  $\hat{\theta}$  is the “true value” in the bootstrap universe.

The percentile-t interval is formed using the distribution of  $T^*$ . This can be calculated via the bootstrap algorithm. On each bootstrap sample the estimator  $\hat{\theta}^*$  and its standard error  $s(\hat{\theta}^*)$  are calculated, and the t-ratio  $T^* = (\hat{\theta}^* - \hat{\theta}) / s(\hat{\theta}^*)$  calculated and stored. This is repeated  $B$  times. The  $\alpha^{th}$  quantile  $q_\alpha^*$  is estimated by the  $\alpha^{th}$  empirical quantile (or any quantile estimator) from the  $B$  bootstrap draws of  $T^*$ .

The bootstrap percentile-t confidence interval is then defined as

$$C^{\text{pt}} = [\hat{\theta} - s(\hat{\theta}) q_{1-\alpha/2}^*, \quad \hat{\theta} - s(\hat{\theta}) q_{\alpha/2}^*].$$

The form may appear unusual when compared with the percentile-type intervals. The left endpoint is determined by the upper quantile of the distribution of  $T^*$ , and the right endpoint is determined by the lower quantile. As we show below, this construction is important for the interval to have correct coverage when the distribution is not symmetric.

When the estimator is asymptotically normal and the standard error a reliable estimator of the standard deviation of the distribution, we would expect the t-ratio  $T$  to be roughly approximated by the normal distribution. In this case we would expect  $q_{0.975}^* \approx -q_{0.025}^* \approx 2$ . Departures from this baseline occur as the distribution becomes skewed or fat-tailed. If the bootstrap quantiles depart substantially from this baseline it is evidence of substantial departure from normality. (It may also indicate a programming error, so in these cases it is wise to triple-check!)

The percentile-t has the following advantages. First, when the standard error  $s(\hat{\theta})$  is reasonably reliable, the percentile-t bootstrap makes use of the information in the standard error, thereby reducing the role of the bootstrap. This can improve the precision of the method relative to other methods. Second, as we show later, the percentile-t intervals achieve higher-order accuracy than the percentile and BC percentile intervals. Third, the percentile-t intervals correspond to the set of parameter values “not rejected” by one-sided t-tests using bootstrap critical values (bootstrap tests are presented in Section 10.23).

The percentile-t interval has the following disadvantages. First, they may be infeasible when standard error formula are unknown. Second, they may be practically infeasible when standard error calculations are computationally costly (since the standard error calculation needs to be performed on each bootstrap sample). Third, the percentile-t may be unreliable if the standard errors  $s(\hat{\theta})$  are unreliable and thus add more noise than clarity. Fourth, the percentile-t interval is not translation preserving, unlike the percentile, BC percentile, and  $BC_a$  percentile intervals.

It is typical to calculate percentile-t interval with t-ratios constructed with conventional asymptotic standard errors. But this is not the only possible implementation. The percentile-t interval can be constructed with any data-dependent measure of scale. For example, if  $\hat{\theta}$  is a two-step estimator for which it is unclear how to construct a correct asymptotic standard error, but we know how to calculate a standard error  $s(\hat{\theta})$  appropriate for the second step alone, then  $s(\hat{\theta})$  can be used for a percentile-t-type interval as described above. It will not possess the higher-order accuracy properties of the following section, but it will satisfy the conditions for first-order validity.

Furthermore, percentile-t intervals can be constructed using bootstrap standard errors. That is, the statistics  $T$  and  $T^*$  can be computed using bootstrap standard errors  $s_{\hat{\theta}}^{\text{boot}}$ . This is computationally

costly, as it requires is called a nested bootstrap. Specifically, for each bootstrap replication, a random sample is drawn, the bootstrap estimate  $\hat{\theta}^*$  computed, and then  $B$  additional bootstrap sub-samples drawn from the bootstrap sample to compute the bootstrap standard error for the bootstrap estimate  $\hat{\theta}^*$ . Effectively  $B^2$  bootstrap samples are drawn and estimated, which increases the computational requirement by an order of magnitude.

We now describe the distribution theory for first-order validity of the percentile-t bootstrap.

First, consider the smooth function model, where  $\hat{\theta} = g(\hat{\mu})$  and  $s(\hat{\theta}) = \sqrt{\frac{1}{n}\hat{\mathbf{G}}'\hat{\mathbf{V}}\hat{\mathbf{G}}}$  with bootstrap analogs  $\hat{\theta}^* = g(\hat{\mu}^*)$  and  $s(\hat{\theta}^*) = \sqrt{\frac{1}{n}\hat{\mathbf{G}}^{*'}\hat{\mathbf{V}}^*\hat{\mathbf{G}}^*}$ . From Theorems 6.24, 6.25, 10.11 and 10.12

$$T = \frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\hat{\mathbf{G}}'\hat{\mathbf{V}}\hat{\mathbf{G}}}} \xrightarrow{d} Z$$

and

$$T^* = \frac{\sqrt{n}(\hat{\theta}^* - \hat{\theta})}{\sqrt{\hat{\mathbf{G}}^{*'}\hat{\mathbf{V}}^*\hat{\mathbf{G}}^*}} \xrightarrow{d^*} Z$$

where  $Z \sim N(0, 1)$ . This shows that the sample and bootstrap t-ratios have the same asymptotic distribution.

This motivates considering the broader situation where the sample and bootstrap t-ratios have the same asymptotic distribution, but not necessarily normal. Thus assume that

$$T \xrightarrow{d} \xi \tag{10.30}$$

$$T^* \xrightarrow{d^*} \xi \tag{10.31}$$

for some continuous distribution  $\xi$ . (10.31) implies that the quantiles of  $T^*$  converge in probability to those of  $\xi$ , that is  $q_\alpha^* \xrightarrow{p} q_\alpha$  where  $q_\alpha$  is the  $\alpha^{th}$  quantile of  $\xi$ . This and (10.30) imply

$$\begin{aligned} \mathbb{P}(\theta \in C^{\text{pt}}) &= \mathbb{P}(\hat{\theta} - s(\hat{\theta})q_{1-\alpha/2}^* \leq \theta \leq \hat{\theta} - s(\hat{\theta})q_{\alpha/2}^*) \\ &= \mathbb{P}(q_{\alpha/2}^* \leq T \leq q_{1-\alpha/2}^*) \\ &\longrightarrow \mathbb{P}(q_{\alpha/2} \leq \xi \leq q_{1-\alpha/2}) \\ &= 1 - \alpha. \end{aligned}$$

Thus the percentile-t is asymptotically valid.

**Theorem 10.18** If (10.30) and (10.31) hold where  $\xi$  is continuously distributed, then

$$\mathbb{P}(\theta \in C^{\text{pt}}) \longrightarrow 1 - \alpha$$

as  $n \rightarrow \infty$ .

The bootstrap percentile-t intervals for the four estimators are reported in Table 13.2. They are similar but somewhat different from the percentile-type intervals, and generally wider. The largest difference arises with the interval for  $\sigma^2$ , which is noticeably wider than the other intervals.

## 10.22 Percentile-t Asymptotic Refinement

The percentile-t interval can be viewed as the intersection of two one-sided confidence intervals. In our discussion of Edgeworth expansions for the coverage probability of one-sided asymptotic confidence intervals (following Theorem 7.15 in the context of functions of regression coefficients) we found

that one-sided asymptotic confidence intervals have accuracy to order  $O(n^{-1/2})$ . We now show that the percentile-t interval has improved accuracy.

Theorems 6.35 and 6.37 showed that the Cornish-Fisher expansion for the quantile  $q_\alpha$  of a t-ratio  $T$  in the smooth function model takes the form

$$q_\alpha = z_\alpha + n^{-1/2} p_{11}(z_\alpha) + O(n^{-1})$$

where  $p_{11}(x)$  is an even polynomial of order 2 with coefficients depending on the moments of  $\mathbf{h}(\mathbf{y})$  up to order 8. The bootstrap quantile  $q_\alpha^*$  has a similar Cornish-Fisher expansion

$$q_\alpha^* = z_\alpha + n^{-1/2} p_{11}^*(z_\alpha) + O_p(n^{-1})$$

where  $p_{11}^*(x)$  is the same as  $p_{11}(x)$  except that the moments of  $\mathbf{h}(\mathbf{y})$  are replaced by the corresponding sample moments. Sample moments are estimated at the rate  $n^{-1/2}$ . Thus we can replace  $p_{11}^*$  with  $p_{11}$  without affecting the order of this expansion:

$$\begin{aligned} q_\alpha^* &= z_\alpha + n^{-1/2} p_{11}(z_\alpha) + O_p(n^{-1}) \\ &= q_\alpha + O_p(n^{-1}). \end{aligned}$$

This shows that the bootstrap quantiles  $q_\alpha^*$  of the studentized t-ratio are within  $O_p(n^{-1})$  of the exact quantiles  $q_\alpha$ .

By the Edgeworth expansion Delta method (Theorem 6.36),  $T$  and  $T + (q_\alpha - q_\alpha^*) = T + O_p(n^{-1})$  have the same Edgeworth expansion to order  $O(n^{-1})$ . Thus

$$\begin{aligned} \mathbb{P}(T \leq q_\alpha^*) &= \mathbb{P}(T + (q_\alpha - q_\alpha^*) \leq q_\alpha) \\ &= \mathbb{P}(T \leq q_\alpha) + O(n^{-1}) \\ &= \alpha + O(n^{-1}). \end{aligned}$$

Thus the coverage of the percentile-t interval is

$$\begin{aligned} \mathbb{P}(\theta \in C^{\text{pt}}) &= \mathbb{P}(q_{\alpha/2}^* \leq T \leq q_{1-\alpha/2}^*) \\ &= \mathbb{P}(q_{\alpha/2} \leq T \leq q_{1-\alpha/2}) + O(n^{-1}) \\ &= 1 - \alpha + O(n^{-1}). \end{aligned}$$

This is an improved rate of convergence relative to the one-sided asymptotic confidence interval.

**Theorem 10.19** Under the assumptions of Theorem 6.35,

$$\mathbb{P}(\theta \in C^{\text{pt}}) = 1 - \alpha + O(n^{-1}).$$

The following definition of the accuracy of a confidence interval is useful.

**Definition 10.5** A confidence set  $C$  for  $\theta$  is  $k^{\text{th}}$ -order accurate if

$$\mathbb{P}(\theta \in C) = 1 - \alpha + O(n^{-k/2}).$$

Examining our results, we find that one-sided asymptotic confidence intervals are first-order accurate, but percentile-t intervals are second-order accurate. When a bootstrap confidence interval (or test) achieves high-order accuracy than the analogous asymptotic interval (or test), we say that the bootstrap method achieves an **asymptotic refinement**. Here, we have shown that the percentile-t interval achieves an asymptotic refinement.

In order to achieve this asymptotic refinement, it is important that the t-ratio  $T$  (and its bootstrap counter-part  $T^*$ ) are constructed with asymptotically valid standard errors. This is because the first term in the Edgeworth expansion is the standard normal distribution, and this requires that the t-ratio is asymptotically normal. This also has the practical finite-sample implication that the accuracy of the percentile-t interval in practice depends on the accuracy of the standard errors used to construct the t-ratio.

We do not go through the details, but normal-approximation bootstrap intervals, percentile bootstrap intervals, and bias-corrected percentile bootstrap intervals are all first-order accurate, and do not achieve an asymptotic refinement.

The  $BC_a$  interval, however, can be shown to be asymptotically equivalent to the percentile-t interval, and thus achieves an asymptotic refinement. We do not make this demonstration here as it is too advanced. For a demonstration see Section 3.10.4 of Hall (1992).

### Peter Hall

Peter Gavin Hall (1951-2016) of Australia was one of the most influential and prolific theoretical statisticians in history. He made wide-ranging contributions. Some of the most relevant for econometrics are theoretical investigations of bootstrap methods and nonparametric kernel methods.

## 10.23 Bootstrap Hypothesis Tests

To test the hypothesis  $H_0 : \theta = \theta_0$  against  $H_1 : \theta \neq \theta_0$  the most common approach is a t-test. We reject  $H_0$  in favor of  $H_1$  for large absolute values of the t-statistic

$$T = \frac{\hat{\theta} - \theta_0}{s(\hat{\theta})},$$

where  $\hat{\theta}$  is an estimator of  $\theta$  and  $s(\hat{\theta})$  is a standard error for  $\hat{\theta}$ . For a bootstrap test we use the bootstrap algorithm to calculate the critical value.

The bootstrap algorithm samples with replacement from the dataset. Given a bootstrap sample the bootstrap estimator  $\hat{\theta}^*$  and standard error  $s(\hat{\theta}^*)$  are calculated. Given these values the bootstrap t-statistic is

$$T^* = \frac{\hat{\theta}^* - \hat{\theta}}{s(\hat{\theta}^*)}.$$

There are two important features about the bootstrap t-statistic. First,  $T^*$  is centered at the sample estimate  $\hat{\theta}$ , not at the hypothesized value  $\theta_0$ . This is done because  $\hat{\theta}$  is the true value in the bootstrap universe, and the distribution of the t-statistic must be centered at the true value within the bootstrap sampling framework. Second,  $T^*$  is calculated using the bootstrap standard error  $s(\hat{\theta}^*)$ . This allows the bootstrap to incorporate the randomness in standard error estimation.

The failure to properly center the bootstrap statistic at  $\hat{\theta}$  is a common error in applications. Often this is because the hypothesis to be tested is  $H_0 : \theta = 0$ , so the test statistic is  $T = \hat{\theta}/s(\hat{\theta})$ . This intuitively suggests the bootstrap statistic  $T^* = \hat{\theta}^*/s(\hat{\theta}^*)$ , but this is wrong. The correct bootstrap statistic is  $T^* = (\hat{\theta}^* - \hat{\theta})/s(\hat{\theta}^*)$ .

The bootstrap algorithm creates  $B$  draws  $T^*(b) = (\hat{\theta}^*(b) - \hat{\theta}) / s(\hat{\theta}^*(b))$ ,  $b = 1, \dots, B$ . The bootstrap  $100\alpha\%$  critical value is  $q_{1-\alpha}^*$ , where  $q_\alpha^*$  is the  $\alpha^{th}$  quantile of the absolute values of the bootstrap t-ratios  $|T^*(b)|$ . For a  $100\alpha\%$  test we reject  $H_0 : \theta = \theta_0$  in favor of  $H_1 : \theta \neq \theta_0$  if  $|T| > q_{1-\alpha}^*$  and fail to reject if  $|T| \leq q_{1-\alpha}^*$ .

It is generally better to report p-values rather than critical values. Recall that a p-value is  $p = 1 - G_n(|T|)$  where  $G_n(u)$  is the null distribution of the statistic  $|T|$ . The bootstrap p-value is defined as  $p^* = 1 - G_n^*(|T|)$ , where  $G_n^*(u)$  is the bootstrap distribution of  $|T^*|$ . This is estimated from the bootstrap algorithm as

$$p^* = \frac{1}{B} \sum_{b=1}^B \mathbf{1}(|T^*(b)| > |T|),$$

the percentage of bootstrap t-statistics that are larger than the observed t-statistic. Intuitively, we want to know how “unusual” is the observed statistic  $T$  when the null hypothesis is true. The bootstrap algorithm generates a large number of independent draws from the distribution  $T^*$  (which is an approximation to the unknown distribution of  $T$ ). If the percentage of the  $|T^*|$  that exceed  $|T|$  is very small (say 1%) this tells us that  $|T|$  is an unusually large value. However, if the percentage is larger, say 15%, then we cannot interpret  $|T|$  as unusually large.

If desired, the bootstrap test can be implemented as a one-sided test. In this case the statistic is the signed version of the t-ratio, and bootstrap critical values are calculated from the upper tail of the distribution for the alternative  $H_1 : \theta > \theta_0$ , and from the lower tail for the alternative  $H_1 : \theta < \theta_0$ . There is a connection between the one-sided tests and the percentile-t confidence interval. The latter is the set of parameter values  $\theta$  which are not rejected by either one-sided  $100\alpha/2\%$  bootstrap t-test.

Bootstrap tests can also be conducted with other statistics. When standard errors are not available or are not reliable, we can use the non-studentized statistic  $T = \hat{\theta} - \theta_0$ . The bootstrap version is  $T^* = \hat{\theta}^* - \hat{\theta}$ . Let  $q_\alpha^*$  be the  $\alpha^{th}$  quantile of the bootstrap statistics  $|\hat{\theta}^*(b) - \hat{\theta}|$ . A bootstrap  $100\alpha\%$  test rejects  $H_0 : \theta = \theta_0$  if  $|\hat{\theta} - \theta_0| > q_{1-\alpha}^*$ . The bootstrap p-value is

$$p^* = \frac{1}{B} \sum_{b=1}^B \mathbf{1}(|\hat{\theta}^*(b) - \hat{\theta}| > |\hat{\theta} - \theta_0|).$$

**Theorem 10.20** If (10.30) and (10.31) hold where  $\xi$  is continuously distributed, then the bootstrap critical value satisfies

$$q_{1-\alpha}^* \xrightarrow{P} q_{1-\alpha}$$

where  $q_{1-\alpha}$  is the  $1 - \alpha^{th}$  quantile of  $|\xi|$ . The bootstrap test “Reject  $H_0$  in favor of  $H_1$  if  $|T| > q_{1-\alpha}^*$ ” has asymptotic size  $\alpha$ :

$$\mathbb{P}(|T| > q_{1-\alpha}^* | H_0) \longrightarrow \alpha$$

as  $n \rightarrow \infty$ .

In the smooth function model the t-test (with correct standard errors) has the following performance.

**Theorem 10.21** Under the assumptions of Theorem 6.35,

$$q_{1-\alpha}^* = \bar{z}_{1-\alpha} + o_p(n^{-1})$$

where  $\bar{z}_\alpha = \Phi^{-1}((1+\alpha)/2)$  is the  $\alpha^{th}$  quantile of  $|Z|$ . The asymptotic test “Reject  $H_0$  in favor of  $H_1$  if  $|T| > \bar{z}_{1-\alpha}$ ” has accuracy

$$\mathbb{P}(|T| > \bar{z}_{1-\alpha} | H_0) = 1 - \alpha + O(n^{-1})$$

and the bootstrap test “Reject  $H_0$  in favor of  $H_1$  if  $|T| > q_{1-\alpha}^*$ ” has accuracy

$$\mathbb{P}(|T| > q_{1-\alpha}^* | H_0) = 1 - \alpha + o(n^{-1}).$$

This shows that the bootstrap test achieves a refinement relative to the asymptotic test.

The reasoning is as follows. We have shown that the Edgeworth expansion for the absolute t-ratio takes the form

$$\mathbb{P}(|T| \leq x) = 2\Phi(x) - 1 + n^{-1}2p_2(x) + o(n^{-1}).$$

This means the asymptotic test has accuracy of order  $O(n^{-1})$ .

Given the Edgeworth expansion, the Cornish-Fisher expansion for the  $\alpha^{th}$  quantile  $q_\alpha$  of the distribution of  $|T|$  takes the form

$$q_\alpha = \bar{z}_\alpha + n^{-1} p_{21}(\bar{z}_\alpha) + o(n^{-1}).$$

The bootstrap quantile  $q_\alpha^*$  has the Cornish-Fisher expansion

$$\begin{aligned} q_\alpha^* &= \bar{z}_\alpha + n^{-1} p_{21}^*(\bar{z}_\alpha) + o(n^{-1}) \\ &= \bar{z}_\alpha + n^{-1} p_{21}(\bar{z}_\alpha) + o_p(n^{-1}) \\ &= q_\alpha + o_p(n^{-1}) \end{aligned}$$

where  $p_{21}^*(x)$  is the same as  $p_{21}(x)$  except that the moments of  $\mathbf{h}(\mathbf{y})$  are replaced by the corresponding sample moments. The bootstrap test has rejection probability, using the Edgeworth expansion Delta method (Theorem 6.36)

$$\begin{aligned} \mathbb{P}(|T| > q_{1-\alpha}^* | H_0) &= \mathbb{P}(|T| + (q_{1-\alpha} - q_{1-\alpha}^*) > q_{1-\alpha}) \\ &= \mathbb{P}(|T| > q_{1-\alpha}) + o(n^{-1}) \\ &= 1 - \alpha + o(n^{-1}) \end{aligned}$$

as claimed.

## 10.24 Wald-Type Bootstrap Tests

If  $\boldsymbol{\theta}$  is a vector, then to test  $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$  against  $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$  at size  $\alpha$ , a common test is based on the Wald statistic

$$W = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \hat{V}_{\hat{\boldsymbol{\theta}}}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

where  $\hat{\boldsymbol{\theta}}$  is an estimator of  $\boldsymbol{\theta}$  and  $\hat{V}_{\hat{\boldsymbol{\theta}}}$  is a covariance matrix estimator. For a bootstrap test we use the bootstrap algorithm to calculate the critical value.

The bootstrap algorithm samples with replacement from the dataset. Given a bootstrap sample the bootstrap estimator  $\hat{\boldsymbol{\theta}}^*$  and covariance matrix estimator  $\hat{V}_{\hat{\boldsymbol{\theta}}}^*$  are calculated. Given these values the bootstrap Wald statistic is

$$W^* = (\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})' \hat{V}_{\hat{\boldsymbol{\theta}}}^{*-1} (\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}).$$

As for the t-test, it is essential that the bootstrap Wald statistic  $W^*$  is centered at the sample estimator  $\hat{\boldsymbol{\theta}}$  instead of the hypothesized value  $\boldsymbol{\theta}_0$ . This is because  $\hat{\boldsymbol{\theta}}$  is the true value in the bootstrap universe.

Based on  $B$  bootstrap replications we calculate the  $\alpha^{th}$  quantile  $q_\alpha^*$  of the distribution of the bootstrap Wald statistics  $W^*$ . The bootstrap test rejects  $\mathbb{H}_0$  in favor of  $\mathbb{H}_1$  if  $W > q_{1-\alpha}^*$ . More commonly, we calculate a bootstrap p-value. This is

$$p^* = \frac{1}{B} \sum_{b=1}^B \mathbf{1}(W^*(b) > W).$$

The asymptotic performance of the Wald test mimics that of the t-test. In general, the bootstrap Wald test is first-order correct (achieves the correct size asymptotically), and under conditions for which an Edgeworth expansion exists, has accuracy

$$\mathbb{P}(W > q_{1-\alpha}^* | \mathbb{H}_0) = 1 - \alpha + o(n^{-1})$$

and thus achieves a refinement relative to the asymptotic Wald test.

If a reliable covariance matrix estimator  $\hat{V}_{\hat{\boldsymbol{\theta}}}$  is not available, a Wald-type test can be implemented with any positive-definite weight matrix instead of  $\hat{V}_{\hat{\boldsymbol{\theta}}}$ . This includes simple choices such as the identity matrix. The bootstrap algorithm can be used to calculate critical values and p-values for the test. So long as the estimator  $\hat{\boldsymbol{\theta}}$  has an asymptotic distribution, this bootstrap test will be asymptotically first-order valid. The test will not achieve an asymptotic refinement but provides a simple method to test hypotheses when covariance matrix estimates are not available.

## 10.25 Criterion-Based Bootstrap Tests

A criterion-based estimator takes the form

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} J(\boldsymbol{\beta})$$

for some criterion function  $J(\boldsymbol{\beta})$ . This includes least-squares, maximum likelihood, minimum distance, and GMM. Given a hypothesis  $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$  where  $\boldsymbol{\theta} = \mathbf{r}(\boldsymbol{\beta})$ , the restricted estimator which satisfies  $\mathbb{H}_0$  is

$$\tilde{\boldsymbol{\beta}} = \underset{\mathbf{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0}{\operatorname{argmin}} J(\boldsymbol{\beta}).$$

A criterion-based statistic to test  $\mathbb{H}_0$  is

$$\begin{aligned} J &= \min_{\mathbf{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0} J(\boldsymbol{\beta}) - \min_{\boldsymbol{\beta}} J(\boldsymbol{\beta}) \\ &= J(\tilde{\boldsymbol{\beta}}) - J(\hat{\boldsymbol{\beta}}). \end{aligned}$$

A criterion-based test rejects  $\mathbb{H}_0$  for large values of  $J$ . A bootstrap test uses the bootstrap algorithm to calculate the critical value.

In this context we need to be a bit thoughtful about how to construct bootstrap versions of  $J$ . It might seem natural to construct the exact same statistic on the bootstrap samples as on the original sample, but this is incorrect. It makes the same error as calculating a t-ratio or Wald statistic centered at the hypothesized value. In the bootstrap universe, the true value of  $\boldsymbol{\theta}$  is not  $\boldsymbol{\theta}_0$ , rather it is  $\hat{\boldsymbol{\theta}} = \mathbf{r}(\hat{\boldsymbol{\beta}})$ . Thus when using the nonparametric bootstrap, we want to impose the constraint  $\mathbf{r}(\boldsymbol{\beta}) = \mathbf{r}(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\theta}}$  to obtain the bootstrap version of  $J$ .

Thus, the correct way to calculate a bootstrap version of  $J$  is as follows. Generate a bootstrap sample by random sampling from the dataset. Let  $J^*(\boldsymbol{\beta})$  be the bootstrap version of the criterion. On a bootstrap sample calculate the unrestricted estimator

$$\hat{\boldsymbol{\beta}}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} J^*(\boldsymbol{\beta})$$

and the restricted version

$$\tilde{\beta}^* = \underset{r(\beta) = \hat{\theta}}{\operatorname{argmin}} J^*(\beta)$$

where  $\hat{\theta} = r(\hat{\beta})$ . The bootstrap statistic is

$$\begin{aligned} J^* &= \min_{r(\beta) = \hat{\theta}} J^*(\beta) - \min_{\beta} J^*(\beta) \\ &= J^*(\tilde{\beta}^*) - J^*(\hat{\beta}^*). \end{aligned}$$

Calculate  $J^*$  on each bootstrap sample. Take the  $1 - \alpha^{th}$  quantile  $q_{1-\alpha}^*$ . The bootstrap test rejects  $H_0$  in favor of  $H_1$  if  $J > q_{1-\alpha}^*$ . The bootstrap p-value is

$$p^* = \frac{1}{B} \sum_{b=1}^B \mathbf{1}(J^*(b) > J).$$

Special cases of criterion-based tests are minimum distance tests, F tests, and likelihood ratio tests. Take the F test for a linear hypothesis  $\mathbf{R}'\beta = \boldsymbol{\theta}_0$ . The F statistic is

$$F = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/q}{\hat{\sigma}^2/(n-k)}$$

where  $\hat{\sigma}^2$  is the unrestricted estimator of the error variance,  $\tilde{\sigma}^2$  is the restricted estimator,  $q$  is the number of restrictions and  $k$  is the number of estimated coefficients. The bootstrap version of the F statistic is

$$F^* = \frac{(\tilde{\sigma}^{*2} - \hat{\sigma}^{*2})/q}{\hat{\sigma}^{*2}/(n-k)}$$

where  $\hat{\sigma}^{*2}$  is the unrestricted estimator on the bootstrap sample, and  $\tilde{\sigma}^{*2}$  is the restricted estimator which imposes the restriction  $\mathbf{R}'\beta = \hat{\theta} = \mathbf{R}'\hat{\beta}$ .

Take the likelihood ratio (LR) test for the hypothesis  $r(\beta) = \boldsymbol{\theta}_0$ . The LR test statistic is

$$LR = 2(\log L(\hat{\beta}) - \log L(\tilde{\beta}))$$

where  $\hat{\beta}$  is the unrestricted MLE and  $\tilde{\beta}$  is the restricted MLE (imposing  $r(\beta) = \boldsymbol{\theta}_0$ ). The bootstrap version is

$$LR^* = 2(\log L^*(\hat{\beta}^*) - \log L^*(\tilde{\beta}^*))$$

where  $\log L^*(\beta)$  is the log-likelihood function calculated on the bootstrap sample,  $\hat{\beta}^*$  is the unrestricted maximizer, and  $\tilde{\beta}^*$  is the restricted maximizer imposing the restriction  $r(\beta) = r(\hat{\beta})$ .

## 10.26 Parametric Bootstrap

Throughout this chapter we have described the most popular form of the bootstrap known as the nonparametric bootstrap. However there are other forms of the bootstrap algorithm including the parametric bootstrap. This is appropriate when there is a full parametric model for the distribution, as in likelihood estimation.

First, consider the context where the model specifies the full distribution of the random vector  $\mathbf{y}$ , e.g.  $\mathbf{y} \sim F(\mathbf{y} | \beta)$  where the distribution function  $F$  is known but the parameter  $\beta$  is unknown. Let  $\hat{\beta}$  be an estimator of  $\beta$ , such as the maximum likelihood estimator. The parametric bootstrap algorithm generates bootstrap observations  $\mathbf{y}_i^*$  by drawing random vectors from the distribution function  $F(\mathbf{y} | \hat{\beta})$ . When this is done, the true value of  $\beta$  in the bootstrap universe is  $\hat{\beta}$ . Everything which has been discussed in the chapter can be applied using this bootstrap algorithm.

Second, consider the context where the model specifies the conditional distribution of the random vector  $\mathbf{y}$  given the random vector  $\mathbf{x}$ , e.g.  $\mathbf{y} | \mathbf{x} \sim F(\mathbf{y} | \mathbf{x}, \beta)$ . An example is the normal linear regression

model, where  $\mathbf{y} | \mathbf{x} \sim N(\mathbf{x}'\boldsymbol{\beta}, \sigma^2)$ . In this context we can hold the regressors  $\mathbf{x}_i$  fixed and then draw the bootstrap observations  $\mathbf{y}_i^*$  from the conditional distribution  $F(\mathbf{y} | \mathbf{x}_i, \tilde{\boldsymbol{\beta}})$ . In the example of the normal regression model this is equivalent to drawing a normal error  $e_i^* \sim N(0, \hat{\sigma}^2)$  and then setting  $y_i^* = \mathbf{x}_i'\hat{\boldsymbol{\beta}} + e_i^*$ . Again, in this algorithm the true value of  $\boldsymbol{\beta}$  is  $\hat{\boldsymbol{\beta}}$  and everything which is discussed in this chapter can be applied as before.

Third, consider tests of the hypothesis  $\mathbf{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$ . In this context we can also construct a restricted estimator  $\tilde{\boldsymbol{\beta}}$  (for example the restricted MLE) which satisfies the hypothesis  $\mathbf{r}(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\theta}_0$ . Then we can alternatively generate bootstrap samples by simulating from the distribution  $\mathbf{y}_i^* \sim F(\mathbf{y} | \tilde{\boldsymbol{\beta}})$ , or in the conditional context from  $\mathbf{y}_i^* \sim F(\mathbf{y} | \mathbf{x}_i, \tilde{\boldsymbol{\beta}})$ . When this is done, the true value of  $\boldsymbol{\beta}$  in the bootstrap is  $\tilde{\boldsymbol{\beta}}$  which satisfies the hypothesis. So in this context the correct values of the bootstrap statistics are

$$\begin{aligned} T^* &= \frac{\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0}{s(\hat{\boldsymbol{\theta}}^*)} \\ W^* &= (\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0)' \hat{V}_{\hat{\boldsymbol{\theta}}}^{*-1} (\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0) \\ J^* &= \min_{\mathbf{r}(\boldsymbol{\beta})=\boldsymbol{\theta}_0} J^*(\boldsymbol{\beta}) - \min_{\boldsymbol{\beta}} J^*(\boldsymbol{\beta}) \\ LR^* &= 2 \left( \max_{\boldsymbol{\beta}} \log L^*(\boldsymbol{\beta}) - \max_{\mathbf{r}(\boldsymbol{\beta})=\boldsymbol{\theta}_0} \log L^*(\boldsymbol{\beta}) \right) \end{aligned}$$

and

$$F^* = \frac{(\tilde{\sigma}^{*2} - \hat{\sigma}^{*2})/q}{\hat{\sigma}^{*2}/(n-k)}$$

where  $\hat{\sigma}^{*2}$  is the unrestricted estimator on the bootstrap sample, and  $\tilde{\sigma}^{*2}$  is the restricted estimator which imposes the restriction  $\mathbf{R}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$ .

The primary advantage of the parametric bootstrap (relative to the nonparametric bootstrap) is that it will be more accurate when the parametric model is correct. This may be quite important in small samples. The primary disadvantage of the parameric bootstrap is that it can be inaccurate when the parametric model is incorrect.

## 10.27 How Many Bootstrap Replications?

How many bootstrap replications should be used? There is no universally correct answer as there is a trade-off between accuracy and computation cost. Computation cost is essentially linear in  $B$ . Accuracy (either standard errors or p-values) is proportional to  $B^{-1/2}$ . Improved accuracy can be obtained but only at a higher computational cost.

In most empirical research, most calculations are quick and investigatory, not requiring full accuracy. But final results (those going into the final version of the paper) should be accurate. Thus it seems reasonable to use asymptotic and/or bootstrap methods with a modest number of replications for daily calculations, but use a much larger  $B$  for the final version.

In particular, for final calculations,  $B = 10,000$  is desired, with  $B = 1000$  a minimal choice. In contrast, for daily quick calculations values as low as  $B = 100$  may be sufficient for rough estimates.

A useful way to think about the accuracy of bootstrap methods stems from the calculation of p-values. The bootstrap p-value  $p^*$  is an average of  $B$  Bernoulli draws. The variance of the simulation estimator of  $p^*$  is  $p^*(1-p^*)/B$ , which is bounded below  $1/4B$ . To calculate the p-value within, say, 0.01 of the true value with 95% probability requires a standard error below 0.005. This is ensured if  $B \geq 10,000$ .

Stata by default sets  $B = 50$ . This is useful for verification that a program runs, but is a poor choice for empirical reporting. Make sure that you set  $B$  to the value you want.

## 10.28 Setting the Bootstrap Seed

Computers do not generate true random numbers, but rather pseudo-random numbers generated by a deterministic algorithm. The algorithms generate sequences which are indistinguishable from random sequences, so this is not a worry for bootstrap applications.

The methods, however, necessarily require a starting value known as a “seed”. Most packages implement this with a default seed which is reset each time the statistical package is started. This means if you start the package fresh, run a bootstrap program (e.g. a do file in Stata), exit the package, restart the package and then rerun the bootstrap program, you should obtain exactly the same results. If you instead run the bootstrap program (e.g. do file) twice sequentially without restarting the package, the seed is not reset so a different set of pseudo-random numbers will be generated, and the results from the two runs will be different.

Packages allow users to set their own seed. (In Stata, the command is `set seed #` where # is a number. In Matlab the command is `rng(#)`.) If the seed is set to a specific number at the start of a file, then the exact same pseudo-random numbers will be generated each time the program is run. If this is the case, the results of a bootstrap calculation (standard error or test) will be identical across computer runs.

The fact that the bootstrap results can be fixed by setting the seed in the replication file has motivated many researchers to follow this choice. They set the seed at the start of the replication file so that repeated executions result in the same numerical findings.

Fixing seeds, however, should be done cautiously. It may be a wise choice for a final calculation (when a paper is finished) but is an unwise choice for daily calculations. If you use a small number of replications, say  $B = 100$ , in your preliminary work, the bootstrap calculations will be quite inaccurate. But as you run your results again and again (as is typical in empirical projects) you will find the same numerical standard errors and test results, giving you a false sense of stability and accuracy. If instead a different seed is used each time the program is run then the bootstrap results will vary across runs, and you will observe that the results vary across these runs, giving you important and meaningful information about the (lack of) accuracy in your results. One way to ensure this in Matlab is to use the command `rng('shuffle')` which sets the seed according to the current clock.

These considerations lead to a recommended hybrid approach. For daily empirical investigations, do not fix the bootstrap set in your program, unless you have it set by the clock. For your final calculations set the seed to a specific arbitrary choice, and also set  $B = 10,000$  so that the results are insensitive to the seed.

## 10.29 Bootstrap Regression

A major focus of this textbook has been on the least-squares estimator  $\hat{\beta}$  in the projection model. The bootstrap can be used to calculate standard errors and confidence intervals for smooth functions of the coefficient estimates.

The nonparametric bootstrap algorithm, as described before, samples observations randomly with replacement from the dataset, creating the bootstrap sample  $\{(y_1^*, \mathbf{x}_1^*), \dots, (y_n^*, \mathbf{x}_n^*)\}$ , or in matrix notation  $(\mathbf{y}^*, \mathbf{X}^*)$ . It is important to recognize that entire observations (pairs of  $y_i$  and  $\mathbf{x}_i$ ) are sampled. This is often called the **pairs bootstrap**.

Given this bootstrap sample, we calculate the regression estimator

$$\hat{\beta}^* = (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} (\mathbf{X}^{*\prime} \mathbf{y}^*). \quad (10.32)$$

This is repeated  $B$  times. The bootstrap standard errors are the standard deviations across the draws, and confidence intervals are constructed from the empirical quantiles across the draws.

What is the nature of the bootstrap distribution of  $\hat{\beta}^*$ ? It is useful to start with the distribution of the bootstrap observations  $(y_i^*, \mathbf{x}_i^*)$ , which is the discrete distribution which puts mass  $1/n$  on each observation pair  $(y_i, \mathbf{x}_i)$ . The bootstrap universe can be thought of as the empirical scatter plot of the

observations. The true value of the projection coefficient in this bootstrap universe is

$$(\mathbb{E}^* (\mathbf{x}_i^* \mathbf{x}_i^{*\prime}))^{-1} (\mathbb{E}^* (\mathbf{x}_i^* y_i^*)) = \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^* \mathbf{x}_i' \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^* y_i \right) = \hat{\boldsymbol{\beta}}.$$

We see that the true value in the bootstrap distribution is the least-squares estimate  $\hat{\boldsymbol{\beta}}$ .

The bootstrap observations satisfy the projection equation

$$\begin{aligned} y_i^* &= \mathbf{x}_i^{*\prime} \hat{\boldsymbol{\beta}} + e_i^* \\ \mathbb{E}^* (\mathbf{x}_i^* e_i^*) &= \mathbf{0}. \end{aligned} \tag{10.33}$$

For each bootstrap pair  $(y_i^*, \mathbf{x}_i^*) = (y_j, \mathbf{x}_j)$  the true error  $e_i^* = \hat{e}_j$  equals the least-squares residual from the original dataset. This is because each bootstrap pair corresponds to an actual observation.

A technical problem (which is typically ignored) is that it is possible for  $\mathbf{X}^{*\prime} \mathbf{X}^*$  to be singular in a simulated bootstrap sample, in which case the least-squares estimator  $\hat{\boldsymbol{\beta}}^*$  cannot be defined. Indeed, the probability is always positive that  $\mathbf{X}^{*\prime} \mathbf{X}^*$  is singular. For example, the probability that a bootstrap sample consists entirely of one observation repeated  $n$  times is  $n^{-(n-1)}$ . This is a small probability, but positive. A more significant example is sparse dummy variable designs where it is possible to draw an entire sample with only one observed value for the dummy variable. For example, if a sample has  $n = 20$  observations with a dummy variable with treatment (equals 1) for only three of the 20 observations, the probability is 4% that a bootstrap sample contains entirely non-treated values (all 0's). 4% is quite high!

The standard approach to circumvent this problem is to compute  $\hat{\boldsymbol{\beta}}^*$  only if  $\mathbf{X}^{*\prime} \mathbf{X}^*$  is non-singular as defined by a conventional numerical tolerance and treat it as missing otherwise. A better solution is to define a tolerance which bounds  $\mathbf{X}^{*\prime} \mathbf{X}^*$  away from non-singularity. Define the ratio of the smallest eigenvalue of the bootstrap design matrix to that of the data design matrix

$$\lambda^* = \frac{\lambda_{\min}(\mathbf{X}^{*\prime} \mathbf{X}^*)}{\lambda_{\min}(\mathbf{X}' \mathbf{X})}.$$

If, in a given bootstrap replication,  $\lambda^* < \tau$  is smaller than a given tolerance (Shao and Tu (1995, p. 291) recommend  $\tau = 1/2$ ) then the estimator can be treated as missing, or we can define the trimming rule

$$\hat{\boldsymbol{\beta}}^* = \begin{cases} \hat{\boldsymbol{\beta}}^* & \text{if } \lambda^* \geq \tau \\ \hat{\boldsymbol{\beta}} & \text{if } \lambda^* < \tau. \end{cases} \tag{10.34}$$

This ensures that the bootstrap estimator  $\hat{\boldsymbol{\beta}}^*$  will be well behaved.

### 10.30 Bootstrap Regression Asymptotic Theory

Define the least-squares estimator  $\hat{\boldsymbol{\beta}}$ , its bootstrap version  $\hat{\boldsymbol{\beta}}^*$  as in (10.32), and the transformations  $\hat{\boldsymbol{\theta}} = \mathbf{g}(\hat{\boldsymbol{\beta}})$  and  $\hat{\boldsymbol{\theta}}^* = \mathbf{r}(\hat{\boldsymbol{\beta}}^*)$  for some smooth transformation  $\mathbf{r}$ . Let  $\hat{V}_{\boldsymbol{\beta}}$  and  $\hat{V}_{\boldsymbol{\theta}}$  denote heteroskedasticity-robust covariance matrix estimators for  $\hat{\boldsymbol{\beta}}$  and  $\hat{\boldsymbol{\theta}}$ , and let  $\hat{V}_{\boldsymbol{\beta}}^*$  and  $\hat{V}_{\boldsymbol{\theta}}^*$  be their bootstrap versions. When  $\theta$  is scalar define the standard errors  $s(\hat{\theta}) = \sqrt{n^{-1} \hat{V}_{\theta}}$  and  $s(\hat{\theta}^*) = \sqrt{n^{-1} \hat{V}_{\theta}^*}$ . Define the t-ratios  $T = (\hat{\theta} - \theta) / s(\hat{\theta})$  and bootstrap version  $T^* = (\hat{\theta}^* - \hat{\theta}) / s(\hat{\theta}^*)$ . We are interested in the asymptotic distributions of  $\hat{\boldsymbol{\beta}}^*$ ,  $\hat{\boldsymbol{\theta}}^*$  and  $T^*$ .

Since the bootstrap observations satisfy the model (10.33), we see by standard calculations that

$$\sqrt{n} (\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) = \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^* \mathbf{x}_i'^* \right)^{-1} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_i^* e_i^* \right).$$

By the bootstrap WLLN

$$\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^* \mathbf{x}_i^{*'} \xrightarrow{p^*} \mathbb{E}(\mathbf{x}_i \mathbf{x}_i') = \mathbf{Q}$$

and by the bootstrap CLT

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_i^* e_i^* \xrightarrow{d^*} N(\mathbf{0}, \boldsymbol{\Omega})$$

where  $\boldsymbol{\Omega} = \mathbb{E}(\mathbf{x}_i \mathbf{x}_i' e_i^2)$ . Again applying the bootstrap WLLN we obtain

$$\hat{V}_{\beta} \xrightarrow{p^*} V_{\beta} = \mathbf{Q}^{-1} \boldsymbol{\Omega} \mathbf{Q}^{-1}$$

and

$$\hat{V}_{\theta} \xrightarrow{p^*} V_{\theta} = \mathbf{R}' V_{\beta} \mathbf{R}$$

where  $\mathbf{R} = \mathbf{R}(\beta)$ .

Combining with the bootstrap CMT and delta method we establish the asymptotic distribution of the bootstrap regression estimator.

**Theorem 10.22** Under Assumption 7.2, as  $n \rightarrow \infty$

$$\sqrt{n} (\hat{\beta}^* - \hat{\beta}) \xrightarrow{d^*} N(\mathbf{0}, V_{\beta}).$$

If Assumption 7.3 also holds then

$$\sqrt{n} (\hat{\theta}^* - \hat{\theta}) \xrightarrow{d^*} N(\mathbf{0}, V_{\theta}).$$

If Assumption 7.4 also holds then

$$T^* \xrightarrow{d^*} N(0, 1).$$

This means that the bootstrap confidence interval and testing methods all apply for inference on  $\beta$  and  $\theta$ . This includes the percentile, BC percentile,  $BC_a$ , and percentile-t intervals, and hypothesis tests based on t-tests, Wald tests, MD tests, LR tests and F tests.

To justify the use of bootstrap standard errors we also need to verify the uniform square integrability of  $\hat{\beta}^*$  and  $\hat{\theta}^*$ . This is technically challenging because the least-squares estimator involves division (matrix inversion) which is not a globally continuous function. A partial solution is to use the trimmed estimator (10.34). This bounds the moments of  $\hat{\beta}^*$  by those of  $n^{-1} \sum_{i=1}^n \mathbf{x}_i^* e_i^*$ . Since this is a sample mean, Theorem 10.14 applies and  $\hat{V}_{\beta}^*$  is bootstrap consistent for  $V_{\beta}$ . However, this does not ensure that  $\hat{V}_{\theta}^*$  will be consistent for  $\hat{V}_{\theta}$  unless the function  $r(\mathbf{u})$  satisfies the conditions of Theorem 10.14. For general applications we should use a trimmed estimator for the bootstrap variance. For some  $\tau_n = O(e^{n/8})$  define

$$\begin{aligned} z_n^* &= \sqrt{n} (\hat{\theta}^* - \hat{\theta}) \\ z^{**} &= z^* \mathbf{1}(\|z_n^*\| \leq \tau_n) \\ \bar{z}^{**} &= \frac{1}{B} \sum_{b=1}^B z^{**}(b) \\ \hat{V}_{\theta}^{\text{boot}, \tau} &= \frac{1}{B-1} \sum_{b=1}^B (\bar{z}^{**}(b) - \bar{z}^{**})(\bar{z}^{**}(b) - \bar{z}^{**})'. \end{aligned}$$

The matrix  $\widehat{V}_{\boldsymbol{\theta}}^{\text{boot}}$  is a trimmed bootstrap estimator of the variance of  $z_n = \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ . The associated bootstrap standard error for  $\widehat{\boldsymbol{\theta}}$  (in the scalar case) is  $s(\widehat{\boldsymbol{\theta}}) = \sqrt{n^{-1}\widehat{V}_{\boldsymbol{\theta}}^{\text{boot}}}$ .

By an application of Theorems 10.15 and 10.16, we find that this estimator  $\widehat{V}_{\boldsymbol{\theta}}^{\text{boot}}$  is consistent for the asymptotic variance.

**Theorem 10.23** Under Assumption 7.2 and 7.3, as  $n \rightarrow \infty$

$$\widehat{V}_{\boldsymbol{\theta}}^{\text{boot},\tau} \xrightarrow{P^*} V_{\boldsymbol{\theta}}.$$

Programs such as Stata use the untrimmed estimator  $\widehat{V}_{\boldsymbol{\theta}}^{\text{boot}}$  rather than the trimmed estimator  $\widehat{V}_{\boldsymbol{\theta}}^{\text{boot},\tau}$ . This means that we should be cautious about interpreting reported bootstrap standard errors especially for nonlinear functions such as ratios.

### 10.31 Wild Bootstrap

Take the linear regression model

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(e_i | \mathbf{x}_i) &= 0. \end{aligned}$$

What is special about this model is the conditional mean restriction. The nonparametric bootstrap (which samples the pairs  $(y_i^*, \mathbf{x}_i^*)$  i.i.d. from the original observations) does not make use of this restriction. Consequently the bootstrap distribution for  $(y_i^*, \mathbf{x}_i^*)$  does not satisfy the conditional mean restriction, and therefore does not satisfy the linear regression assumption. To improve the precision of the bootstrap method it seems reasonable to impose the conditional mean restriction on the bootstrap distribution.

A natural approach is to hold the regressors  $\mathbf{x}_i$  fixed and then draw the errors  $e_i^*$  in some way which imposes a conditional mean of zero. The simplest approach is to draw the errors independent from the regressors, perhaps from the empirical distribution of the residuals. This procedure is known as the **residual bootstrap**. However, this imposes independence of the errors from the regressors, which is much stronger than the conditional mean assumption. This is generally undesirable.

A method which imposes the conditional mean restriction while allowing general heteroskedasticity is the **wild bootstrap**. It was proposed by Liu (1988) and extended by Mammon (1993). The method uses auxiliary random variables  $\xi_i^*$  which are i.i.d., mean zero, and variance 1. The bootstrap observations are then generated as

$$\begin{aligned} y_i^* &= \mathbf{x}'_i \widehat{\boldsymbol{\beta}} + e_i^* \\ e_i^* &= \widehat{e}_i \xi_i^* \end{aligned}$$

where the regressors  $\mathbf{x}_i$  are held fixed at their sample values,  $\widehat{\boldsymbol{\beta}}$  is the sample least-squares estimator, and  $\widehat{e}_i$  are the least-squares residuals, which are also held fixed at their sample values.

This algorithm generates bootstrap errors  $e_i^*$  which are conditionally mean zero. Thus the bootstrap pairs  $(y_i^*, \mathbf{x}_i)$  satisfy a linear regression, with the “true” coefficient of  $\widehat{\boldsymbol{\beta}}$ . The conditional variance of the wild bootstrap errors  $e_i^*$  are

$$\mathbb{E}^*(e_i^{*2} | \mathbf{x}_i) = \widehat{e}_i^2.$$

This means that the conditional variance of the bootstrap estimator  $\widehat{\boldsymbol{\beta}}^*$  is

$$\mathbb{E}^*\left((\widehat{\boldsymbol{\beta}}^* - \widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}^* - \widehat{\boldsymbol{\beta}})' | \mathbf{X}\right) = (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \widehat{e}_i^2 \right) (\mathbf{X}' \mathbf{X})^{-1}$$

which is the White estimator of the variance of  $\hat{\beta}$ . Thus the wild bootstrap replicates the appropriate first and second moments of the distribution.

Two distributions have been proposed for the auxiliary variables  $\xi_i^*$  both of which are two-point discrete distributions. The first are **Rademacher** random variables, which satisfy

$$\begin{aligned}\mathbb{P}(\xi_i^* = 1) &= \frac{1}{2} \\ \mathbb{P}(\xi_i^* = -1) &= \frac{1}{2}.\end{aligned}$$

The second is the Mammen (1993) two-point distribution

$$\begin{aligned}\mathbb{P}\left(\xi_i^* = \frac{1+\sqrt{5}}{2}\right) &= \frac{\sqrt{5}-1}{2\sqrt{5}} \\ \mathbb{P}\left(\xi_i^* = \frac{1-\sqrt{5}}{2}\right) &= \frac{\sqrt{5}+1}{2\sqrt{5}}.\end{aligned}$$

The reasoning behind the Mammen distribution is that this choice implies  $\mathbb{E}(\xi_i^{*3}) = 1$ , which implies that the third central moment of  $\hat{\beta}^*$  matches the natural nonparametric estimator of the third central moment of  $\hat{\beta}$ . Since the wild bootstrap matches the first three moments, the percentile-t interval and one-sided t-tests can be shown to achieve asymptotic refinements.

The reasoning behind the Rademacher distribution is that this choice implies  $\mathbb{E}(\xi_i^{*4}) = 1$ , which implies that the fourth central moment of  $\hat{\beta}^*$  matches the natural nonparametric estimator of the fourth central moment of  $\hat{\beta}$ . If the regression errors  $e_i$  are symmetrically distributed (so the third moment is zero) then the first four moments are matched. In this case the wild bootstrap should have even better performance, and additionally two-sided t-tests can be shown to achieve an asymptotic refinement. When the regression error is not symmetrically distributed these asymptotic refinements are not achieved. However, simulation evidence for one-sided t-tests presented in Davidson and Flachaire (2008) suggest that the Rademacher distribution (used with the restricted wild bootstrap) overall has the best performance and is the preferred choice.

For hypothesis testing improved precision can be obtained by the **restricted wild bootstrap**. Consider tests of the hypothesis

$$\mathbb{H}_0 : \mathbf{r}(\boldsymbol{\beta}) = \mathbf{0}.$$

Let  $\tilde{\beta}$  be a CLS or EMD estimator of  $\beta$  subject to the restriction  $\mathbf{r}(\tilde{\beta}) = \mathbf{0}$ . Let  $\tilde{e}_i = y_i - \mathbf{x}'_i \tilde{\beta}$  be the constrained residuals. The restricted wild bootstrap algorithm generates observations as

$$\begin{aligned}y_i^* &= \mathbf{x}'_i \tilde{\beta} + e_i^* \\ e_i^* &= \tilde{e}_i \xi_i^*.\end{aligned}$$

With this modification,  $\tilde{\beta}$  is the true value in the bootstrap universe, so the null hypothesis  $\mathbb{H}_0$  holds. Thus bootstrap tests are constructed the same as for the parametric bootstrap using a restricted parameter estimator.

## 10.32 Bootstrap for Clustered Observations

Bootstrap methods can also be applied in the context of clustered observations, though the methodological literature is relatively thin. Here we review methods discussed in Cameron, Gelbach and Miller (2008).

Let  $\mathbf{y}_g = (y_{1g}, \dots, y_{n_g g})'$  and  $\mathbf{X}_g = (\mathbf{x}_{1g}, \dots, \mathbf{x}_{n_g g})'$  denote the  $n_g \times 1$  vector of dependent variables and  $n_g \times k$  matrix of regressors for the  $g^{th}$  cluster. A linear regression model using cluster notation is

$$\mathbf{y}_g = \mathbf{X}_g \boldsymbol{\beta} + \mathbf{e}_g$$

where  $\mathbf{e}_g = (e_{1g}, \dots, e_{n_g g})'$  is a  $n_g \times 1$  error vector. The sample has  $G$  cluster pairs  $(\mathbf{y}_g, \mathbf{X}_g)$ .

The **pairs cluster bootstrap** samples  $G$  cluster pairs  $(\mathbf{y}_g, \mathbf{X}_g)$  to create the bootstrap sample. Least-squares is applied to the bootstrap sample to obtain the coefficient estimators. By repeating  $B$  times, bootstrap standard errors for coefficients estimates, or functions of the coefficient estimates, can be calculated. Percentile, BC percentile, and  $BC_a$  confidence intervals can be calculated.

The  $BC_a$  interval requires an estimator of the acceleration coefficient  $a$  which is a scaled jackknife estimate of the third moment of the estimator. In the context of clustered observations the delete-cluster jackknife should be used for estimation of  $a$ .

Furthermore, on each bootstrap sample the cluster-robust standard errors can be calculated and used to compute bootstrap t-ratios, from which percentile-t confidence intervals can be calculated.

The **wild cluster bootstrap** fixes the clusters and regressors, and generates the bootstrap observations as

$$\begin{aligned}\mathbf{y}_g^* &= \mathbf{X}_g \hat{\boldsymbol{\beta}} + \mathbf{e}_g^* \\ \mathbf{e}_g^* &= \hat{\mathbf{e}}_i \xi_g^*\end{aligned}$$

where  $\xi_g^*$  is a scalar auxiliary random variable as described in the previous section. Notice that  $\xi_g^*$  is interacted with the entire vector of residuals from cluster  $g$ . Cameron, Gelbach and Miller (2008) follow the recommendation of Davidson and Flachaire (2008) and use Rademacher random variables for  $\xi_g^*$ .

For hypothesis testing, Cameron, Gelbach and Miller (2008) recommend the **restricted wild cluster bootstrap**. For tests of

$$\mathbb{H}_0 : \mathbf{r}(\boldsymbol{\beta}) = \mathbf{0}$$

let  $\tilde{\boldsymbol{\beta}}$  be a CLS or EMD estimator of  $\boldsymbol{\beta}$  subject to the restriction  $\mathbf{r}(\tilde{\boldsymbol{\beta}}) = \mathbf{0}$ . Let  $\tilde{\mathbf{e}}_g = \mathbf{y}_g - \mathbf{X}_g \tilde{\boldsymbol{\beta}}$  be the constrained cluster-level residuals. The restricted wild cluster bootstrap algorithm generates observations as

$$\begin{aligned}\mathbf{y}_g^* &= \mathbf{X}_g \tilde{\boldsymbol{\beta}} + \mathbf{e}_g^* \\ \mathbf{e}_g^* &= \tilde{\mathbf{e}}_i \xi_g^*.\end{aligned}$$

On each bootstrap sample the test statistic for  $\mathbb{H}_0$  (t-ratio, Wald, LR, or F) is applied. Since the bootstrap algorithm satisfies  $\mathbb{H}_0$  these statistics are centered at the hypothesized value. p-values are then calculated conventionally and used to assess the significance of the test statistic.

There are several reasons why conventional asymptotic approximations may work poorly with clustered observations. First, while the sample size  $n$  may be large, the effective sample size is the number of clusters  $G$ . This is because when the dependence structure within each cluster is unconstrained the central limit theorem effectively treats each cluster as a single observation. Thus, if  $G$  is small we should treat inference as a small sample problem. Second, cluster-robust covariance matrix estimation explicitly treats each cluster as a single observation. Consequently the accuracy of normal approximations to t-ratios and Wald statistics is more accurately viewed as a small sample distribution problem. Third, when cluster sizes  $n_g$  are heterogeneous, this means that the estimation problems just described also involve heterogeneous variances. Specifically, heterogeneous cluster sizes induces a high degree of effective heteroskedasticity (since the variance of a within-cluster sum is proportional to  $n_g$ ). When  $G$  is small this means that cluster-robust inference is similar to finite-sample inference with a small heteroskedastic sample. Fourth, interest often concerns treatment which is applied at the level of a cluster (such as the effect of tracking discussed in Section 4.21). If the number of treated clusters is small, this is equivalent to estimation with a highly sparse dummy variable design, in which case cluster-robust covariance matrix estimation can be unreliable.

These concerns suggest that conventional normal approximations may be poor in the context of clustered observations with a small number of groups  $G$ , motivating instead the use of bootstrap methods. However, these concerns also can cause challenges with the accuracy of bootstrap approximations. When the number of clusters  $G$  is small, the cluster sizes  $n_g$  heterogeneous, or the number of treated

clusters small, bootstrap methods may also be inaccurate. In such cases inference should proceed cautiously.

To illustrate the use of the pairs cluster bootstrap, Table 10.4 reports the estimates of the example from Section 4.21 of the effect of tracking on testscores from Duflo, Dupas and Kremer (2011). In addition to the asymptotic cluster standard error, we report the cluster jackknife and cluster bootstrap standard errors, as well as three percentile-type confidence intervals and using 10,000 bootstrap replications. In this example the asymptotic, jackknife, and cluster bootstrap standard errors are identical, which reflects the good balance of this particular regression design.

Table 10.4: Comparison of Methods for Estimate of Effect of Tracking

Coefficient on <i>Tracking</i>	0.138
Asymptotic cluster s.e.	(0.078)
Jackknife cluster s.e.	(0.078)
Cluster Bootstrap s.e.	(0.078)
95% Percentile Interval	[-0.013, 0.291]
95% BC Percentile Interval	[-0.015, 0.289]
95% BC <sub>a</sub> Percentile Interval	[-0.018, 0.286]

In Stata, to obtain cluster bootstrap standard errors and confidence intervals use the options `cluster(id)` `vce(bootstrap, reps(#))`, where `id` is the cluster variable and `#` is the number of bootstrap replications.

### 10.33 Technical Proofs\*

**Proof of Theorem 10.1:** We present a case for the one-dimensional case. Fix  $\varepsilon > 0$  and set  $J = 1/\varepsilon$ . Define the left-limits  $F(u-) = \lim_{t \uparrow u} F(t) = \mathbb{E}(\mathbf{1}(y < u))$ . We can find points  $-\infty = u_0 < u_1 < \dots < u_J = \infty$  such that

$$F(u_j-) - F(u_{j-1}) \leq \varepsilon. \quad (10.35)$$

By the WLLN and  $J$  is fixed there is an  $n$  sufficiently large such that with probability exceeding  $1 - \varepsilon$ ,

$$\max_{j \leq J} |F_n(u_j-) - F(u_j-)| = \max_{j \leq J} \left| \frac{1}{n} \sum_{i=1}^n (\mathbf{1}(y_i < u_j) - \mathbb{E}(\mathbf{1}(y < u_j))) \right| \leq \varepsilon \quad (10.36)$$

and

$$\max_{j \leq J} |F_n(u_j) - F(u_j)| = \max_{j \leq J} \left| \frac{1}{n} \sum_{i=1}^n (\mathbf{1}(y_i \leq u_j) - \mathbb{E}(\mathbf{1}(y \leq u_j))) \right| \leq \varepsilon. \quad (10.37)$$

Since both  $F_n(u)$  and  $F(u)$  are weakly monotonically increasing, for any  $u$  satisfying  $u_{j-1} \leq u < u_j$

$$F_n(u) - F(u) \leq F_n(u_j-) - F(u_{j-1}) \leq F_n(u_j-) - F(u_j-) + \varepsilon \leq 2\varepsilon.$$

The second inequality is (10.35) and the final inequality holds on the event (10.36).

Similarly, on the event (10.37)

$$F_n(u) - F(u) \geq F_n(u_{j-1}) - F(u_j-) \geq F_n(u_{j-1}) - F(u_{j-1}) - \varepsilon \geq -2\varepsilon.$$

We have shown that for any  $u$ ,  $|F_n(u) - F(u)| \leq 2\varepsilon$  with probability exceeding  $1 - \varepsilon$ . Since  $\varepsilon$  is arbitrary this shows  $\sup_u |F_n(u) - F(u)| \xrightarrow{P} 0$  as  $n \rightarrow \infty$ . ■

**Proof of Theorem 10.4:** Fix  $\varepsilon > 0$ .

Set  $\delta_1 = F(q_\alpha) - F(q_\alpha - \varepsilon)$ . Note that  $\delta_1 > 0$  by the definition of  $q_\alpha$  and the assumption  $\alpha = F(q_\alpha) > 0$ . The WLLN implies that

$$F_n(q_\alpha - \varepsilon) - F(q_\alpha - \varepsilon) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(y_i \leq q_\alpha - \varepsilon) - \mathbb{E}(\mathbf{1}(y \leq q_\alpha - \varepsilon)) \xrightarrow{p} 0$$

which means that there is a  $\bar{n}_1 < \infty$  such that for all  $n \geq \bar{n}_1$

$$\mathbb{P}(|F(q_\alpha - \varepsilon) - F_n(q_\alpha - \varepsilon)| > \delta_1/2) \leq \varepsilon.$$

Assume as well that  $\bar{n}_1 > 2/\delta_1$ . The inequality  $\hat{q}_\alpha \geq y_{(j-1)}$  means that  $\hat{q}_\alpha < q_\alpha - \varepsilon$  implies

$$F_n(q_\alpha - \varepsilon) \geq (j-1)/n \geq \alpha - 1/n.$$

Thus for all  $n \geq \bar{n}_1$

$$\begin{aligned} \mathbb{P}(\hat{q}_\alpha < q_\alpha - \varepsilon) &\leq \mathbb{P}(F_n(q_\alpha - \varepsilon) \geq \alpha - 1/n) \\ &= \mathbb{P}(F_n(q_\alpha - \varepsilon) - F(q_\alpha - \varepsilon) \geq \delta_1 - 1/n) \\ &\leq \mathbb{P}(|F_n(q_\alpha - \varepsilon) - F(q_\alpha - \varepsilon)| > \delta_1/2) \leq \varepsilon. \end{aligned}$$

Now set  $\delta_2 = F^+(q_\alpha^+ + \varepsilon) - F^+(q_\alpha^+)$ . Note that  $\delta_2 > 0$  by the definition of  $q_\alpha^+$  and the assumption  $\alpha = F^+(q_\alpha^+) < 1$ . The WLLN implies that

$$F_n^+(q_\alpha + \varepsilon) - F^+(q_\alpha + \varepsilon) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(y_i < q_\alpha + \varepsilon) - \mathbb{E}(\mathbf{1}(y < q_\alpha + \varepsilon)) \xrightarrow{p} 0$$

which means that there is a  $\bar{n}_2 < \infty$  such that for all  $n \geq \bar{n}_2$

$$\mathbb{P}(|F_n^+(q_\alpha + \varepsilon) - F^+(q_\alpha + \varepsilon)| > \delta_2/2) \leq \varepsilon.$$

Again assume that  $\bar{n}_2 > 2/\delta_2$ . The inequality  $\hat{q}_\alpha \leq y_{(j+1)}$  means that  $\hat{q}_\alpha > q_\alpha^+ + \varepsilon$  implies

$$F_n^+(q_\alpha^+ + \varepsilon) \leq j/n \leq \alpha + 1/n.$$

Thus for all  $n \geq \bar{n}_2$

$$\begin{aligned} \mathbb{P}(\hat{q}_\alpha > q_\alpha^+ + \varepsilon) &\leq \mathbb{P}(F_n^+(q_\alpha^+ + \varepsilon) \leq \alpha + 1/n) \\ &\leq \mathbb{P}(F^+(q_\alpha^+ + \varepsilon) - F_n^+(q_\alpha^+ + \varepsilon) > \delta_2/2) \\ &\leq \mathbb{P}(|F_n^+(q_\alpha + \varepsilon) - F^+(q_\alpha + \varepsilon)| > \delta_2/2) \leq \varepsilon. \end{aligned}$$

We have shown that for all  $n \geq \max[\bar{n}_1, \bar{n}_2]$

$$\mathbb{P}(q_\alpha - \varepsilon \leq \hat{q}_\alpha \leq q_\alpha^+ + \varepsilon) \geq 1 - 2\varepsilon$$

which establishes the result. ■

**Proof of Theorem 10.5:** Fix  $\varepsilon > 0$ . Since  $\mathbf{z}_n \xrightarrow{p} \mathbf{z}$  there is an  $n$  sufficiently large such that

$$\mathbb{P}(\|\mathbf{z}_n - \mathbf{z}\| > \varepsilon) < \varepsilon.$$

Since the event  $\|\mathbf{z}_n - \mathbf{z}\| > \varepsilon$  is non-random under the conditional probability  $\mathbb{P}^*$ , for such  $n$ ,

$$\mathbb{P}^*(\|\mathbf{z}_n - \mathbf{z}\| > \varepsilon) = \begin{cases} 0 & \text{with probability exceeding } 1 - \varepsilon \\ 1 & \text{with probability less than } \varepsilon \end{cases}.$$

Since  $\varepsilon$  is arbitrary we conclude  $\mathbb{P}^*(\|\mathbf{z}_n - \mathbf{z}\| > \varepsilon) \xrightarrow{p} 0$  as required. ■

**Proof of Theorem 10.6:** Fix  $\varepsilon > 0$ . By Markov's inequality (B.35), the facts (10.12) and (10.13), and finally the Marcinkiewicz WLLN (Theorem 6.39) with  $r = 2$  and  $y_i = \|\mathbf{y}_i\|$ ,

$$\begin{aligned}\mathbb{P}^*(\|\bar{\mathbf{y}}^* - \bar{\mathbf{y}}\| > \varepsilon) &\leq \varepsilon^{-2} \mathbb{E}^* \|\bar{\mathbf{y}}^* - \bar{\mathbf{y}}\|^2 \\ &= \varepsilon^{-2} \text{tr}(\text{var}^*(\bar{\mathbf{y}}^*)) \\ &= \varepsilon^{-2} \text{tr}\left(\frac{1}{n} \hat{\Sigma}\right) \\ &\leq \varepsilon^{-2} n^{-2} \sum_{i=1}^n \mathbf{y}'_i \mathbf{y}_i \\ &\xrightarrow{p} 0.\end{aligned}$$

This establishes that  $\bar{\mathbf{y}}^* - \bar{\mathbf{y}} \xrightarrow{p^*} \mathbf{0}$ .

Since  $\bar{\mathbf{y}} - \boldsymbol{\mu} \xrightarrow{p} 0$  by the WLLN,  $\bar{\mathbf{y}} - \boldsymbol{\mu} \xrightarrow{p^*} 0$  by Theorem 10.5. Since  $\bar{\mathbf{y}}^* - \boldsymbol{\mu} = \bar{\mathbf{y}}^* - \bar{\mathbf{y}} + \bar{\mathbf{y}} - \boldsymbol{\mu}$ , we deduce that  $\bar{\mathbf{y}}^* - \boldsymbol{\mu} \xrightarrow{p^*} \mathbf{0}$ . ■

**Proof of Theorem 10.8:** We verify conditions for the multivariate Lindeberg CLT (Theorem 6.15). (We cannot use the Lindeberg–Lévy CLT since the conditional distribution depends on  $n$ .) Conditional on  $F_n$ , the bootstrap draws  $\mathbf{y}_i^* - \bar{\mathbf{y}}$  are i.i.d. with mean  $\mathbf{0}$  and variance matrix  $\hat{\Sigma}$ . Set  $\nu_n^2 = \lambda_{\min}(\hat{\Sigma})$ . Note that by the WLLN,  $\nu_n^2 \xrightarrow{p} \nu^2 = \lambda_{\min}(\Sigma) > 0$ . Thus for  $n$  sufficiently large,  $\nu_n^2 > 0$  with high probability. Fix  $\varepsilon > 0$ . Equation (6.14) equals

$$\begin{aligned}\frac{1}{n\nu_n^2} \sum_{i=1}^n \mathbb{E}^* (\|\mathbf{y}_i^* - \bar{\mathbf{y}}\|^2 \mathbf{1}(\|\mathbf{y}_i^* - \bar{\mathbf{y}}\|^2 \geq \varepsilon n\nu_n^2)) &= \frac{1}{\nu_n^2} \mathbb{E}^* (\|\mathbf{y}_i^* - \bar{\mathbf{y}}\|^2 \mathbf{1}(\|\mathbf{y}_i^* - \bar{\mathbf{y}}\|^2 \geq \varepsilon n\nu_n^2)) \\ &\leq \frac{1}{\varepsilon n\nu_n^4} \mathbb{E}^* \|\mathbf{y}_i^* - \bar{\mathbf{y}}\|^4 \\ &\leq \frac{2^4}{\varepsilon n\nu_n^4} \mathbb{E}^* \|\mathbf{y}_i^*\|^4 \\ &= \frac{2^4}{\varepsilon n^2 \nu_n^4} \sum_{i=1}^n \|\mathbf{y}_i\|^4 \\ &\xrightarrow{p} 0.\end{aligned}$$

The second inequality uses Minkowski's inequality (B.33), Liapunov's inequality (B.34) and the  $c_r$  inequality (B.6). The following equality is  $\mathbb{E}^* \|\mathbf{y}_i^*\|^4 = n^{-1} \sum_{i=1}^n \|\mathbf{y}_i\|^4$ , which is similar to (10.10). The final convergence holds by the Marcinkiewicz WLLN (Theorem 6.39) with  $r = 2$  and  $y_i = \|\mathbf{y}_i\|^2$ . The conditions for Theorem 6.15 hold and we conclude

$$\hat{\Sigma}^{-1/2} \sqrt{n}(\bar{\mathbf{y}}^* - \bar{\mathbf{y}}) \xrightarrow{d^*} \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Since  $\hat{\Sigma} \xrightarrow{p^*} \Sigma$  we deduce that

$$\sqrt{n}(\bar{\mathbf{y}}^* - \bar{\mathbf{y}}) \xrightarrow{d^*} \mathcal{N}(\mathbf{0}, \Sigma)$$

as claimed. ■

**Proof of Theorem 10.14:** For notational simplicity assume  $\theta$  and  $\mu$  are scalar. Set  $h_i = h(y_i)$ . The assumption that the  $p^{th}$  derivative of  $g(u)$  is bounded implies  $|g^{(p)}(u)| \leq C$  for some  $C < \infty$ . Taking a  $p^{th}$  order Taylor series expansion

$$\hat{\theta}^* - \hat{\theta} = g(\bar{h}^*) - g(\bar{h}) = \sum_{j=1}^{p-1} \frac{g^{(j)}(\bar{h})}{j!} (\bar{h}^* - \bar{h})^j + \frac{g^{(p)}(\zeta_n^*)}{p!} (\bar{h}^* - \bar{h})^p$$

where  $\zeta_n^*$  lies between  $\bar{h}^*$  and  $\bar{h}$ . This implies

$$|z_n^*| = \sqrt{n} |\hat{\theta}^* - \hat{\theta}| \leq \sqrt{n} \sum_{j=1}^p c_j |\bar{h}^* - \bar{h}|^j$$

where  $c_j = |g^{(j)}(\bar{h})| / j!$  for  $j < p$  and  $c_p = C/p!$ . We find that the fourth central moment of the normalized bootstrap estimator  $z_n^* = \sqrt{n}(\hat{\theta}^* - \hat{\theta})$  satisfies the bound

$$\mathbb{E}^*(z_n^*)^4 \leq \sum_{r=4}^{4p} a_r n^2 \mathbb{E}^* |\bar{h}^* - \bar{h}|^r \quad (10.38)$$

where the coefficients  $a_r$  are products of the coefficients  $c_j$  and hence each  $O_p(1)$ . We see that  $\mathbb{E}^*(z_n^*)^4 = O_p(1)$  if  $n^2 \mathbb{E}^* |\bar{h}^* - \bar{h}|^r = O_p(1)$  for  $r = 4, \dots, 4p$ .

We show this holds for any  $r \geq 4$  using Rosenthal's inequality (B.50), which states that for each  $r$  there is a constant  $R_r < \infty$  such that

$$\begin{aligned} n^2 \mathbb{E}^* |\bar{h}^* - \bar{h}|^r &= n^{2-r} \mathbb{E}^* \left| \sum_{i=1}^n (h_i^* - \bar{h}) \right|^r \\ &\leq n^{2-r} R_r \left\{ \left( n \mathbb{E}^* (h_i^* - \bar{h})^2 \right)^{r/2} + n \mathbb{E}^* |h_i^* - \bar{h}|^r \right\} \\ &= R_r \left\{ n^{2-r/2} \hat{\sigma}^r + \frac{1}{n^{r-2}} \sum_{i=1}^n |h_i - \bar{h}|^r \right\}. \end{aligned} \quad (10.39)$$

Since  $\mathbb{E}(h_i^2) < \infty$ ,  $\hat{\sigma}^2 = O_p(1)$ , so the first term in (10.39) is  $O_p(1)$ . Also, by the Marcinkiewicz WLLN (Theorem 6.39),  $n^{-r/2} \sum_{i=1}^n |h_i - \bar{h}|^r = o_p(1)$  for any  $r \geq 1$ , so the second term in (10.39) is  $o_p(1)$  for  $r \geq 4$ . Thus for all  $r \geq 4$ , (10.39) is  $O_p(1)$  and thus (10.38) is  $O_p(1)$ . We deduce that  $z_n^*$  is uniformly square integrable, and the bootstrap estimate of variance is consistent.

This argument can be extended to vector-valued means and estimates. ■

**Proof of Theorem 10.16:** We show that  $\mathbb{E}^* \|z_n^{**}\|^4 = O_p(1)$ . By Theorem 6.31 this implies that  $z_n^{**}$  is uniformly square integrable. Since  $z_n^{**} \xrightarrow{d^*} Z$ , Theorem 6.32 implies that  $\text{var}(z_n^{**}) \rightarrow \text{var}(Z) = V_\beta$  as stated.

Set  $\mathbf{h}_i = \mathbf{h}(\mathbf{y}_i)$ . Since  $\mathbf{G}(\mathbf{u}) = \frac{\partial}{\partial \mathbf{u}} \mathbf{g}(\mathbf{u})'$  is continuous in a neighborhood of  $\mu$ , there exists  $\eta > 0$  and  $M < \infty$  such that  $\|\mathbf{u} - \mu\| \leq 2\eta$  implies  $\text{tr}(\mathbf{G}(\mathbf{u})' \mathbf{G}(\mathbf{u})) \leq M$ . By the WLLN and bootstrap WLLN there is an  $n$  sufficiently large such that  $\|\bar{\mathbf{h}}_n - \mu\| \leq \eta$  and  $\|\bar{\mathbf{h}}_n^* - \bar{\mathbf{h}}_n\| \leq \eta$  with probability exceeding  $1 - \eta$ . On this event,  $\|\mathbf{u} - \bar{\mathbf{h}}_n\| \leq \eta$  implies  $\text{tr}(\mathbf{G}(\mathbf{u})' \mathbf{G}(\mathbf{u})) \leq M$ . Using the mean-value theorem at a point  $\zeta_n^*$  intermediate between  $\bar{\mathbf{h}}_n^*$  and  $\bar{\mathbf{h}}_n$

$$\begin{aligned} \|z_n^{**}\|^4 \mathbf{1}(\|\bar{\mathbf{h}}_n^* - \bar{\mathbf{h}}_n\| \leq \eta) &\leq n^2 \|\mathbf{g}(\bar{\mathbf{h}}_n^*) - \mathbf{g}(\bar{\mathbf{h}}_n)\|^4 \mathbf{1}(\|\bar{\mathbf{h}}_n^* - \bar{\mathbf{h}}_n\| \leq \eta) \\ &\leq n^2 \|\mathbf{G}(\zeta_n^*)' (\bar{\mathbf{h}}_n^* - \bar{\mathbf{h}}_n)\|^4 \\ &\leq M^2 n^2 \|\bar{\mathbf{h}}_n^* - \bar{\mathbf{h}}_n\|^4. \end{aligned}$$

Then

$$\begin{aligned} \mathbb{E}^* \|z_n^{**}\|^4 &\leq \mathbb{E}^* \left( \|z_n^{**}\|^4 \mathbf{1}(\|\bar{\mathbf{h}}_n^* - \bar{\mathbf{h}}_n\| \leq \eta) \right) + \tau_n^4 \mathbb{E}^* \left( \mathbf{1}(\|\bar{\mathbf{h}}_n^* - \bar{\mathbf{h}}_n\| > \eta) \right) \\ &\leq M^2 n^2 \mathbb{E}^* \|\bar{\mathbf{h}}_n^* - \bar{\mathbf{h}}_n\|^4 + \tau_n^4 \mathbb{P}^* \left( \|\bar{\mathbf{h}}_n^* - \bar{\mathbf{h}}_n\| > \eta \right). \end{aligned} \quad (10.40)$$

In (10.17) we showed that the first term in (10.40) is  $O_p(1)$  in the scalar case. The vector case follows by element-by-element expansion.

Now take the second term in (10.40). We apply Bernstein's inequality for vectors (B.39). Note that  $\bar{\mathbf{h}}_n^* - \bar{\mathbf{h}}_n = n^{-1} \sum_{i=1}^n \mathbf{u}_i^*$  with  $\mathbf{u}_i^* = \mathbf{h}_i^* - \bar{\mathbf{h}}_n$  with  $j^{th}$  element  $u_{ji}^* = h_{ji}^* - \bar{h}_{jn}$ . The  $\mathbf{u}_i^*$  are i.i.d., mean zero,  $\mathbb{E}^*(u_{ji}^{*2}) = \hat{\sigma}_j^2 = O_p(1)$ , and satisfy the bound  $|u_{ji}^*| \leq 2 \max_{i,j} |h_{ji}| = B_n$ , say. Bernstein's inequality states that

$$\mathbb{P}^*(\|\bar{\mathbf{h}}_n^* - \bar{\mathbf{h}}_n\| > \eta) \leq 2m \exp\left(-n^{1/2} \frac{\eta^2}{2m^2 n^{-1/2} \max_j \hat{\sigma}_j^2 + 2mn^{-1/2} B_n \eta / 3}\right). \quad (10.41)$$

Theorem (6.31) shows that  $n^{-1/2} B_n = o_p(1)$ . Thus the expression in the denominator of the parentheses in (10.41) is  $o_p(1)$  as  $n \rightarrow \infty$ . It follows that for  $n$  sufficiently large (10.41) is  $O_p(\exp(-n^{1/2}))$ . Hence the second term in (10.40) is  $O_p(\exp(-n^{1/2})) o_p(\exp(-n^{1/2})) = o_p(1)$  by the assumption on  $\tau_n$ .

We have shown that the two terms in (10.40) are each  $O_p(1)$ . This completes the proof.  $\blacksquare$

## Exercises

**Exercise 10.1** Find the jackknife estimator of variance of the estimator  $\hat{\mu}_r = n^{-1} \sum_{i=1}^n y_i^r$  for  $\mu_r = \mathbb{E}(y_i^r)$ .

**Exercise 10.2** Show that if the jackknife estimator of variance of  $\hat{\beta}$  is  $\hat{V}_{\hat{\beta}}^{\text{jack}}$ , then the jackknife estimator of variance of  $\hat{\theta} = \mathbf{a} + \mathbf{C}\hat{\beta}$  is  $\hat{V}_{\hat{\theta}}^{\text{jack}} = \mathbf{C}\hat{V}_{\hat{\beta}}^{\text{jack}}\mathbf{C}'$ .

**Exercise 10.3** A two-step estimator such as (12.51) is  $\hat{\beta} = (\sum_{i=1}^n \hat{\mathbf{w}}_i \hat{\mathbf{w}}_i')^{-1} (\sum_{i=1}^n \hat{\mathbf{w}}_i y_i)$  where  $\hat{\mathbf{w}}_i = \hat{\mathbf{A}}' \mathbf{z}_i$  and  $\hat{\mathbf{A}} = (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X}$ . Describe how to construct the jackknife estimator of variance of  $\hat{\beta}$ .

**Exercise 10.4** Let  $\hat{F}(\mathbf{x})$  denote the EDF of a random sample. Show that

$$\sqrt{n} (\hat{F}(\mathbf{x}) - F(\mathbf{x})) \xrightarrow{d} N(0, F(\mathbf{x})(1 - F(\mathbf{x}))).$$

**Exercise 10.5** Show that if the bootstrap estimator of variance of  $\hat{\beta}$  is  $\hat{V}_{\hat{\beta}}^{\text{boot}}$ , then the bootstrap estimator of variance of  $\hat{\theta} = \mathbf{a} + \mathbf{C}\hat{\beta}$  is  $\hat{V}_{\hat{\theta}}^{\text{boot}} = \mathbf{C}\hat{V}_{\hat{\beta}}^{\text{boot}}\mathbf{C}'$ .

**Exercise 10.6** Show that if the percentile interval for  $\beta$  is  $[L, U]$  then the percentile interval for  $a + c\beta$  is  $[a + cL, a + cU]$ .

**Exercise 10.7** Consider the following bootstrap procedure. Using the non-parametric bootstrap, generate bootstrap samples, calculate the estimate  $\hat{\theta}^*$  on these samples and then calculate

$$T^* = (\hat{\theta}^* - \hat{\theta}) / s(\hat{\theta}),$$

where  $s(\hat{\theta})$  is the standard error in the original data. Let  $q_{\alpha/2}^*$  and  $q_{1-\alpha/2}^*$  denote the  $\alpha/2^{\text{th}}$  and  $1 - \alpha/2^{\text{th}}$  quantiles of  $T^*$ , and define the bootstrap confidence interval

$$C = [\hat{\theta} + s(\hat{\theta}) q_{\alpha/2}^*, \quad \hat{\theta} + s(\hat{\theta}) q_{1-\alpha/2}^*].$$

Show that  $C$  exactly equals the percentile interval.

**Exercise 10.8** Prove Theorem 10.10.

**Exercise 10.9** Prove Theorem 10.11.

**Exercise 10.10** Prove Theorem 10.12.

**Exercise 10.11** Let  $y_i$  be i.i.d.,  $\mu = \mathbb{E}(y_i) > 0$ , and  $\theta = \mu^{-1}$ . Let  $\hat{\mu} = \bar{Y}_n$  be the sample mean and  $\hat{\theta} = \hat{\mu}^{-1}$ .

- (a) Is  $\hat{\theta}$  unbiased for  $\theta$ ?
- (b) If  $\hat{\theta}$  is biased, can you determine the direction of the bias  $\mathbb{E}(\hat{\theta} - \theta)$  (up or down)?
- (c) Is the percentile interval appropriate in this context for confidence interval construction?

**Exercise 10.12** Consider the following bootstrap procedure for a regression of  $y_i$  on  $\mathbf{x}_i$ . Let  $\hat{\beta}$  denote the OLS estimator from the regression of  $\mathbf{y}$  on  $\mathbf{X}$ , and  $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\beta}$  the OLS residuals.

- (a) Draw a random vector  $(\mathbf{x}^*, e^*)$  from the pair  $\{(\mathbf{x}_i, \hat{e}_i) : i = 1, \dots, n\}$ . That is, draw a random integer  $i'$  from  $[1, 2, \dots, n]$ , and set  $\mathbf{x}^* = \mathbf{x}_{i'}$  and  $e^* = \hat{e}_{i'}$ . Set  $y^* = \mathbf{x}^{*'} \hat{\beta} + e^*$ . Draw (with replacement)  $n$  such vectors, creating a random bootstrap data set  $(\mathbf{y}^*, \mathbf{X}^*)$ .

- (b) Regress  $\mathbf{y}^*$  on  $\mathbf{X}^*$ , yielding OLS estimates  $\hat{\boldsymbol{\beta}}^*$  and any other statistic of interest.

Show that this bootstrap procedure is (numerically) identical to the non-parametric bootstrap.

**Exercise 10.13** Take  $p^*$  as defined in (10.22) for the BC percentile interval. Show that it is invariant to replacing  $\theta$  with  $g(\theta)$  for any strictly monotonically increasing transformation  $g(\theta)$ . Does this extend to  $z_0^*$  as defined in (10.23)?

**Exercise 10.14** Show that if the percentile-t interval for  $\beta$  is  $[L, U]$  then the percentile-t interval for  $a + c\beta$  is  $[a + bL, a + bU]$ .

**Exercise 10.15** You want to test  $H_0 : \theta = 0$  against  $H_1 : \theta > 0$ . The test for  $H_0$  is to reject if  $T_n = \hat{\theta}/s(\hat{\theta}) > c$  where  $c$  is picked so that Type I error is  $\alpha$ . You do this as follows. Using the nonparametric bootstrap, you generate bootstrap samples, calculate the estimates  $\hat{\theta}^*$  on these samples and then calculate

$$T^* = \hat{\theta}^*/s(\hat{\theta}^*).$$

Let  $q_{1-\alpha}^*$  denote the  $1 - \alpha^{th}$  quantile of  $T^*$ . You replace  $c$  with  $q_{1-\alpha}^*$ , and thus reject  $H_0$  if  $T_n = \hat{\theta}/s(\hat{\theta}) > q_{1-\alpha}^*$ . What is wrong with this procedure?

**Exercise 10.16** Suppose that in an application,  $\hat{\theta} = 1.2$  and  $s(\hat{\theta}) = .2$ . Using the nonparametric bootstrap, 1000 samples are generated from the bootstrap distribution, and  $\hat{\theta}^*$  is calculated on each sample. The  $\hat{\theta}^*$  are sorted, and the  $0.025^{th}$  and  $0.975^{th}$  quantiles of the  $\hat{\theta}^*$  are .75 and 1.3, respectively.

- (a) Report the 95% percentile interval for  $\theta$ .
- (c) With the given information, can you calculate the 95% BC percentile interval or percentile-t interval for  $\theta$ ?

**Exercise 10.17** Take the normal regression model

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ e_i | \mathbf{x}_i &\sim N(0, \sigma^2) \end{aligned}$$

where we know the MLE are the least-squares estimators  $\hat{\boldsymbol{\beta}}$  and  $\hat{\sigma}^2$ .

- (a) Describe the parametric regression bootstrap for this model. Show that the conditional distribution of the bootstrap observations is  $y_i^* | F_n \sim N(\mathbf{x}'_i \hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ .
- (b) Show that the distribution of the bootstrap least-squares estimator is  $\hat{\boldsymbol{\beta}}^* | F_n \sim N(\hat{\boldsymbol{\beta}}, (\mathbf{X}' \mathbf{X})^{-1} \hat{\sigma}^2)$ .
- (c) (optional) Show that the distribution of the bootstrap t-ratio with a homoskedastic standard error is  $T^* \sim t_{n-k}$ .

**Exercise 10.18** Consider the model

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(e_i | \mathbf{x}_i) &= 0 \end{aligned}$$

with  $y_i$  scalar and  $\mathbf{x}_i$  a  $k$  vector. You have a random sample  $(y_i, \mathbf{x}_i : i = 1, \dots, n)$ . You are interested in estimating the regression function  $m(\mathbf{x}) = E(y_i | \mathbf{x}_i = \mathbf{x})$  at a fixed vector  $x$  and constructing a 95% confidence interval.

- (a) Write down the standard estimator and asymptotic confidence interval for  $m(\mathbf{x})$ .
- (b) Describe the percentile bootstrap confidence interval for  $m(\mathbf{x})$ .

- (c) Describe the percentile-t bootstrap confidence interval for  $m(\mathbf{x})$ .

**Exercise 10.19** The observed data is  $\{y_i, x_i\} \in \mathbb{R} \times \mathbb{R}^k$ ,  $k > 1$ ,  $i = 1, \dots, n$ . Take the model

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(x_i e_i) &= 0 \\ \mu_3 &= \mathbb{E}(e_i^3) \end{aligned}$$

- (a) Write down an estimator for  $\mu_3$ .
- (b) Explain how to use the percentile method to construct a 90% confidence interval for  $\mu_3$  in this specific model.

**Exercise 10.20** Take the model

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(x_i e_i) &= 0 \\ \mathbb{E}(e_i^2) &= \sigma^2 \end{aligned}$$

Describe the bootstrap percentile confidence interval for  $\sigma^2$ .

**Exercise 10.21** The model is

$$\begin{aligned} y_i &= \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) &= 0 \end{aligned}$$

with  $x_{2i}$  scalar. Describe how to test  $H_0 : \beta_2 = 0$  against  $H_1 : \beta_2 \neq 0$  using the nonparametric bootstrap.

**Exercise 10.22** The model is

$$\begin{aligned} y_i &= \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + x_{2i} \beta_2 + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) &= 0 \end{aligned}$$

with both  $\mathbf{x}_{1i}$  and  $\mathbf{x}_{1i}$   $k \times 1$ . Describe how to test  $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$  against  $H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$  using the nonparametric bootstrap.

**Exercise 10.23** Suppose a PhD student has a sample  $(y_i, x_i, z_i : i = 1, \dots, n)$  and estimates by OLS the equation

$$y_i = z_i \hat{\alpha} + x'_i \hat{\beta} + \hat{e}_i$$

where  $\alpha$  is the coefficient of interest and she is interested in testing  $H_0 : \alpha = 0$  against  $H_1 : \alpha \neq 0$ . She obtains  $\hat{\alpha} = 2.0$  with standard error  $s(\hat{\alpha}) = 1.0$  so the value of the t-ratio for  $H_0$  is  $T = \hat{\alpha}/s(\hat{\alpha}) = 2.0$ . To assess significance, the student decides to use the bootstrap. She uses the following algorithm

1. Samples  $(y_i^*, x_i^*, z_i^*)$  randomly from the observations. (Random sampling with replacement). Creates a random sample with  $n$  observations.
2. On this pseudo-sample, estimates the equation

$$y_i^* = z_i^* \hat{\alpha}^* + x_i^{*\prime} \hat{\beta}^* + \hat{e}_i^*$$

by OLS and computes standard errors, including  $s(\hat{\alpha}^*)$ . The t-ratio for  $H_0$ ,  $T^* = \hat{\alpha}^*/s(\hat{\alpha}^*)$  is computed and stored.

3. This is repeated  $B = 10,000$  times.

4. The 0.95<sup>th</sup> empirical quantile  $q_{.95}^*$  of the bootstrap absolute t-ratios  $|T^*|$  is computed. It is  $q_{.95}^* = 3.5$ .
5. The student notes that while  $|T| = 2 > 1.96$  (and thus an asymptotic 5% size test rejects  $H_0$ ),  $|T| = 2 < q_{.95}^* = 3.5$  and thus the bootstrap test does not reject  $H_0$ . As the bootstrap is more reliable, the student concludes that  $H_0$  cannot be rejected in favor of  $H_1$ .

Question: Do you agree with the student's method and reasoning? Do you see an error in her method?

**Exercise 10.24** Take the model

$$\begin{aligned}y_i &= x_{1i}\beta_1 + x_{2i}\beta_2 + e_i \\ \mathbb{E}(x_i e_i) &= 0\end{aligned}$$

The parameter of interest is  $\theta = \beta_1\beta_2$ . Show how to construct a confidence interval for  $\theta$  using the following three methods.

- (a) Asymptotic Theory.
- (b) Percentile Bootstrap.
- (c) Percentile-t Bootstrap.

Your answer should be specific to this problem, not general.

**Exercise 10.25** Take the model

$$\begin{aligned}y_i &= x_{1i}\beta_1 + x_{2i}\beta_2 + e_i \\ \mathbb{E}(x_i e_i) &= 0 \\ \theta &= \frac{\beta_1}{\beta_2}.\end{aligned}$$

Assume that the observations  $(y_i, x_{1i}, x_{2i})$  are i.i.d. across  $i = 1, \dots, n$ . Describe how you would construct the percentile-t bootstrap confidence interval for  $\theta$ .

**Exercise 10.26** The model is i.i.d. data,  $i = 1, \dots, n$ ,

$$\begin{aligned}y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(e_i | \mathbf{x}_i) &= 0.\end{aligned}$$

Does the presence of conditional heteroskedasticity invalidate the application of the nonparametric bootstrap? Explain.

**Exercise 10.27** The RESET specification test for nonlinearity in a random sample (due to Ramsey (1969)) is the following.

The null hypothesis is a linear regression

$$\begin{aligned}y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(e_i | \mathbf{x}_i) &= 0.\end{aligned}$$

The parameter  $\boldsymbol{\beta}$  is estimated by OLS yielding predicted values  $\hat{y}_i$ . Then a second-stage least-squares regression is estimated including both  $\mathbf{x}_i$  and  $\hat{y}_i$

$$y_i = \mathbf{x}'_i \tilde{\boldsymbol{\beta}} + (\hat{y}_i)^2 \tilde{\gamma} + \tilde{e}_i$$

The RESET test statistic  $R$  is the squared t-ratio on  $\tilde{\gamma}$ .

A colleague suggests obtaining the critical value for the test using the bootstrap. He proposes the following bootstrap implementation.

- Draw  $n$  observations  $(y_i^*, \mathbf{x}_i^*)$  randomly from the observed sample pairs  $(y_i, \mathbf{x}_i)$  to create a bootstrap sample.
- Compute the statistic  $R^*$  on this bootstrap sample as described above.
- Repeat this  $B$  times. Sort the bootstrap statistics  $R^*$ , take the  $0.95^{th}$  quantile and use this as the critical value.
- Reject the null hypothesis if  $R$  exceeds this critical value, otherwise do not reject.

Is this procedure a correct implementation of the bootstrap in this context? If not, propose a modified bootstrap.

**Exercise 10.28** The model is

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) \neq 0,$$

so the regressor  $\mathbf{x}_i$  is endogenous. We know that in this case, the least-squares estimator is biased for the parameter  $\boldsymbol{\beta}$ . We also know that the nonparametric BC percentile interval is (generally) a good method for confidence interval construction in the presence of bias. Explain whether or not you expect the BC percentile interval applied to the least-squares estimator will have accurate coverage in the presence of endogeneity.

**Exercise 10.29** In Exercise 9.26 you estimated a cost function for 145 electric companies and tested the restriction  $\theta = \beta_3 + \beta_4 + \beta_5 = 1$ .

- Estimate the regression by unrestricted least-squares, and report standard errors calculated by asymptotic, jackknife and the bootstrap.
- Estimate  $\theta = \beta_3 + \beta_4 + \beta_5$ , and report standard errors calculated by asymptotic, jackknife and the bootstrap.
- Report confidence intervals for  $\theta$  using the percentile and  $BC_a$  methods

**Exercise 10.30** In Exercise 9.27 you estimated the Mankiw, Romer, and Weil (1992) unrestricted regression. Let  $\theta$  be the sum of the second, third and fourth coefficients.

- Estimate the regression by unrestricted least-squares, and report standard errors calculated by asymptotic, jackknife and the bootstrap.
- Estimate  $\theta$  and report standard errors calculated by asymptotic, jackknife and the bootstrap.
- Report confidence intervals for  $\theta$  using the percentile and BC methods.

**Exercise 10.31** In Exercise 7.29 you estimated a wage regression with the CPS dataset and the subsample of white Male Hispanics. Further restrict the sample to those never-married and live in the Midwest region. (This sample has 99 observations.) As in subquestion (b), let  $\theta$  be the ratio of the return to one year of education to the return of one year of experience.

- Estimate  $\theta$  and report standard errors calculated by asymptotic, jackknife and the bootstrap.
- Explain the discrepancy between the standard errors.
- Report confidence intervals for  $\theta$  using the BC percentile method.

**Exercise 10.32** In Exercise 4.26 you extended the work from Duflo, Dupas and Kremer (2011). Repeat that regression, now calculating the standard error as well by cluster bootstrap. Report a  $BC_a$  confidence interval for each coefficient.

## **Part III**

# **Multiple Equation Models**

# Chapter 11

## Multivariate Regression

### 11.1 Introduction

**Multivariate regression** is a system of regression equations. Multivariate regression is used as reduced form models for instrumental variable estimation (explored in Chapter 12), vector autoregressions (explored in Chapter 15), demand systems (demand for multiple goods), and other contexts.

Multivariate regression is also called by the name **systems of regression equations**. Closely related is the method of **Seemingly Unrelated Regressions** (SUR) which we introduce in Section 11.7.

Most of the tools of single equation regression generalize naturally to multivariate regression. A major difference is a new set of notation to handle matrix estimates.

### 11.2 Regression Systems

A system of linear regressions takes the form

$$y_{ji} = \mathbf{x}'_{ji} \boldsymbol{\beta}_j + e_{ji} \quad (11.1)$$

for variables  $j = 1, \dots, m$  and observations  $i = 1, \dots, n$ , where the regressor vectors  $\mathbf{x}_{ji}$  are  $k_j \times 1$  and  $e_{ji}$  is an error. The coefficient vectors  $\boldsymbol{\beta}_j$  are  $k_j \times 1$ . The total number of coefficients are  $\bar{k} = \sum_{j=1}^n k_j$ . The regression system specializes to univariate regression when  $m = 1$ .

It is typical to treat the observations as independent across observations  $i$  but correlated across variables  $j$ . As an example, the observations  $y_{ji}$  could be expenditures by household  $i$  on good  $j$ . The standard assumptions are that households are mutually independent, but expenditures by an individual household are correlated across goods.

To describe the dependence between the dependent variables, we can define the  $m \times 1$  error vector  $\mathbf{e}_i = (e_{1i}, \dots, e_{mi})'$  and its  $m \times m$  variance matrix

$$\boldsymbol{\Sigma} = \mathbb{E}(\mathbf{e}_i \mathbf{e}_i').$$

The diagonal elements are the variances of the errors  $e_{ji}$ , and the off-diagonals are the covariances across variables. It is typical to allow  $\boldsymbol{\Sigma}$  to be unconstrained.

We can group the  $m$  equations (11.1) into a single equation as follows. Let  $\mathbf{y}_i = (y_{1i}, \dots, y_{mi})'$  be the  $m \times 1$  vector of dependent variables, define the  $m \times \bar{k}$  matrix of regressors

$$\overline{\mathbf{X}}_i = \begin{pmatrix} \mathbf{x}'_{1i} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \mathbf{x}'_{2i} & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}'_{mi} \end{pmatrix},$$

and define the  $\bar{k} \times 1$  stacked coefficient vector

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_m \end{pmatrix}.$$

Then the  $m$  regression equations can jointly be written as

$$\mathbf{y}_i = \bar{\mathbf{X}}_i \boldsymbol{\beta} + \mathbf{e}_i. \quad (11.2)$$

The entire system can be written in matrix notation by stacking the variables. Define

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_n \end{pmatrix}, \quad \bar{\mathbf{X}} = \begin{pmatrix} \bar{\mathbf{X}}_1 \\ \vdots \\ \bar{\mathbf{X}}_n \end{pmatrix}$$

which are  $mn \times 1$ ,  $mn \times 1$ , and  $mn \times \bar{k}$ , respectively. The system can be written as

$$\mathbf{y} = \bar{\mathbf{X}} \boldsymbol{\beta} + \mathbf{e}.$$

In many applications the regressor vectors  $\mathbf{x}_{ji}$  are common across the variables  $j$ , so  $\mathbf{x}_{ji} = \mathbf{x}_i$  and  $k_j = k$ . By this we mean that the same variables enter each equation with no exclusion restrictions. Several important simplifications occur in this context. One is that we can write (11.2) using the notation

$$\mathbf{y}_i = \mathbf{B}' \mathbf{x}_i + \mathbf{e}_i \quad (11.3)$$

where  $\mathbf{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_m)$  is  $k \times m$ . Another is that we can write the system in the  $n \times m$  matrix notation

$$\mathbf{Y} = \mathbf{X} \mathbf{B} + \mathbf{E}$$

where

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}'_1 \\ \vdots \\ \mathbf{y}'_n \end{pmatrix}, \quad \mathbf{E} = \begin{pmatrix} \mathbf{e}'_1 \\ \vdots \\ \mathbf{e}'_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}'_1 \\ \vdots \\ \mathbf{x}'_n \end{pmatrix}.$$

Another convenient implication of common regressors is that we have the simplification

$$\bar{\mathbf{X}}_i = \begin{pmatrix} \mathbf{x}'_i & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}'_i & & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}'_i \end{pmatrix} = \mathbf{I}_m \otimes \mathbf{x}'_i$$

where  $\otimes$  is the Kronecker product (see Appendix A.21).

### 11.3 Least-Squares Estimator

Consider estimating each equation (11.1) by least-squares. This takes the form

$$\hat{\boldsymbol{\beta}}_j = \left( \sum_{i=1}^n \mathbf{x}_{ji} \mathbf{x}'_{ji} \right)^{-1} \left( \sum_{i=1}^n \mathbf{x}_{ji} y_{ji} \right).$$

The combined estimate of  $\boldsymbol{\beta}$  is the stacked vector

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \vdots \\ \hat{\boldsymbol{\beta}}_m \end{pmatrix}.$$

It turns that we can write this estimator using the systems notation

$$\hat{\beta} = (\bar{X}' \bar{X})^{-1} (\bar{X}' \mathbf{y}) = \left( \sum_{i=1}^n \bar{X}_i' \bar{X}_i \right)^{-1} \left( \sum_{i=1}^n \bar{X}_i' \mathbf{y}_i \right). \quad (11.4)$$

To see this, observe that

$$\begin{aligned} \bar{X}' \bar{X} &= \begin{pmatrix} \bar{X}_1' & \cdots & \bar{X}_n' \end{pmatrix} \begin{pmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_n \end{pmatrix} \\ &= \sum_{i=1}^n \bar{X}_i' \bar{X}_i \\ &= \sum_{i=1}^n \begin{pmatrix} \mathbf{x}_{1i} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \mathbf{x}_{2i} & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}_{mi} \end{pmatrix} \begin{pmatrix} \mathbf{x}'_{1i} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \mathbf{x}'_{2i} & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}'_{mi} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^n \mathbf{x}_{1i} \mathbf{x}'_{1i} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \sum_{i=1}^n \mathbf{x}_{2i} \mathbf{x}'_{2i} & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \sum_{i=1}^n \mathbf{x}_{mi} \mathbf{x}'_{mi} \end{pmatrix} \end{aligned}$$

and

$$\begin{aligned} \bar{X}' \mathbf{y} &= \begin{pmatrix} \bar{X}_1' & \cdots & \bar{X}_n' \end{pmatrix} \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{pmatrix} \\ &= \sum_{i=1}^n \bar{X}_i' \mathbf{y}_i \\ &= \sum_{i=1}^n \begin{pmatrix} \mathbf{x}_{1i} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \mathbf{x}_{2i} & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}_{mi} \end{pmatrix} \begin{pmatrix} y_{1i} \\ \vdots \\ y_{mi} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^n \mathbf{x}_{1i} y_{1i} \\ \vdots \\ \sum_{i=1}^n \mathbf{x}_{mi} y_{mi} \end{pmatrix}. \end{aligned}$$

Hence

$$\begin{aligned} (\bar{X}' \bar{X})^{-1} (\bar{X}' \mathbf{y}) &= \left( \sum_{i=1}^n \bar{X}_i' \bar{X}_i \right)^{-1} \left( \sum_{i=1}^n \bar{X}_i' \mathbf{y}_i \right) \\ &= \begin{pmatrix} (\sum_{i=1}^n \mathbf{x}_{1i} \mathbf{x}'_{1i})^{-1} (\sum_{i=1}^n \mathbf{x}_{1i} y_{1i}) \\ \vdots \\ (\sum_{i=1}^n \mathbf{x}_{mi} \mathbf{x}'_{mi})^{-1} (\sum_{i=1}^n \mathbf{x}_{mi} y_{mi}) \end{pmatrix} \\ &= \hat{\beta} \end{aligned}$$

as claimed.

The  $m \times 1$  residual vector for the  $i^{th}$  observation is

$$\hat{\mathbf{e}}_i = \mathbf{y}_i - \bar{X}_i' \hat{\beta}$$

and the least-squares estimator of the  $m \times m$  error variance matrix is

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \hat{\mathbf{e}}_i \hat{\mathbf{e}}_i'. \quad (11.5)$$

In the case of common regressors, observe that

$$\hat{\boldsymbol{\beta}}_j = \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{x}_i y_{ji} \right)$$

and

$$\hat{\mathbf{B}} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \dots, \hat{\boldsymbol{\beta}}_m) = (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{Y}). \quad (11.6)$$

In Stata, multivariate regression can be implemented using the `mvreg` command.

## 11.4 Mean and Variance of Systems Least-Squares

We can calculate the finite-sample mean and variance of  $\hat{\boldsymbol{\beta}}$  under the conditional mean assumption

$$\mathbb{E}(\mathbf{e}_i | \mathbf{x}_i) = \mathbf{0} \quad (11.7)$$

where  $\mathbf{x}_i$  is the union of the regressors  $\mathbf{x}_{ji}$ . Equation (11.7) is equivalent to  $\mathbb{E}(y_{ji} | \mathbf{x}_i) = \mathbf{x}'_{ji} \boldsymbol{\beta}_j$ , or that the regression model is correctly specified.

We can center the estimator as

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\bar{\mathbf{X}}' \bar{\mathbf{X}})^{-1} (\bar{\mathbf{X}}' \mathbf{e}) = \left( \sum_{i=1}^n \bar{\mathbf{X}}'_i \bar{\mathbf{X}}_i \right)^{-1} \left( \sum_{i=1}^n \bar{\mathbf{X}}'_i \mathbf{e}_i \right).$$

Taking conditional expectations, we find  $\mathbb{E}(\hat{\boldsymbol{\beta}} | \mathbf{X}) = \boldsymbol{\beta}$ . Consequently, systems least-squares is unbiased under correct specification.

To compute the variance of the estimator, define the conditional covariance matrix of the errors of the  $i^{th}$  observation

$$\mathbb{E}(\mathbf{e}_i \mathbf{e}'_i | \mathbf{x}_i) = \boldsymbol{\Sigma}_i$$

which in general is unrestricted. Observe that if the observations are mutually independent, then

$$\begin{aligned} \mathbb{E}(\mathbf{e} \mathbf{e}' | \mathbf{X}) &= \mathbb{E}\left(\begin{pmatrix} \mathbf{e}_1 \mathbf{e}_1 & \mathbf{e}_1 \mathbf{e}_2 & \cdots & \mathbf{e}_1 \mathbf{e}_n \\ \vdots & \ddots & & \vdots \\ \mathbf{e}_n \mathbf{e}_1 & \mathbf{e}_n \mathbf{e}_2 & \cdots & \mathbf{e}_n \mathbf{e}_n \end{pmatrix} | \mathbf{X} \right) \\ &= \begin{pmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_n \end{pmatrix}. \end{aligned}$$

Also, by independence across observations,

$$\text{var}\left(\sum_{i=1}^n \bar{\mathbf{X}}'_i \mathbf{e}_i | \mathbf{X}\right) = \sum_{i=1}^n \text{var}(\bar{\mathbf{X}}'_i \mathbf{e}_i | \mathbf{x}_i) = \sum_{i=1}^n \bar{\mathbf{X}}'_i \boldsymbol{\Sigma}_i \bar{\mathbf{X}}_i.$$

It follows that

$$\text{var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) = (\bar{\mathbf{X}}' \bar{\mathbf{X}})^{-1} \left( \sum_{i=1}^n \bar{\mathbf{X}}'_i \boldsymbol{\Sigma}_i \bar{\mathbf{X}}_i \right) (\bar{\mathbf{X}}' \bar{\mathbf{X}})^{-1}.$$

When the regressors are common so that  $\bar{\mathbf{X}}_i = \mathbf{I}_m \otimes \mathbf{x}'_i$  then the covariance matrix can be written as

$$\text{var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) = (\mathbf{I}_m \otimes (\mathbf{X}' \mathbf{X})^{-1}) \left( \sum_{i=1}^n (\boldsymbol{\Sigma}_i \otimes \mathbf{x}_i \mathbf{x}'_i) \right) (\mathbf{I}_m \otimes (\mathbf{X}' \mathbf{X})^{-1}).$$

Alternatively, if the errors are conditionally homoskedastic

$$\mathbb{E}(\mathbf{e}_i \mathbf{e}'_i | \mathbf{x}_i) = \boldsymbol{\Sigma} \quad (11.8)$$

then the covariance matrix takes the form

$$\text{var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) = (\bar{\mathbf{X}}' \bar{\mathbf{X}})^{-1} \left( \sum_{i=1}^n \bar{\mathbf{X}}_i' \boldsymbol{\Sigma} \bar{\mathbf{X}}_i \right) (\bar{\mathbf{X}}' \bar{\mathbf{X}})^{-1}.$$

If both simplifications (common regressors and conditional homoskedasticity) hold then we have the considerable simplification

$$\text{var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) = \boldsymbol{\Sigma} \otimes (\mathbf{X}' \mathbf{X})^{-1}.$$

## 11.5 Asymptotic Distribution

For an asymptotic distribution it is sufficient to consider the equation-by-equation projection model in which case

$$\mathbb{E}(\mathbf{x}_{ji} e_{ji}) = \mathbf{0}. \quad (11.9)$$

First, consider consistency. Since  $\hat{\boldsymbol{\beta}}_j$  are the standard least-squares estimators, they are consistent for the projection coefficients  $\boldsymbol{\beta}_j$ .

Second, consider the asymptotic distribution. Again by our single equation theory it is immediate that the  $\hat{\boldsymbol{\beta}}_j$  are asymptotically normally distributed. But our previous theory does not provide a joint distribution of the  $\hat{\boldsymbol{\beta}}_j$  across  $j$ . For this we need a joint theory for the stacked estimates  $\hat{\boldsymbol{\beta}}$ , which we now provide.

Since the vector

$$\bar{\mathbf{X}}_i' \mathbf{e}_i = \begin{pmatrix} \mathbf{x}_{1i} e_{1i} \\ \vdots \\ \mathbf{x}_{mi} e_{mi} \end{pmatrix}$$

is i.i.d. across  $i$  and mean zero under (11.9), the central limit theorem implies

$$\left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \bar{\mathbf{X}}_i' \mathbf{e}_i \right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$$

where

$$\boldsymbol{\Omega} = \mathbb{E}(\bar{\mathbf{X}}_i' \mathbf{e}_i \mathbf{e}_i' \bar{\mathbf{X}}_i) = \mathbb{E}(\bar{\mathbf{X}}_i' \boldsymbol{\Sigma} \bar{\mathbf{X}}_i).$$

The matrix  $\boldsymbol{\Omega}$  is the covariance matrix of the variables  $\mathbf{x}_{ji} e_{ji}$  across equations. Under conditional homoskedasticity (11.8) the matrix  $\boldsymbol{\Omega}$  simplifies to

$$\boldsymbol{\Omega} = \mathbb{E}(\bar{\mathbf{X}}_i' \boldsymbol{\Sigma} \bar{\mathbf{X}}_i) \quad (11.10)$$

(see Exercise 11.1). When the regressors are common then it simplifies to

$$\boldsymbol{\Omega} = \mathbb{E}(\mathbf{e}_i \mathbf{e}_i' \otimes \mathbf{x}_i \mathbf{x}_i') \quad (11.11)$$

(see Exercise 11.2) and under both conditions (homoskedasticity and common regressors) it simplifies to

$$\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \mathbb{E}(\mathbf{x}_i \mathbf{x}_i') \quad (11.12)$$

(see Exercise 11.3).

Applied to the centered and normalized estimator we obtain the asymptotic distribution.

**Theorem 11.1** Under Assumption 7.2,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, V_{\boldsymbol{\beta}})$$

where

$$V_{\boldsymbol{\beta}} = \mathbf{Q}^{-1} \boldsymbol{\Omega} \mathbf{Q}^{-1}$$

$$\mathbf{Q} = \mathbb{E}(\bar{\mathbf{X}}_i' \bar{\mathbf{X}}_i) = \begin{pmatrix} \mathbb{E}(\mathbf{x}_{1i} \mathbf{x}'_{1i}) & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbb{E}(\mathbf{x}_{ni} \mathbf{x}'_{ni}) \end{pmatrix}.$$

For a proof, see Exercise 11.4.

When the regressors are common then the matrix  $\mathbf{Q}$  simplifies as

$$\mathbf{Q} = \mathbf{I}_m \otimes \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i) \quad (11.13)$$

(See Exercise 11.5).

If both the regressors are common and the errors are conditionally homoskedastic (11.8) then we have the simplification

$$V_{\boldsymbol{\beta}} = \boldsymbol{\Sigma} \otimes (\mathbb{E}(\mathbf{x}_i \mathbf{x}'_i))^{-1} \quad (11.14)$$

(see Exercise 11.6).

Sometimes we are interested in parameters  $\boldsymbol{\theta} = r(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m) = r(\boldsymbol{\beta})$  which are functions of the coefficients from multiple equations. In this case the least-squares estimate of  $\boldsymbol{\theta}$  is  $\hat{\boldsymbol{\theta}} = r(\hat{\boldsymbol{\beta}})$ . The asymptotic distribution of  $\hat{\boldsymbol{\theta}}$  can be obtained from Theorem 11.1 by the delta method.

**Theorem 11.2** Under Assumptions 7.2 and 7.3,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, V_{\boldsymbol{\theta}})$$

where

$$V_{\boldsymbol{\theta}} = \mathbf{R}' V_{\boldsymbol{\beta}} \mathbf{R}$$

$$\mathbf{R} = \frac{\partial}{\partial \boldsymbol{\beta}} r(\boldsymbol{\beta})'.$$

For a proof, see Exercise 11.7.

Theorem 11.2 is an example where multivariate regression is fundamentally distinct from univariate regression. Only by treating the least-squares estimates as a joint estimator can we obtain a distributional theory for an estimator  $\hat{\boldsymbol{\theta}}$  which is a function of estimates from multiple equations and thereby construct standard errors, confidence intervals, and hypothesis tests.

## 11.6 Covariance Matrix Estimation

From the finite sample and asymptotic theory we can construct appropriate estimators for the variance of  $\hat{\boldsymbol{\beta}}$ . In the general case we have

$$\hat{V}_{\hat{\boldsymbol{\beta}}} = (\bar{\mathbf{X}}' \bar{\mathbf{X}})^{-1} \left( \sum_{i=1}^n \bar{\mathbf{X}}'_i \hat{\mathbf{e}}_i \hat{\mathbf{e}}'_i \bar{\mathbf{X}}_i \right) (\bar{\mathbf{X}}' \bar{\mathbf{X}})^{-1}.$$

Under conditional homoskedasticity (11.8) an appropriate estimator is

$$\widehat{V}_{\widehat{\beta}}^0 = \left( \overline{\mathbf{X}}' \overline{\mathbf{X}} \right)^{-1} \left( \sum_{i=1}^n \overline{\mathbf{X}}_i' \widehat{\Sigma} \overline{\mathbf{X}}_i \right) \left( \overline{\mathbf{X}}' \overline{\mathbf{X}} \right)^{-1}.$$

When the regressors are common then these estimators equal

$$\widehat{V}_{\widehat{\beta}} = \left( \mathbf{I}_m \otimes (\mathbf{X}' \mathbf{X})^{-1} \right) \left( \sum_{i=1}^n (\widehat{\mathbf{e}}_i \widehat{\mathbf{e}}_i' \otimes \mathbf{x}_i \mathbf{x}_i') \right) \left( \mathbf{I}_m \otimes (\mathbf{X}' \mathbf{X})^{-1} \right)$$

and

$$\widehat{V}_{\widehat{\beta}}^0 = \widehat{\Sigma} \otimes (\mathbf{X}' \mathbf{X})^{-1},$$

respectively.

Covariance matrix estimators for  $\widehat{\theta}$  are found as

$$\begin{aligned}\widehat{V}_{\widehat{\theta}} &= \widehat{\mathbf{R}}' \widehat{V}_{\widehat{\beta}} \widehat{\mathbf{R}} \\ \widehat{V}_{\widehat{\theta}}^0 &= \widehat{\mathbf{R}}' \widehat{V}_{\widehat{\beta}}^0 \widehat{\mathbf{R}} \\ \widehat{\mathbf{R}} &= \frac{\partial}{\partial \beta} \mathbf{r}(\widehat{\beta})'.\end{aligned}$$

**Theorem 11.3** Under Assumption 7.2,

$$n \widehat{V}_{\widehat{\beta}} \xrightarrow{p} V_{\beta}$$

and

$$n \widehat{V}_{\widehat{\beta}}^0 \xrightarrow{p} V_{\beta}^0.$$

For a proof, see Exercise 11.8.

## 11.7 Seemingly Unrelated Regression

Consider the systems regression model under the conditional mean and conditional homoskedasticity assumptions

$$\begin{aligned}\mathbf{y}_i &= \overline{\mathbf{X}}_i \beta + \mathbf{e}_i \\ \mathbb{E}(\mathbf{e}_i | \mathbf{x}_i) &= \mathbf{0} \\ \mathbb{E}(\mathbf{e}_i \mathbf{e}_i' | \mathbf{x}_i) &= \Sigma\end{aligned}\tag{11.15}$$

Since the errors are correlated across equations we can consider estimation by Generalized Least Squares (GLS). To derive the estimator, premultiply (11.15) by  $\Sigma^{-1/2}$  so that the transformed error vector is i.i.d. with covariance matrix  $\mathbf{I}_m$ . Then apply least-squares and rearrange to find

$$\widehat{\beta}_{\text{gls}} = \left( \sum_{i=1}^n \overline{\mathbf{X}}_i' \Sigma^{-1} \overline{\mathbf{X}}_i \right)^{-1} \left( \sum_{i=1}^n \overline{\mathbf{X}}_i' \Sigma^{-1} \mathbf{y}_i \right).\tag{11.16}$$

(see Exercise 11.9). Another approach is to take the vector representation

$$\mathbf{y} = \overline{\mathbf{X}} \beta + \mathbf{e}$$

and calculate that the equation error  $\mathbf{e}$  has variance  $\mathbb{E}(\mathbf{e}\mathbf{e}') = \mathbf{I}_n \otimes \boldsymbol{\Sigma}$ . Premultiply the equation by  $\mathbf{I}_n \otimes \boldsymbol{\Sigma}^{-1/2}$  so that the transformed error has variance matrix  $\mathbf{I}_{nm}$  and then apply least-squares to find

$$\hat{\boldsymbol{\beta}}_{\text{gls}} = \left( \bar{\mathbf{X}}' (\mathbf{I}_n \otimes \boldsymbol{\Sigma}^{-1}) \bar{\mathbf{X}} \right)^{-1} \left( \bar{\mathbf{X}}' (\mathbf{I}_n \otimes \boldsymbol{\Sigma}^{-1}) \mathbf{y} \right) \quad (11.17)$$

(see Exercise 11.10).

Expressions (11.16) and (11.17) are algebraically equivalent. To see the equivalence, observe that

$$\begin{aligned} \bar{\mathbf{X}}' (\mathbf{I}_n \otimes \boldsymbol{\Sigma}^{-1}) \bar{\mathbf{X}} &= \begin{pmatrix} \bar{\mathbf{X}}'_1 & \dots & \bar{\mathbf{X}}'_n \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}^{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \boldsymbol{\Sigma}^{-1} & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}^{-1} \end{pmatrix} \begin{pmatrix} \bar{\mathbf{X}}_1 \\ \vdots \\ \bar{\mathbf{X}}_n \end{pmatrix} \\ &= \sum_{i=1}^n \bar{\mathbf{X}}'_i \boldsymbol{\Sigma}^{-1} \bar{\mathbf{X}}_i \end{aligned}$$

and

$$\begin{aligned} \bar{\mathbf{X}}' (\mathbf{I}_n \otimes \boldsymbol{\Sigma}^{-1}) \mathbf{y} &= \begin{pmatrix} \bar{\mathbf{X}}'_1 & \dots & \bar{\mathbf{X}}'_n \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}^{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \boldsymbol{\Sigma}^{-1} & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{pmatrix} \\ &= \sum_{i=1}^n \bar{\mathbf{X}}'_i \boldsymbol{\Sigma}^{-1} \mathbf{y}_i. \end{aligned}$$

Since  $\boldsymbol{\Sigma}$  is unknown it must be replaced by an estimator. Using  $\hat{\boldsymbol{\Sigma}}$  from (11.5) we obtain a feasible GLS estimator.

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{sur}} &= \left( \sum_{i=1}^n \bar{\mathbf{X}}'_i \hat{\boldsymbol{\Sigma}}^{-1} \bar{\mathbf{X}}_i \right)^{-1} \left( \sum_{i=1}^n \bar{\mathbf{X}}'_i \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{y}_i \right) \\ &= \left( \bar{\mathbf{X}}' (\mathbf{I}_n \otimes \hat{\boldsymbol{\Sigma}}^{-1}) \bar{\mathbf{X}} \right)^{-1} \left( \bar{\mathbf{X}}' (\mathbf{I}_n \otimes \hat{\boldsymbol{\Sigma}}^{-1}) \mathbf{y} \right). \end{aligned} \quad (11.18)$$

This is known as the **Seemingly Unrelated Regression (SUR)** estimator, and was introduced by Zellner (1962).

The estimator  $\hat{\boldsymbol{\Sigma}}$  can be updated by calculating the SUR residuals  $\hat{\mathbf{e}}_i = \mathbf{y}_i - \bar{\mathbf{X}}'_i \hat{\boldsymbol{\beta}}_{\text{sur}}$  and the covariance matrix estimate  $\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^n \hat{\mathbf{e}}_i \hat{\mathbf{e}}_i'$ . Substituted into (11.18) we find an iterated SUR estimator, and this can be iterated until convergence.

Under conditional homoskedasticity (11.8) we can derive its asymptotic distribution.

**Theorem 11.4** Under Assumption 7.2 and (11.8)

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{sur}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, V_{\boldsymbol{\beta}}^*)$$

where

$$V_{\boldsymbol{\beta}}^* = \left( \mathbb{E}(\bar{\mathbf{X}}'_i \boldsymbol{\Sigma}^{-1} \bar{\mathbf{X}}_i) \right)^{-1}.$$

For a proof, see Exercise 11.11.

Under these assumptions (in particular conditional homoskedasticity), SUR is more efficient than least-squares.

**Theorem 11.5** Under Assumption 7.2 and (11.8)

$$\begin{aligned} V_{\beta}^* &= \left( \mathbb{E} \left( \bar{\mathbf{X}}_i' \Sigma^{-1} \bar{\mathbf{X}}_i \right) \right)^{-1} \\ &\leq \left( \mathbb{E} \left( \bar{\mathbf{X}}_i' \bar{\mathbf{X}}_i \right) \right)^{-1} \mathbb{E} \left( \bar{\mathbf{X}}_i' \Sigma \bar{\mathbf{X}}_i \right) \left( \mathbb{E} \left( \bar{\mathbf{X}}_i' \bar{\mathbf{X}}_i \right) \right)^{-1} \\ &= V_{\beta} \end{aligned}$$

and thus  $\hat{\beta}_{\text{sur}}$  is asymptotically more efficient than  $\hat{\beta}_{\text{ols}}$ .

For a proof, see Exercise 11.12.

An appropriate estimator of the variance of  $\hat{\beta}_{\text{sur}}$  is

$$\hat{V}_{\hat{\beta}} = \left( \sum_{i=1}^n \bar{\mathbf{X}}_i' \hat{\Sigma}^{-1} \bar{\mathbf{X}}_i \right)^{-1}.$$

**Theorem 11.6** Under Assumption 7.2 and (11.8)

$$n \hat{V}_{\hat{\beta}} \xrightarrow{p} V_{\beta}.$$

For a proof, see Exercise 11.13.

In Stata, the seemingly unrelated regressions estimator is implemented using the `sureg` command.

### Arnold Zellner

Arnold Zellner (1927-2000) of the United States was a founding father of the econometrics field. He was a pioneer in Bayesian econometrics. One of his core contributions was the method of Seemingly Unrelated Regressions.

## 11.8 Equivalence of SUR and Least-Squares

When the regressors are common across equations  $\mathbf{x}_{ji} = \mathbf{x}_i$  it turns out that the SUR estimator simplifies to least-squares.

To see this, recall that when regressors are common this implies that  $\bar{\mathbf{X}}_i = \mathbf{I}_m \otimes \mathbf{x}'_i$ . Then

$$\begin{aligned} \bar{\mathbf{X}}_i' \hat{\Sigma}^{-1} &= (\mathbf{I}_m \otimes \mathbf{x}_i)' \hat{\Sigma}^{-1} \\ &= \hat{\Sigma}^{-1} \otimes \mathbf{x}_i \\ &= (\hat{\Sigma}^{-1} \otimes \mathbf{I}_k)(\mathbf{I}_m \otimes \mathbf{x}_i) \\ &= (\hat{\Sigma}^{-1} \otimes \mathbf{I}_k) \bar{\mathbf{X}}_i'. \end{aligned}$$

Thus

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{sur}} &= \left( \sum_{i=1}^n \bar{\mathbf{X}}_i' \hat{\boldsymbol{\Sigma}}^{-1} \bar{\mathbf{X}}_i \right)^{-1} \left( \sum_{i=1}^n \bar{\mathbf{X}}_i' \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{y}_i \right) \\ &= \left( (\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_k) \sum_{i=1}^n \bar{\mathbf{X}}_i \bar{\mathbf{X}}_i' \right)^{-1} \left( (\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_k) \sum_{i=1}^n \bar{\mathbf{X}}_i' \mathbf{y}_i \right) \\ &= \left( \sum_{i=1}^n \bar{\mathbf{X}}_i \bar{\mathbf{X}}_i' \right)^{-1} \left( \sum_{i=1}^n \bar{\mathbf{X}}_i' \mathbf{y}_i \right) \\ &= \hat{\boldsymbol{\beta}}_{\text{ols}}.\end{aligned}$$

A model where regressors are not common across equations is nested within a model with the union of all regressors included in all equations. Thus the model with regressors common across equations is a fully unrestricted model, and a model where the regressors differ across equations is a restricted model. Thus the above result shows that the SUR estimator reduces to least-squares in the absence of restrictions, but SUR can differ from least-squares otherwise.

## 11.9 Maximum Likelihood Estimator

Take the linear model under the assumption that the error is independent of the regressors and multivariate normally distributed. Thus

$$\begin{aligned}\mathbf{y}_i &= \bar{\mathbf{X}}_i \boldsymbol{\beta} + \mathbf{e}_i \\ \mathbf{e}_i &\sim N(\mathbf{0}, \boldsymbol{\Sigma}).\end{aligned}$$

In this case we can consider the maximum likelihood estimator (MLE) of the coefficients.

It is convenient to reparameterize the covariance matrix in terms of its inverse, thus  $\mathbf{S} = \boldsymbol{\Sigma}^{-1}$ . With this reparameterization, the conditional density of  $\mathbf{y}_i$  given  $\mathbf{X}_i$  equals

$$f(\mathbf{y}_i | \mathbf{X}_i) = \frac{\det(\mathbf{S})^{1/2}}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})' \mathbf{S} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})\right).$$

The log-likelihood function for the sample is

$$\log L(\boldsymbol{\beta}, \mathbf{S}) = -\frac{nm}{2} \log(2\pi) + \frac{n}{2} \log \det(\mathbf{S}) - \frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{X}}_i \boldsymbol{\beta})' \mathbf{S} (\mathbf{y}_i - \bar{\mathbf{X}}_i \boldsymbol{\beta}).$$

The maximum likelihood estimator  $(\hat{\boldsymbol{\beta}}_{\text{mle}}, \hat{\mathbf{S}}_{\text{mle}})$  maximizes the log-likelihood function. The first order conditions are

$$\begin{aligned}0 &= \frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta}, \mathbf{S}) \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \mathbf{S}=\hat{\mathbf{S}}} \\ &= \sum_{i=1}^n \bar{\mathbf{X}}_i \hat{\mathbf{S}} (\mathbf{y}_i - \bar{\mathbf{X}}_i \hat{\boldsymbol{\beta}})\end{aligned}$$

and

$$\begin{aligned}0 &= \frac{\partial}{\partial \mathbf{S}} \log L(\boldsymbol{\beta}, \mathbf{S}) \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \mathbf{S}=\hat{\mathbf{S}}} \\ &= \frac{n}{2} \hat{\mathbf{S}}^{-1} - \frac{1}{2} \text{tr}\left(\sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{X}}_i \hat{\boldsymbol{\beta}})' (\mathbf{y}_i - \bar{\mathbf{X}}_i \hat{\boldsymbol{\beta}})\right).\end{aligned}$$

The second equation uses the matrix results  $\frac{\partial}{\partial \mathbf{S}} \log \det(\mathbf{S}) = \mathbf{S}^{-1}$  and  $\frac{\partial}{\partial \mathbf{B}} \text{tr}(\mathbf{AB}) = \mathbf{A}'$  from Appendix A.20.

Solving and making the substitution  $\widehat{\Sigma} = \widehat{S}^{-1}$  we obtain

$$\begin{aligned}\widehat{\beta}_{\text{mle}} &= \left( \sum_{i=1}^n \bar{X}'_i \widehat{\Sigma}^{-1} \bar{X}_i \right)^{-1} \left( \sum_{i=1}^n \bar{X}'_i \widehat{\Sigma}^{-1} y_i \right) \\ \widehat{\Sigma}_{\text{mle}} &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{X}_i \widehat{\beta}) (y_i - \bar{X}_i \widehat{\beta})'.\end{aligned}$$

Notice that each equation refers to the other. Hence these are not closed-form expressions, but can be solved via iteration. The solution is identical to the iterated SUR estimator. Thus the SUR estimator (iterated) is identical to the MLE under normality.

Recall that the SUR estimator simplifies to OLS when the regressors are common across equations. The same occurs for the MLE. Thus when  $\bar{X}_i = I_m \otimes x'_i$  we find that  $\widehat{\beta}_{\text{mle}} = \widehat{\beta}_{\text{ols}}$  and  $\widehat{\Sigma}_{\text{mle}} = \widehat{\Sigma}_{\text{ols}}$ .

## 11.10 Restricted Estimation

In many multivariate regression applications it is desired to impose restrictions on the coefficients. In particular, cross-equation restrictions (for example, imposing Slutsky symmetry on a demand system) can be quite important, and can only be imposed by a multivariate estimation method. Estimation subject to restrictions can be done by minimum distance, maximum likelihood, or the generalized method of moments.

Minimum distance is a straightforward application of the methods of Chapter 8 to the estimators presented in this chapter, since such methods apply to any asymptotically normal unrestricted estimator.

Imposing restrictions on maximum likelihood is also straightforward. The likelihood is maximized subject to the imposed restrictions. One important example is explored in detail in the following section.

Generalized method of moments estimation of multivariate regression subject to restrictions will be explored in Section 13.18. This is a particularly simple and straightforward way to estimate restricted multivariate regression models, and is our generally preferred approach.

## 11.11 Reduced Rank Regression

One context where systems estimation is important is when it is desired to impose or test restrictions across equations. Restricted systems are commonly estimated by maximum likelihood under normality. In this section we explore one important special case of restricted multivariate regression known as reduced rank regression. The model was originally proposed by Anderson (1951) and extended by Johansen (1995).

The unrestricted model is

$$\begin{aligned}y_i &= \mathbf{B}' \mathbf{x}_i + \mathbf{C}' \mathbf{z}_i + \mathbf{e}_i \\ \mathbb{E}(\mathbf{e}_i \mathbf{e}'_i | \mathbf{x}_i, \mathbf{z}_i) &= \Sigma\end{aligned}\tag{11.19}$$

where  $\mathbf{B}$  is  $k \times m$ ,  $\mathbf{C}$  is  $\ell \times m$ , and  $\mathbf{x}_i$  and  $\mathbf{z}_i$  are regressors. We separate the regressors  $\mathbf{x}_i$  and  $\mathbf{z}_i$  because the coefficient matrix  $\mathbf{B}$  will be restricted while  $\mathbf{C}$  will be unrestricted.

The matrix  $\mathbf{B}$  is full rank if

$$\text{rank}(\mathbf{B}) = \min(k, m).$$

The reduced rank restriction is that

$$\text{rank}(\mathbf{B}) = r < \min(k, m)$$

for some known  $r$ .

The reduced rank restriction implies that we can write the coefficient matrix  $\mathbf{B}$  in the factored form

$$\mathbf{B} = \mathbf{G} \mathbf{A}'$$

where  $\mathbf{A}$  is  $m \times r$  and  $\mathbf{G}$  is  $k \times r$ . This representation is not unique (as we can replace  $\mathbf{G}$  with  $\mathbf{G}\mathbf{Q}$  and  $\mathbf{A}$  with  $\mathbf{A}\mathbf{Q}^{-1}$  for any invertible  $\mathbf{Q}$  and the same relation holds). Identification therefore requires a normalization of the coefficients. A conventional normalization is

$$\mathbf{G}'\mathbf{D}\mathbf{G} = \mathbf{I}_r$$

for given  $\mathbf{D}$ .

Equivalently, the reduced rank restriction can be imposed by requiring that  $\mathbf{B}$  satisfy the restriction  $\mathbf{B}\mathbf{A}_\perp = \mathbf{G}\mathbf{A}'\mathbf{A}_\perp = \mathbf{0}$  for some  $m \times (m - r)$  coefficient matrix  $\mathbf{A}_\perp$ . Since  $\mathbf{G}$  is full rank this requires that  $\mathbf{A}'\mathbf{A}_\perp = \mathbf{0}$ , hence  $\mathbf{A}_\perp$  is the orthogonal complement to  $\mathbf{A}$ . Note that  $\mathbf{A}_\perp$  is not unique as it can be replaced by  $\mathbf{A}_\perp\mathbf{Q}$  for any  $(m - r) \times (m - r)$  invertible  $\mathbf{Q}$ . Thus if  $\mathbf{A}_\perp$  is to be estimated it requires a normalization.

We discuss methods for estimation of  $\mathbf{G}$ ,  $\mathbf{A}$ ,  $\Sigma$ ,  $\mathbf{C}$ , and  $\mathbf{A}_\perp$ . The standard approach is maximum likelihood under the assumption that  $\mathbf{e}_i \sim N(\mathbf{0}, \Sigma)$ . The log-likelihood function for the sample is

$$\begin{aligned} \log L(\mathbf{G}, \mathbf{A}, \mathbf{C}, \Sigma) &= -\frac{nm}{2} \log(2\pi) - \frac{n}{2} \log \det(\Sigma) \\ &\quad - \frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \mathbf{AG}'\mathbf{x}_i - \mathbf{C}'\mathbf{z}_i)' \Sigma^{-1} (\mathbf{y}_i - \mathbf{AG}'\mathbf{x}_i - \mathbf{C}'\mathbf{z}_i). \end{aligned}$$

Anderson (1951) derived the MLE by imposing the constraint  $\mathbf{B}\mathbf{A}_\perp = \mathbf{0}$  via the method of Lagrange multipliers. This turns out to be algebraically cumbersome.

Johansen (1995) instead proposed a concentration method which turns out to be relatively straightforward. The method is as follows. First, treat  $\mathbf{G}$  as if it is known. Then maximize the log-likelihood with respect to the other parameters. Resubstituting these estimates, we obtain the concentrated log-likelihood function with respect to  $\mathbf{G}$ . This can be maximized to find the MLE for  $\mathbf{G}$ . The other parameter estimates are then obtained by substitution. We now describe these steps in detail.

Given  $\mathbf{G}$ , the likelihood is a normal multivariate regression in the variables  $\mathbf{G}'\mathbf{x}_i$  and  $\mathbf{z}_i$ , so the MLE for  $\mathbf{A}$ ,  $\mathbf{C}$  and  $\Sigma$  are least-squares. In particular, using the Frisch-Waugh-Lovell residual regression formula, we can write the estimators for  $\mathbf{A}$  and  $\Sigma$  as

$$\widehat{\mathbf{A}}(\mathbf{G}) = (\tilde{\mathbf{Y}}'\tilde{\mathbf{X}}\mathbf{G}) (\mathbf{G}'\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\mathbf{G})^{-1}$$

and

$$\widehat{\Sigma}(\mathbf{G}) = \frac{1}{n} \left( \tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} - \tilde{\mathbf{Y}}'\tilde{\mathbf{X}}\mathbf{G} (\mathbf{G}'\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\mathbf{G})^{-1} \mathbf{G}'\tilde{\mathbf{X}}'\tilde{\mathbf{Y}} \right)$$

where

$$\begin{aligned} \tilde{\mathbf{Y}} &= \mathbf{Y} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} \\ \tilde{\mathbf{X}} &= \mathbf{X} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}. \end{aligned}$$

Substituting these estimators into the log-likelihood function, we obtain the concentrated likelihood function, which is a function of  $\mathbf{G}$  only

$$\begin{aligned} \log \tilde{L}(\mathbf{G}) &= \log L(\mathbf{G}, \widehat{\mathbf{A}}(\mathbf{G}), \widehat{\mathbf{C}}(\mathbf{G}), \widehat{\Sigma}(\mathbf{G})) \\ &= \frac{m}{2} (n \log(2\pi) - 1) - \frac{n}{2} \log \det \left( \tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} - \tilde{\mathbf{Y}}'\tilde{\mathbf{X}}\mathbf{G} (\mathbf{G}'\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\mathbf{G})^{-1} \mathbf{G}'\tilde{\mathbf{X}}'\tilde{\mathbf{Y}} \right) \\ &= \frac{m}{2} (n \log(2\pi) - 1) - \frac{n}{2} \log \det(\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}) \frac{\det \left( \mathbf{G}' \left( \tilde{\mathbf{X}}'\tilde{\mathbf{X}} - \tilde{\mathbf{X}}'\tilde{\mathbf{Y}}(\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}})^{-1}\mathbf{Y}'\tilde{\mathbf{X}} \right) \mathbf{G} \right)}{\det(\mathbf{G}'\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\mathbf{G})}. \end{aligned}$$

The third equality uses Theorem A.1.8. The MLE  $\widehat{\mathbf{G}}$  for  $\mathbf{G}$  is the maximizer of  $\log \tilde{L}(\mathbf{G})$ , or equivalently equals

$$\begin{aligned}\widehat{\mathbf{G}} &= \underset{\mathbf{G}}{\operatorname{argmin}} \frac{\det\left(\mathbf{G}'\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} - \tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}\left(\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}\right)^{-1}\mathbf{Y}'\tilde{\mathbf{X}}\right)\mathbf{G}\right)}{\det\left(\mathbf{G}'\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\mathbf{G}\right)} \\ &= \underset{\mathbf{G}}{\operatorname{argmax}} \frac{\det\left(\mathbf{G}'\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}\left(\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}\right)^{-1}\mathbf{Y}'\tilde{\mathbf{X}}\mathbf{G}\right)}{\det\left(\mathbf{G}'\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\mathbf{G}\right)} \\ &= \{\mathbf{v}_1, \dots, \mathbf{v}_r\}\end{aligned}\tag{11.20}$$

which are the generalized eigenvectors of  $\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}\left(\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}\right)^{-1}\mathbf{Y}'\tilde{\mathbf{X}}$  with respect to  $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$  corresponding to the  $r$  largest generalized eigenvalues. (Generalized eigenvalues and eigenvectors are discussed in Section A.14.) The estimator satisfies the normalization  $\widehat{\mathbf{G}}'\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\widehat{\mathbf{G}} = \mathbf{I}_r$ . Letting  $\mathbf{v}_j^*$  denote the eigenvectors of (11.20), we can also express  $\widehat{\mathbf{G}} = \{\mathbf{v}_m^*, \dots, \mathbf{v}_{m-r+1}^*\}$ .

This is computationally straightforward. In MATLAB, for example, the generalized eigenvalues and eigenvectors of a matrix  $\mathbf{A}$  with respect to  $\mathbf{B}$  are found using the command `eig(A, B)`.

Given  $\widehat{\mathbf{G}}$ , the MLE  $\widehat{\mathbf{A}}$ ,  $\widehat{\mathbf{C}}$ ,  $\widehat{\Sigma}$  are found by least-squares regression of  $\mathbf{y}_i$  on  $\widehat{\mathbf{G}}'\mathbf{x}_i$  and  $\mathbf{z}_i$ . In particular,  $\widehat{\mathbf{A}} = \widehat{\mathbf{G}}'\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}$  since  $\widehat{\mathbf{G}}'\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\widehat{\mathbf{G}} = \mathbf{I}_r$ .

We now discuss the estimator  $\widehat{\mathbf{A}}_\perp$  of  $\mathbf{A}_\perp$ . It turns out that

$$\begin{aligned}\widehat{\mathbf{A}}_\perp &= \underset{\mathbf{A}}{\operatorname{argmax}} \frac{\det\left(\mathbf{A}'\left(\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} - \tilde{\mathbf{Y}}'\tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}\right)\mathbf{A}\right)}{\det\left(\mathbf{A}'\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}\mathbf{A}\right)} \\ &= \{\mathbf{w}_1, \dots, \mathbf{w}_{m-r}\}\end{aligned}\tag{11.21}$$

the eigenvectors of  $\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} - \tilde{\mathbf{Y}}'\tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}$  with respect to  $\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}$  associated with the largest  $m-r$  eigenvalues.

By the dual eigenvalue relation (Theorem A.5), the eigenvalue problems in equations (11.20) and (11.21) have the same non-unit eigenvalues  $\lambda_j$ , and the associated eigenvectors  $\mathbf{v}_j^*$  and  $\mathbf{w}_j$  satisfy the relationship

$$\mathbf{w}_j = \lambda_j^{-1/2} \left(\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}\right)^{-1} \tilde{\mathbf{Y}}'\tilde{\mathbf{X}}\mathbf{v}_j^*.$$

Letting  $\Lambda = \operatorname{diag}\{\lambda_m, \dots, \lambda_{m-r+1}\}$  this implies

$$\begin{aligned}\{\mathbf{w}_m, \dots, \mathbf{w}_{m-r+1}\} &= \left(\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}\right)^{-1} \tilde{\mathbf{Y}}'\tilde{\mathbf{X}}\{\mathbf{v}_m^*, \dots, \mathbf{v}_{m-r+1}^*\} \Lambda \\ &= \left(\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}\right)^{-1} \widehat{\mathbf{A}}\Lambda.\end{aligned}$$

The second equality holds since  $\widehat{\mathbf{G}} = \{\mathbf{v}_m^*, \dots, \mathbf{v}_{m-r+1}^*\}$  and  $\widehat{\mathbf{A}} = \tilde{\mathbf{Y}}'\tilde{\mathbf{X}}\widehat{\mathbf{G}}$ . Since the eigenvectors  $\mathbf{w}_j$  satisfy the orthogonality property  $\mathbf{w}_j'\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}\mathbf{w}_\ell = 0$  for  $j \neq \ell$ , it follows that

$$0 = \widehat{\mathbf{A}}_\perp'\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}\{\mathbf{w}_m, \dots, \mathbf{w}_{m-r+1}\} = \widehat{\mathbf{A}}_\perp'\widehat{\mathbf{A}}\Lambda.$$

Since  $\Lambda > 0$  we conclude that  $\widehat{\mathbf{A}}_\perp'\widehat{\mathbf{A}} = 0$  as desired.

The solution  $\widehat{\mathbf{A}}_\perp$  in (11.21) can be represented several ways. One which is computationally convenient is to observe that

$$\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} - \tilde{\mathbf{Y}}'\tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{Y}}'\tilde{\mathbf{X}} = \mathbf{Y}'\mathbf{M}_{\mathbf{X}, \mathbf{Z}}\mathbf{Y} = \tilde{\mathbf{e}}'\tilde{\mathbf{e}}$$

where  $\mathbf{M}_{\mathbf{X}, \mathbf{Z}} = \mathbf{I}_n - (\mathbf{X}, \mathbf{Z})((\mathbf{X}, \mathbf{Z})'(\mathbf{X}, \mathbf{Z}))^{-1}(\mathbf{X}, \mathbf{Z})'$  and  $\tilde{\mathbf{e}} = \mathbf{M}_{\mathbf{X}, \mathbf{Z}}\mathbf{Y}$  is the residual from the unrestricted least-squares regression of  $\mathbf{Y}$  on  $\mathbf{X}$  and  $\mathbf{Z}$ . The first equality follows by the Frisch-Waugh-Lovell theorem.

This shows that  $\widehat{\mathbf{A}}_{\perp}$  are the generalized eigenvectors of  $\tilde{\mathbf{e}}'\tilde{\mathbf{e}}$  with respect to  $\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}$  corresponding to the  $m-r$  largest eigenvalues. In MATLAB, for example, these can be computed using the `eig(A, B)` command.

Another representation is to write  $\mathbf{M}_Z = \mathbf{I}_n - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$  so that

$$\widehat{\mathbf{A}}_{\perp} = \operatorname{argmax}_{\mathbf{A}} \frac{\det(\mathbf{A}'\mathbf{Y}'\mathbf{M}_{X,Z}\mathbf{Y}\mathbf{A})}{\det(\mathbf{A}'\mathbf{Y}'\mathbf{M}_Z\mathbf{Y}\mathbf{A})} = \operatorname{argmin}_{\mathbf{A}} \frac{\det(\mathbf{A}'\mathbf{Y}'\mathbf{M}_Z\mathbf{Y}\mathbf{A})}{\det(\mathbf{A}'\mathbf{Y}'\mathbf{M}_{X,Z}\mathbf{Y}\mathbf{A})}$$

We summarize our findings.

**Theorem 11.7** The MLE for the reduced rank model (11.19) under  $\mathbf{e}_i \sim N(\mathbf{0}, \Sigma)$  is given as follows.  $\widehat{\mathbf{G}}_{\text{mle}} = \{\mathbf{v}_1, \dots, \mathbf{v}_r\}$ , the generalized eigenvectors of  $\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}(\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}})^{-1}\mathbf{Y}'\tilde{\mathbf{X}}$  with respect to  $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$  corresponding to the  $r$  largest eigenvalues.  $\widehat{\mathbf{A}}_{\text{mle}}$ ,  $\widehat{\mathbf{C}}_{\text{mle}}$  and  $\widehat{\Sigma}_{\text{mle}}$  are obtained by the least-squares regression

$$\begin{aligned}\mathbf{y}_i &= \widehat{\mathbf{A}}_{\text{mle}}\widehat{\mathbf{G}}'_{\text{mle}}\mathbf{x}_i + \widehat{\mathbf{C}}'_{\text{mle}}\mathbf{z}_i + \widehat{\mathbf{e}}_i \\ \widehat{\Sigma}_{\text{mle}} &= \frac{1}{n} \sum_{i=1}^n \widehat{\mathbf{e}}_i \widehat{\mathbf{e}}'_i.\end{aligned}$$

$\widehat{\mathbf{A}}_{\perp}$  equals the generalized eigenvectors of  $\tilde{\mathbf{e}}'\tilde{\mathbf{e}}$  with respect to  $\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}$  corresponding to the  $m-r$  smallest eigenvalues.

## 11.12 Principal Component Analysis

Recall in Section 4.21 we described the Duflo, Dupas and Kremer (2011) dataset which contains a sample of Kenyan first grade students and their test scores. Following the authors we had estimated regressions attempting to explain the variable *totalscore*, which was each student's composite test score. However, if you examine the data file you will find a large number of other pieces of information, including each student's score on the separate sections of the test, with the labels *wordscore* (word recognition), *sentscore* (sentence recognition), *letterscore* (letter recognition), *spellscore* (spelling), *additions\_score* (addition), *subtractions\_score* (subtraction), *multiplications\_score* (multiplication). The “total” score sums the scores from the individual sections. Perhaps there is more information in the individual scores. How can we learn about this from the data?

**Principal component analysis (PCA)** addresses this issue by building models consisting of a common component and an idiosyncratic component. Let  $\mathbf{x}_i$  be a  $k \times 1$  vector (for example the seven test sub-scores described above) of observations for individual  $i$ . The elements of  $\mathbf{x}_i$  should be standardized to have mean zero and unit variance. A **single factor model** takes the form

$$\mathbf{x}_i = \mathbf{h}f_i + \mathbf{u}_i \tag{11.22}$$

where  $\mathbf{x}_i$ ,  $\mathbf{h}$  and  $\mathbf{u}_i$  are  $k \times 1$  and  $f_i$  is scalar. The random variable  $f_i$  is known as the **common factor** and the random vector  $\mathbf{u}_i$  is the **individual component**. The vector  $\mathbf{h}$  is called the **factor loadings**. The scale of  $\mathbf{h}$  and  $f_i$  are not separately identified, so a normalization is required. A typical choice is to normalize  $\mathbf{h}$  to have unit length,  $\mathbf{h}'\mathbf{h} = 1$ . The sign of  $\mathbf{h}$  and  $f_i$  are also not separately identified, so another normalization is needed. One choice is to set the sign so that  $\sum_{i=1}^n f_i > 0$ . Let  $\sigma_f^2 = \mathbb{E}(f_i^2)$  be a free parameter.

Economists typically refer to (11.22) as a **factor model**. Other disciplines reserve that label for similar but distinct models.

The way to think about (11.22) in the student test score example is that  $f_i$  is a student's scholastic “aptitude” and the vector  $\mathbf{h}$  describes how scholastic aptitude affects the seven sub-sections of the test. We would expect the elements of  $\mathbf{h}$  to all be positive, indicating that scholastic aptitude is related to improved performance in all seven test areas.

Equation (11.22) decomposes the vector of observables  $\mathbf{x}_i$  into the components  $f_i$  and  $\mathbf{u}_i$ . The model is typically completed by the assumption that the elements of  $f_i$  and  $\mathbf{u}_i$  are mutually uncorrelated, and the elements of  $\mathbf{u}_i$  have common variances so that  $\mathbb{E}(\mathbf{u}_i \mathbf{u}'_i) = \mathbf{I}_k \sigma_u^2$ .

Under the assumptions described above the covariance matrix of  $\mathbf{x}_i$  takes the form

$$\begin{aligned}\boldsymbol{\Sigma}_x &= \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i) \\ &= \mathbf{h} \mathbf{h}' \sigma_f^2 + \mathbf{I}_k \sigma_u^2.\end{aligned}$$

In fact this summarizes the implications of the assumptions. An alternative way of viewing the model (11.22) is that it is equivalent to restricting the covariance matrix  $\boldsymbol{\Sigma}_x$  to take this form.

Notice that since  $\mathbf{h}' \mathbf{h} = 1$

$$\boldsymbol{\Sigma}_x \mathbf{h} = \mathbf{h} \mathbf{h}' \lambda \sigma_f^2 + \mathbf{I}_k \lambda \sigma_u^2 = \mathbf{h} \sigma_f^2 + \mathbf{h} \sigma_u^2 = \mathbf{h} (\sigma_f^2 + \sigma_u^2).$$

This means that  $\mathbf{h}$  is an eigenvector of  $\boldsymbol{\Sigma}_x$  with associated eigenvalue  $\sigma_f^2 + \sigma_u^2$ . Let  $\mathbf{h}_j$  be any other eigenvector of  $\boldsymbol{\Sigma}_x$ . Since  $\mathbf{h}' \mathbf{h}_j = 0$ ,

$$\boldsymbol{\Sigma}_x \mathbf{h}_j = \mathbf{h} \mathbf{h}' \mathbf{h}_j \sigma_f^2 + \mathbf{I}_k \mathbf{h}_j \sigma_u^2 = \mathbf{h}_j \sigma_u^2$$

so its associated eigenvalue is  $\sigma_u^2$ . Thus  $\boldsymbol{\Sigma}_x$  has eigenvalues  $\sigma_f^2 + \sigma_u^2$  and  $\sigma_u^2$ , the latter with multiplicity  $k - 1$ . So  $\mathbf{h}$  is the eigenvector associated with the largest eigenvalue (if  $\sigma_f^2$  is strictly positive). The proportional contribution of this factor to the total variance is  $\lambda_1 / \sum_{m=1}^M \lambda_m$  where  $\lambda_m$  are the eigenvalues of  $\boldsymbol{\Sigma}_x$ .

This suggests that the factor loading can be estimated by the leading eigenvector of the sample covariance matrix  $\widehat{\boldsymbol{\Sigma}}_x = n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i$ . Let  $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \dots \geq \widehat{\lambda}_k$  be the eigenvalues of  $\widehat{\boldsymbol{\Sigma}}_x$  and  $\widehat{\mathbf{h}}_1, \widehat{\mathbf{h}}_2, \dots, \widehat{\mathbf{h}}_k$  the associated eigenvectors. The estimator of  $\mathbf{h}$  is  $\widehat{\mathbf{h}}_1$ .

A multiple factor model takes the form

$$\begin{aligned}\mathbf{x}_i &= \sum_{m=1}^r \mathbf{h}_m f_{mi} + \mathbf{u}_i \\ &= \mathbf{H} \mathbf{f}_i + \mathbf{u}_i\end{aligned}\tag{11.23}$$

where  $\mathbf{h}_m$  are  $k \times 1$  factor loadings and  $f_{mi}$  is scalar. The second line sets  $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_r]$  and  $\mathbf{f}_i = [f_{1i}, \dots, f_{ri}]'$ . The loadings are normalized so that they are mutually orthonormal. The elements of  $f_{mi}$  and  $\mathbf{u}_i$  are assumed to be mutually uncorrelated, and have variances  $\sigma_m^2$  and  $\sigma_u^2$ , respectively. Assume that the factors are ordered so that  $\sigma_1^2 > \sigma_2^2 > \dots > \sigma_r^2 > 0$ .

The covariance matrix takes the form

$$\begin{aligned}\boldsymbol{\Sigma}_x &= \sum_{m=1}^r \mathbf{h}_m \mathbf{h}'_m \sigma_m^2 + \mathbf{I}_k \sigma_u^2 \\ &= \mathbf{H} \boldsymbol{\Sigma}_f \mathbf{H}' + \mathbf{I}_k \sigma_u^2\end{aligned}$$

where  $\boldsymbol{\Sigma}_f = \text{diag}(\sigma_1^2, \dots, \sigma_r^2)$ . We find that for any of the loadings  $\mathbf{h}_j$

$$\boldsymbol{\Sigma}_x \mathbf{h}_j = \mathbf{h}_j (\sigma_j^2 + \sigma_u^2)$$

and is an eigenvector of  $\boldsymbol{\Sigma}_x$ . We see that the first  $r$  eigenvalues are strictly ranked and have associated eigenvectors  $\mathbf{h}_1, \dots, \mathbf{h}_r$ .

This suggests that the factor loadings can be estimated by taking the eigenvectors  $\widehat{\mathbf{h}}_m$  associated with the largest  $r$  eigenvalues of  $\widehat{\boldsymbol{\Sigma}}_x$ . The estimated proportional contribution of the  $m^{th}$  factor is  $\widehat{\lambda}_m / \sum_{j=1}^r \widehat{\lambda}_j$  where  $\widehat{\lambda}_m$  are the eigenvalues of  $\widehat{\boldsymbol{\Sigma}}_x$ .

In practice the number of factors is unknown. There are a number of rules which have been suggested for selection of  $r$ . Essentially, the key is to examine the eigenvalues of  $\widehat{\boldsymbol{\Sigma}}_x$  and determine if there is a clear cut-off between the “large” and “small” eigenvalues.

To illustrate, we use the Duflo, Dupas and Kremer (2011) dataset. In Table 11.1 we display the seven eigenvalues of the sample covariance matrix for the seven test scores described above. The first eigenvalue is 4.0, and is associated with 57% of the variance. The second eigenvalue is 1.0, and is associated

Table 11.1: Eigenvalue Decomposition of Sample Covariance Matrix

	Eigenvalue	Proportion
1	4.02	0.57
2	1.04	0.15
3	0.57	0.08
4	0.52	0.08
5	0.37	0.05
6	0.29	0.04
7	0.19	0.03

Table 11.2: Factor Loadings

	First Factor	Second Factor
words	0.41	-0.32
sentences	0.32	-0.49
letters	0.40	-0.13
spelling	0.43	-0.28
addition	0.38	0.41
subtraction	0.35	0.52
multiplication	0.33	0.36

with 15% of the variance. The remaining eigenvalues are smaller and of similar magnitude to one another. This is consistent with a two-factor specification. In Table 11.2 we display the factor loadings associated with these two eigenvalues. The coefficients in the first loading are all positive and similar in magnitude. This is consistent with a general “scholastic aptitude” factor which affects all test subjects. The coefficients in the second loading have the interesting pattern that the first four (literacy measures) are all negative and the last three (math measures) are all positive with similar magnitudes. This is consistent with a “mathematics ability” factor which positively affects all mathematics subjects relative to literacy subjects. The presence of a mathematics factor does not mean that all students have a spread between these subjects. Instead, it means that some students do better at the mathematics subjects, some do better at the literacy subjects, and that the mathematics and literacy sub-tests are highly correlated. These results are intuitive and credible.

In Stata, principal components analysis can be implemented using the `pca` command. The command automatically standardizes the variables so this does not need to be done by the researcher.

### 11.13 PCA with Additional Regressors

Consider the model

$$\mathbf{x}_i = \mathbf{H}\mathbf{f}_i + \mathbf{B}'\mathbf{z}_i + \mathbf{u}_i$$

where (as in the previous section)  $\mathbf{x}_i$  and  $\mathbf{u}_i$  are  $k \times 1$ ,  $\mathbf{f}_i$  is  $r \times 1$ , and  $\mathbf{H}$  is  $k \times r$ . In addition there is an  $\ell \times 1$  regressor  $\mathbf{z}_i$  and coefficient matrix  $\mathbf{B}$ .

The coefficients  $\mathbf{H}$ ,  $\mathbf{B}$  and factors  $\mathbf{f}_i$  can be estimated by a combination of PCA and least squares. The key is the following two observations:

- Given  $\mathbf{B}$ , the coefficient  $\mathbf{H}$  and factors  $\mathbf{f}_i$  can be estimated by PCA applied to  $\mathbf{x}_i - \mathbf{B}'\mathbf{z}_i$  as described in the previous section.
- Given the factors  $\mathbf{f}_i$ , the coefficients  $\mathbf{H}$ ,  $\mathbf{B}$  can be estimated by multivariate least squares of  $\mathbf{x}_i$  on  $\mathbf{f}_i$  and  $\mathbf{z}_i$ .

To estimate the parameters all that is needed is to iterate between these two steps. Start with a preliminary estimator of  $\mathbf{B}$  obtained by multivariate least squares of  $\mathbf{x}_i$  on  $\mathbf{z}_i$ . Then apply the above two steps and iterate under convergence.

## 11.14 Factor-Augmented Regression

In Section 11.12 we discussed estimation of single-factor (11.22) and multiple-factor (11.23) principal components models. The factor loadings can be used to estimate the underlying factors, and these can be used for regression analysis.

First, let us consider the problem of estimation of the factors  $f_i$  in a single-factor model. Recall that the observables follow the model  $\mathbf{x}_i = \mathbf{h}f_i + \mathbf{u}_i$ . Suppose we know the factor loading  $\mathbf{h}$ . Then an estimator of  $f_i$  is

$$\tilde{f}_i = \frac{\mathbf{h}'\mathbf{x}_i}{\mathbf{h}'\mathbf{h}} = f_i + v_i$$

where

$$v_i = \frac{\mathbf{h}'\mathbf{u}_i}{\mathbf{h}'\mathbf{h}}.$$

In the previous section we used the normalization  $\mathbf{h}'\mathbf{h} = 1$ . For our current treatment it will be convenient to instead use the normalization  $\sigma_f^2 = 1$ .

Notice that  $v_i$  is mean zero and uncorrelated with  $f_i$ . Thus  $\tilde{f}_i$  is unbiased for  $f_i$ . It also has variance (conditional on  $f_i$ )  $\sigma_u^2/\mathbf{h}'\mathbf{h}$ . We then develop an asymptotic framework under which  $\tilde{f}_i$  is consistent for  $f_i$ . This is the “large  $k$ ” framework where there are a large number of covariates included in the vector  $\mathbf{x}_i$ . If  $k \rightarrow \infty$  such that  $\mathbf{h}'\mathbf{h} \rightarrow \infty$  then  $\text{var}(\tilde{f}_i | f_i) = \sigma_u^2/\mathbf{h}'\mathbf{h} \rightarrow 0$ . The condition  $\mathbf{h}'\mathbf{h} \rightarrow \infty$  means that the typical regressor in  $\mathbf{x}_i$  is related to the factor  $f_i$ , so that as the number of regressors grows the information about  $f_i$  grows as well. We deduce that as  $k \rightarrow \infty$

$$\tilde{f}_i \xrightarrow{P} f_i.$$

This convergence is pointwise in  $i$ .

Now suppose that  $\mathbf{h}$  is not observed but we have the estimator  $\hat{\mathbf{h}}$  from the previous section. Our estimator of  $f_i$  is

$$\hat{f}_i = \frac{\hat{\mathbf{h}}'\mathbf{x}_i}{\hat{\mathbf{h}}'\hat{\mathbf{h}}} = \frac{\hat{\mathbf{h}}'\mathbf{h}}{\hat{\mathbf{h}}'\hat{\mathbf{h}}} f_i + \hat{v}_i$$

where

$$\hat{v}_i = \frac{\hat{\mathbf{h}}'\mathbf{u}_i}{\hat{\mathbf{h}}'\hat{\mathbf{h}}}.$$

As  $n \rightarrow \infty$ ,  $\hat{\Sigma}_x \xrightarrow{P} \Sigma_x$  so by the continuous mapping theorem  $\hat{\mathbf{h}} \xrightarrow{P} \mathbf{h}$ . Hence for each  $i$ ,  $\hat{f}_i \xrightarrow{P} \tilde{f}_i$  as  $n \rightarrow \infty$ . From our previous discussion  $\tilde{f}_i \xrightarrow{P} f_i$  as  $k \rightarrow \infty$  so it stands to reason that  $\hat{f}_i \xrightarrow{P} f_i$  as both  $n$  and  $k$  diverge. This is a bit trickier to establish so we won’t go into the technical details. Still, the idea is that if  $n$  is large then the factor loadings will be well estimated, and if  $k$  is large (with informative regressors) then the factors will be precisely estimated as well.

The above discussion extends naturally to the case of the multiple-factor model (11.23).

Now consider a regression problem. Suppose we have the observations  $(y_i, \mathbf{x}_i)$  where the dimension of  $\mathbf{x}_i$  is large and the elements are highly correlated. Rather than considering the regression of  $y_i$  on  $\mathbf{x}_i$  consider the **factor-augmented regression** model

$$\begin{aligned} y_i &= \mathbf{f}_i'\boldsymbol{\beta} + e_i \\ \mathbf{x}_i &= \mathbf{H}\mathbf{f}_i + \mathbf{u}_i \\ \mathbb{E}(\mathbf{f}_i e_i) &= 0 \\ \mathbb{E}(\mathbf{f}_i \mathbf{u}_i') &= 0 \\ \mathbb{E}(\mathbf{u}_i e_i) &= 0 \end{aligned}$$

This model specifies that the influence of  $\mathbf{x}_i$  on  $y_i$  is through the common factors  $\mathbf{f}_i$ . (There could be additional conventional regressors as well; we omit these from our discussion for simplicity.) The idea is that the variation in the regressors is mostly captured by the variation in the factors, so the influence of the regressors can be mostly captured through these factors. This can be viewed as a dimension-reduction technique, as we have reduced the  $k$ -dimensional  $\mathbf{x}_i$  to the  $r$ -dimensional  $\mathbf{f}_i$ .

In most cases it is difficult to interpret the factors  $\mathbf{f}_i$  and hence the coefficient  $\boldsymbol{\beta}$ . In some cases, though, we can interpret the factors. For example, in the context of the empirical example discussed in the previous section, we could interpret the first two factors as a general “scholastic aptitude” and “math ability”.

The model is typically estimated in multiple steps. First, the factor loadings  $\mathbf{H}$  are estimated from the covariance matrix of  $\mathbf{x}_i$ . Second, the factors  $\mathbf{f}_i$  are estimated as described above. Third,  $y_i$  is regressed on the estimated factors to obtain the estimator of  $\boldsymbol{\beta}$ . The latter takes the form

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left( \sum_{i=1}^n \hat{\mathbf{f}}_i \hat{\mathbf{f}}_i' \right)^{-1} \left( \sum_{i=1}^n \hat{\mathbf{f}}_i y_i \right) \\ &= \left( \hat{\mathbf{H}}' \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \hat{\mathbf{H}} \right)^{-1} \left( \hat{\mathbf{H}}' \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i y_i \right)\end{aligned}$$

where  $\hat{\mathbf{H}}$  are the  $r$  eigenvectors of  $\hat{\Sigma}_x$  associated with the largest  $r$  eigenvalues. As  $n \rightarrow \infty$ ,  $\hat{\mathbf{H}} \xrightarrow{p} \mathbf{H}$  so

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} (\mathbf{H}' \mathbb{E}(\mathbf{x}_i \mathbf{x}_i') \mathbf{H})^{-1} (\mathbf{H}' \mathbb{E}(\mathbf{x}_i y_i)). \quad (11.24)$$

If we use the normalization  $\boldsymbol{\Sigma}_f = \mathbf{I}_r$ , then

$$\mathbb{E}(\mathbf{x}_i \mathbf{x}_i') = \mathbf{H} \mathbf{H}' + \mathbf{I}_k \sigma_u^2$$

and

$$\mathbb{E}(\mathbf{x}_i y_i) = \mathbb{E}((\mathbf{H} \mathbf{f}_i + \mathbf{u}_i)(\mathbf{f}_i' \boldsymbol{\beta} + e_i)) = \mathbf{H} \boldsymbol{\beta}.$$

So the right-hand-side of (11.24) equals

$$(\mathbf{H}' (\mathbf{H} \mathbf{H}' + \mathbf{I}_k \sigma_u^2) \mathbf{H})^{-1} (\mathbf{H}' \mathbf{H} \boldsymbol{\beta}) = (\mathbf{H}' \mathbf{H} + \mathbf{I}_k \sigma_u^2)^{-1} \boldsymbol{\beta}.$$

In the single factor case this is

$$\frac{1}{\mathbf{h}' \mathbf{h} + \sigma_u^2} \boldsymbol{\beta}.$$

As  $k \rightarrow \infty$  we suggested that it is reasonable to assume that  $\mathbf{h}' \mathbf{h} \rightarrow \infty$ . In this case the limit of the above expression is  $\boldsymbol{\beta}$ . Thus the factor-augmented least squares estimator is consistent in the “large  $n$  and  $k$ ” framework.

In the multi-factor case the needed assumption for  $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$  is

$$\lambda_{\min}(\mathbf{H}' \mathbf{H}) \rightarrow \infty.$$

In words, the smallest eigenvalue of the factor loading covariance matrix diverges as  $k \rightarrow \infty$ .

For asymptotic normality we need a stronger rate condition. In the single-factor case the needed condition is that  $n^{-1/2} \mathbf{h}' \mathbf{h} \rightarrow \infty$  as  $n, k \rightarrow \infty$ . In the multi-factor case it is  $n^{-1/2} \lambda_{\min}(\mathbf{H}' \mathbf{H}) \rightarrow \infty$ . These are reasonable conditions as  $\mathbf{h}' \mathbf{h}$  and  $\mathbf{H}' \mathbf{H}$  should grow proportionally to  $k$  if all regressors are similarly related to the factors, and if  $k^2/n \rightarrow \infty$ . The latter is a technical condition, but can be interpreted as meaning that  $k$  is large relative to  $\sqrt{n}$ .

In Stata, the factor estimates  $\hat{\mathbf{f}}_i$  as described above can be calculated by first running principal components analysis (pca) and then using the predict command. This creates the estimated factors which can then be used in a regression command.

## Exercises

**Exercise 11.1** Show (11.10) when the errors are conditionally homoskedastic (11.8).

**Exercise 11.2** Show (11.11) when the regressors are common across equations  $\mathbf{x}_{ji} = \mathbf{x}_i$ .

**Exercise 11.3** Show (11.12) when the regressors are common across equations  $\mathbf{x}_{ji} = \mathbf{x}_i$  and the errors are conditionally homoskedastic (11.8).

**Exercise 11.4** Prove Theorem 11.1.

**Exercise 11.5** Show (11.13) when the regressors are common across equations  $\mathbf{x}_{ji} = \mathbf{x}_i$ .

**Exercise 11.6** Show (11.14) when the regressors are common across equations  $\mathbf{x}_{ji} = \mathbf{x}_i$  and the errors are conditionally homoskedastic (11.8).

**Exercise 11.7** Prove Theorem 11.2.

**Exercise 11.8** Prove Theorem 11.3.

**Exercise 11.9** Show that (11.16) follows from the steps described.

**Exercise 11.10** Show that (11.17) follows from the steps described.

**Exercise 11.11** Prove Theorem 11.4.

**Exercise 11.12** Prove Theorem 11.5.

Hint: First, show that it is sufficient to show that

$$\mathbb{E}(\bar{\mathbf{X}}_i' \bar{\mathbf{X}}_i) \left( \mathbb{E}(\bar{\mathbf{X}}_i' \Sigma^{-1} \bar{\mathbf{X}}_i) \right)^{-1} \mathbb{E}(\bar{\mathbf{X}}_i' \bar{\mathbf{X}}_i) \leq \mathbb{E}(\bar{\mathbf{X}}_i' \Sigma \bar{\mathbf{X}}_i).$$

Second, rewrite this equation using the transformations  $\mathbf{U}_i = \Sigma^{1/2} \bar{\mathbf{X}}_i$  and  $\mathbf{V}_i = \Sigma^{1/2} \bar{\mathbf{X}}_i$ , and then apply the matrix Cauchy-Schwarz inequality (B.32).

**Exercise 11.13** Prove Theorem 11.6.

**Exercise 11.14** Take the model

$$\begin{aligned} y_i &= \boldsymbol{\pi}'_i \boldsymbol{\beta} + e_i \\ \boldsymbol{\pi}_i &= \mathbb{E}(\mathbf{x}_i | \mathbf{z}_i) = \boldsymbol{\Gamma}' \mathbf{z}_i \\ \mathbb{E}(e_i | \mathbf{z}_i) &= 0 \end{aligned}$$

where  $y_i$ ,  $i$  is scalar,  $\mathbf{x}_i$  is a  $k$  vector and  $\mathbf{z}_i$  is an  $\ell$  vector.  $\boldsymbol{\beta}$  and  $\boldsymbol{\pi}_i$  are  $k \times 1$  and  $\boldsymbol{\Gamma}$  is  $\ell \times k$ . The sample is  $(y_i, \mathbf{x}_i, \mathbf{z}_i : i = 1, \dots, n)$  with  $\boldsymbol{\pi}_i$  unobserved.

Consider the estimator  $\hat{\boldsymbol{\beta}}$  for  $\boldsymbol{\beta}$  by OLS of  $y_i$  on  $\hat{\boldsymbol{\pi}}_i = \hat{\boldsymbol{\Gamma}}' \mathbf{z}_i$  where  $\hat{\boldsymbol{\Gamma}}$  is the OLS coefficient from the multivariate regression of  $\mathbf{x}_i$  on  $\mathbf{z}_i$

- (a) Show that  $\hat{\boldsymbol{\beta}}$  is consistent for  $\boldsymbol{\beta}$ .
- (b) Find the asymptotic distribution  $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$  as  $n \rightarrow \infty$  assuming that  $\boldsymbol{\beta} = \mathbf{0}$ .
- (c) Why is the assumption  $\boldsymbol{\beta} = \mathbf{0}$  an important simplifying condition in part (b)?
- (d) Using the result in (c), construct an appropriate asymptotic test for the hypothesis  $H_0 : \boldsymbol{\beta} = \mathbf{0}$ .

**Exercise 11.15** The observations are i.i.d.,  $(y_{1i}, y_{2i}, \mathbf{x}_i : i = 1, \dots, n)$ . The dependent variables  $y_{1i}$  and  $y_{2i}$  are real-valued. The regressor  $\mathbf{x}_i$  is a  $k$ -vector. The model is the two-equation system

$$\begin{aligned} y_{1i} &= \mathbf{x}'_i \boldsymbol{\beta}_1 + e_{1i} \\ \mathbb{E}(\mathbf{x}_i e_{1i}) &= 0 \\ y_{2i} &= \mathbf{x}'_i \boldsymbol{\beta}_2 + e_{2i} \\ \mathbb{E}(\mathbf{x}_i e_{2i}) &= 0. \end{aligned}$$

- (a) What are the appropriate estimators  $\hat{\boldsymbol{\beta}}_1$  and  $\hat{\boldsymbol{\beta}}_2$  for  $\boldsymbol{\beta}_1$  and  $\boldsymbol{\beta}_2$ ?
- (b) Find the joint asymptotic distribution of  $\hat{\boldsymbol{\beta}}_1$  and  $\hat{\boldsymbol{\beta}}_2$ .
- (c) Describe a test for  $H_0: \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ .

# Chapter 12

## Instrumental Variables

### 12.1 Introduction

The concepts of **endogeneity** and **instrumental variable** are fundamental to econometrics, and mark a substantial departure from other branches of statistics. The ideas of endogeneity arise naturally in economics from models of simultaneous equations, most notably the classic supply/demand model of price determination.

The identification problem in simultaneous equations dates back to Philip Wright (1915) and Working (1927). The method of instrumental variables first appears in an Appendix of a 1928 book by Philip Wright, though the authorship is sometimes credited to his son Sewell Wright. The label “instrumental variables” was introduced by Reiersøl (1945). An excellent review of the history of instrumental variables and this controversy is Stock and Trebbi (2003).

### 12.2 Overview

We say that there is **endogeneity** in the linear model

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + e_i \quad (12.1)$$

if  $\boldsymbol{\beta}$  is the parameter of interest and

$$\mathbb{E}(\mathbf{x}_i e_i) \neq \mathbf{0}. \quad (12.2)$$

This is a core problem in econometrics and largely differentiates econometrics from many branches of statistics. To distinguish (12.1) from the regression and projection models, we will call (12.1) a **structural equation** and  $\boldsymbol{\beta}$  a **structural parameter**. When (12.2) holds, it is typical to say that  $\mathbf{x}_i$  is **endogenous** for  $\boldsymbol{\beta}$ .

Endogeneity cannot happen if the coefficient is defined by linear projection. Indeed, we can define the linear projection coefficient  $\boldsymbol{\beta}^* = \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i)^{-1} \mathbb{E}(\mathbf{x}_i y_i)$  and linear projection equation

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta}^* + e_i^* \\ \mathbb{E}(\mathbf{x}_i e_i^*) &= \mathbf{0}. \end{aligned}$$

However, under endogeneity (12.2) the projection coefficient  $\boldsymbol{\beta}^*$  does not equal the structural parameter. Indeed,

$$\begin{aligned} \boldsymbol{\beta}^* &= (\mathbb{E}(\mathbf{x}_i \mathbf{x}'_i))^{-1} \mathbb{E}(\mathbf{x}_i y_i) \\ &= (\mathbb{E}(\mathbf{x}_i \mathbf{x}'_i))^{-1} \mathbb{E}(\mathbf{x}_i (\mathbf{x}'_i \boldsymbol{\beta} + e_i)) \\ &= \boldsymbol{\beta} + (\mathbb{E}(\mathbf{x}_i \mathbf{x}'_i))^{-1} \mathbb{E}(\mathbf{x}_i e_i) \\ &\neq \boldsymbol{\beta} \end{aligned}$$

the final relation since  $\mathbb{E}(\mathbf{x}_i e_i) \neq \mathbf{0}$ .

Thus endogeneity requires that the coefficient be defined differently than projection. We describe such definitions as **structural**. We will present three examples in the following section.

Endogeneity implies that the least-squares estimator is inconsistent for the structural parameter. Indeed, under i.i.d. sampling, least-squares is consistent for the projection coefficient, and thus is inconsistent for  $\beta$ .

$$\hat{\beta} \xrightarrow{P} (\mathbb{E}(\mathbf{x}_i \mathbf{x}'_i))^{-1} \mathbb{E}(\mathbf{x}_i y_i) = \beta^* \neq \beta.$$

The inconsistency of least-squares is typically referred to as **endogeneity bias** or **estimation bias** due to endogeneity. (This is an imperfect label as the actual issue is inconsistency, not bias.)

As the structural parameter  $\beta$  is the parameter of interest, endogeneity requires the development of alternative estimation methods. We discuss those in later sections.

## 12.3 Examples

The concept of endogeneity may be easiest to understand by example. We discuss three distinct examples. In each case it is important to see how the structural parameter  $\beta$  is defined independently from the linear projection model.

**Example: Measurement error in the regressor.** Suppose that  $(y_i, z_i)$  are joint random variables,  $\mathbb{E}(y_i | z_i) = z'_i \beta$  is linear,  $\beta$  is the structural parameter, and  $z_i$  is not observed. Instead we observe  $\mathbf{x}_i = z_i + \mathbf{u}_i$  where  $\mathbf{u}_i$  is a  $k \times 1$  measurement error, independent of  $e_i$  and  $z_i$ . This is an example of a latent variable model, where “latent” refers to a structural variable which is unobserved.

The model  $\mathbf{x}_i = z_i + \mathbf{u}_i$  with  $z_i$  and  $\mathbf{u}_i$  independent and  $\mathbb{E}(\mathbf{u}_i) = \mathbf{0}$  is known as **classical measurement error**. This means that  $\mathbf{x}_i$  is a noisy but unbiased measure of  $z_i$ .

By substitution we can express  $y_i$  as a function of the observed variable  $\mathbf{x}_i$ .

$$\begin{aligned} y_i &= z'_i \beta + e_i \\ &= (\mathbf{x}_i - \mathbf{u}_i)' \beta + e_i \\ &= \mathbf{x}'_i \beta + v_i \end{aligned}$$

where  $v_i = e_i - \mathbf{u}'_i \beta$ . This means that  $(y_i, \mathbf{x}_i)$  satisfy the linear equation

$$y_i = \mathbf{x}'_i \beta + v_i$$

with an error  $v_i$ . But this error is not a projection error. Indeed,

$$\mathbb{E}(\mathbf{x}_i v_i) = \mathbb{E}[(\mathbf{x}_i - \mathbf{u}_i)(e_i - \mathbf{u}'_i \beta)] = -\mathbb{E}(\mathbf{u}_i \mathbf{u}'_i) \beta \neq \mathbf{0}$$

if  $\beta \neq 0$  and  $\mathbb{E}(\mathbf{u}_i \mathbf{u}'_i) \neq 0$ . As we learned in the previous section, if  $\mathbb{E}(\mathbf{x}_i v_i) \neq 0$  then least-squares estimation will be inconsistent.

We can calculate the form of the projection coefficient (which is consistently estimated by least-squares). For simplicity suppose that  $k = 1$ . We find

$$\beta^* = \beta + \frac{\mathbb{E}(x_i v_i)}{\mathbb{E}(x_i^2)} = \beta \left( 1 - \frac{\mathbb{E}(u_i^2)}{\mathbb{E}(x_i^2)} \right).$$

Since  $\mathbb{E}(u_i^2)/\mathbb{E}(x_i^2) < 1$  the projection coefficient shrinks the structural parameter  $\beta$  towards zero. This is called **measurement error bias** or **attenuation bias**.

**Example: Supply and Demand.** The variables  $q_i$  and  $p_i$  (quantity and price) are determined jointly by the demand equation

$$q_i = -\beta_1 p_i + e_{1i}$$

and the supply equation

$$q_i = \beta_2 p_i + e_{2i}.$$

Assume that  $\mathbf{e}_i = \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix}$  is i.i.d.,  $\mathbb{E}(\mathbf{e}_i) = \mathbf{0}$  and  $\mathbb{E}(\mathbf{e}_i \mathbf{e}'_i) = \mathbf{I}_2$  (the latter for simplicity). The question is: if we regress  $q_i$  on  $p_i$ , what happens?

It is helpful to solve for  $q_i$  and  $p_i$  in terms of the errors. In matrix notation,

$$\begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix} \begin{pmatrix} q_i \\ p_i \end{pmatrix} = \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix}$$

so

$$\begin{aligned} \begin{pmatrix} q_i \\ p_i \end{pmatrix} &= \begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix}^{-1} \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \\ &= \begin{bmatrix} \beta_2 & \beta_1 \\ 1 & -1 \end{bmatrix} \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \left( \frac{1}{\beta_1 + \beta_2} \right) \\ &= \begin{pmatrix} (\beta_2 e_{1i} + \beta_1 e_{2i}) / (\beta_1 + \beta_2) \\ (e_{1i} - e_{2i}) / (\beta_1 + \beta_2) \end{pmatrix}. \end{aligned}$$

The projection of  $q_i$  on  $p_i$  yields

$$\begin{aligned} q_i &= \beta^* p_i + e_i^* \\ \mathbb{E}(p_i e_i^*) &= 0 \end{aligned}$$

where

$$\beta^* = \frac{\mathbb{E}(p_i q_i)}{\mathbb{E}(p_i^2)} = \frac{\beta_2 - \beta_1}{2}.$$

Thus the projection coefficient  $\beta^*$  equals neither the demand slope  $\beta_1$  nor the supply slope  $\beta_2$ , but equals an average of the two. (The fact that it is a simple average is an artifact of the simple covariance structure.)

Hence the OLS estimate satisfies  $\hat{\beta} \xrightarrow{P} \beta^*$ , and the limit does not equal either  $\beta_1$  or  $\beta_2$ . The fact that the limit is neither the supply nor demand slope is called **simultaneous equations bias**. This occurs generally when  $y_i$  and  $x_i$  are jointly determined, as in a market equilibrium.

Generally, when both the dependent variable and a regressor are simultaneously determined, then the variables should be treated as endogenous.

**Example: Choice Variables as Regressors.** Take the classic wage equation

$$\log(wage) = \beta education + e$$

with  $\beta$  the average causal effect of education on wages. If wages are affected by unobserved ability, and individuals with high ability self-select into higher education, then  $e$  contains unobserved ability, so *education* and  $e$  will be positively correlated. Hence *education* is endogenous. The positive correlation means that the linear projection coefficient  $\beta^*$  will be upward biased relative to the structural coefficient  $\beta$ . Thus least-squares (which is estimating the projection coefficient) will tend to over-estimate the causal effect of education on wages.

This type of endogeneity occurs generally when  $y$  and  $x$  are both choices made by an economic agent, even if they are made at different points in time.

Generally, when both the dependent variable and a regressor are choice variables made by the same agent, the variables should be treated as endogenous.

## 12.4 Instruments

We have defined endogeneity as the context where the regressor is correlated with the equation error. In most applications we only treat a subset of the regressors as endogenous; most of the regressors will be treated as **exogenous**, meaning that they are assumed uncorrelated with the equation error. To be specific, we make the partition

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_{1i} \\ \mathbf{x}_{2i} \end{pmatrix} \begin{matrix} k_1 \\ k_2 \end{matrix} \quad (12.3)$$

and similarly

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \begin{matrix} k_1 \\ k_2 \end{matrix}$$

so that the **structural equation** is

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ &= \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + e_i. \end{aligned} \quad (12.4)$$

The regressors are assumed to satisfy

$$\begin{aligned} \mathbb{E}(\mathbf{x}_{1i} e_i) &= \mathbf{0} \\ \mathbb{E}(\mathbf{x}_{2i} e_i) &\neq \mathbf{0}. \end{aligned}$$

We call  $\mathbf{x}_{1i}$  **exogenous** and  $\mathbf{x}_{2i}$  **endogenous** for the structural parameter  $\boldsymbol{\beta}$ . As the dependent variable  $y_i$  is also endogenous, we sometimes differentiate  $\mathbf{x}_{2i}$  by calling  $\mathbf{x}_{2i}$  the **endogenous right-hand-side variables**.

In matrix notation we can write (12.4) as

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \\ &= \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \mathbf{e}. \end{aligned}$$

The endogenous regressors  $\mathbf{x}_{2i}$  are the critical variables discussed in the examples of the previous section – simultaneous variables, choice variables, mis-measured regressors – that are potentially correlated with the equation error  $e_i$ . In most applications the number  $k_2$  of variables treated as endogenous is small (1 or 2). The exogenous variables  $\mathbf{x}_{1i}$  are the remaining regressors (including the equation intercept) and can be low or high dimensional.

To consistently estimate  $\boldsymbol{\beta}$  we require additional information. One type of information which is commonly used in economic applications are what we call **instruments**.

**Definition 12.1** The  $\ell \times 1$  random vector  $\mathbf{z}_i$  is an **instrumental variable** for (12.4) if

$$\mathbb{E}(\mathbf{z}_i e_i) = \mathbf{0} \quad (12.5)$$

$$\mathbb{E}(\mathbf{z}_i \mathbf{z}'_i) > 0 \quad (12.6)$$

$$\text{rank}(\mathbb{E}(\mathbf{z}_i \mathbf{x}'_i)) = k. \quad (12.7)$$

There are three components to the definition as given. The first (12.5) is that the instruments are uncorrelated with the regression error. The second (12.6) is a normalization which excludes linearly redundant instruments. The third (12.7) is often called the **relevance condition** and is essential for the identification of the model, as we discuss later. A necessary condition for (12.7) is that  $\ell \geq k$ .

Condition (12.5) – that the instruments are uncorrelated with the equation error, is often described as that they are **exogenous** in the sense that they are determined outside the model for  $y_i$ .

Notice that the regressors  $\mathbf{x}_{1i}$  satisfy condition (12.5) and thus should be included as instrumental variables. It is thus a subset of the variables  $\mathbf{z}_i$ . Notationally we make the partition

$$\mathbf{z}_i = \begin{pmatrix} \mathbf{z}_{1i} \\ \mathbf{z}_{2i} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_{1i} \\ \mathbf{z}_{2i} \end{pmatrix} \begin{matrix} k_1 \\ \ell_2 \end{matrix}. \quad (12.8)$$

Here,  $\mathbf{x}_{1i} = \mathbf{z}_{1i}$  are the **included exogenous variables**, and  $\mathbf{z}_{2i}$  are the **excluded exogenous variables**. That is,  $\mathbf{z}_{2i}$  are variables which could be included in the equation for  $y_i$  (in the sense that they are uncorrelated with  $e_i$ ) yet can be excluded, as they would have true zero coefficients in the equation.

Many authors simply label  $\mathbf{x}_{1i}$  as the “exogenous variables”,  $\mathbf{x}_{2i}$  as the “endogenous variables”, and  $\mathbf{z}_{2i}$  as the “instrumental variables”.

We say that the model is **just-identified** if  $\ell = k$  (and  $\ell_2 = k_2$ ) and **over-identified** if  $\ell > k$  (and  $\ell_2 > k_2$ ).

What variables can be used as instrumental variables? From the definition  $\mathbb{E}(\mathbf{z}_i e_i) = \mathbf{0}$  we see that the instrument must be uncorrelated with the equation error, meaning that it is excluded from the structural equation as mentioned above. From the rank condition (12.7) it is also important that the instrumental variable be correlated with the endogenous variables  $\mathbf{x}_{2i}$  after controlling for the other exogenous variables  $\mathbf{x}_{1i}$ . These two requirements are typically interpreted as requiring that the instruments be determined outside the system for  $(y_i, \mathbf{x}_{2i})$ , causally determine  $\mathbf{x}_{2i}$ , but do not causally determine  $y_i$  except through  $\mathbf{x}_{2i}$ .

Let's take the three examples given above.

**Measurement error in the regressor.** When  $\mathbf{x}_i$  is a mis-measured version of  $\mathbf{z}_i$ , a common choice for an instrument  $\mathbf{z}_{2i}$  is an alternative measurement of  $\mathbf{z}_i$ . For this  $\mathbf{z}_{2i}$  to satisfy the property of an instrumental variable the measurement error in  $\mathbf{z}_{2i}$  must be independent of that in  $\mathbf{x}_i$ .

**Supply and Demand.** An appropriate instrument for price  $p_i$  in a demand equation is a variable  $z_{2i}$  which influences supply but not demand. Such a variable affects the equilibrium values of  $p_i$  and  $q_i$  but does not directly affect price except through quantity. Variables which affect supply but not demand are typically related to production costs.

An appropriate instrument for price in a supply equation is a variable which influences demand but not supply. Such a variable affects the equilibrium values of price and quantity but only affects price through quantity.

**Choice Variable as Regressor.** An ideal instrument affects the choice of the regressor (education) but does not directly influence the dependent variable (wages) except through the indirect effect on the regressor. We will discuss an example in the next section.

## 12.5 Example: College Proximity

In a influential paper, David Card (1995) suggested if a potential student lives close to a college, this reduces the cost of attendance and thereby raises the likelihood that the student will attend college. However, college proximity does not directly affect a student's skills or abilities, so should not have a direct effect on his or her market wage. These considerations suggest that college proximity can be used as an instrument for education in a wage regression. We use the simplest model reported in Card's paper to illustrate the concepts of instrumental variables throughout the chapter.

Card used data from the National Longitudinal Survey of Young Men (NLSYM) for 1976. A baseline least-squares wage regression for his data set is reported in the first column of Table 12.1. The dependent variable is the log of weekly earnings. The regressors are *education* (years of schooling), *experience* (years of work experience, calculated as *age* (years) less *education*+6), *experience*<sup>2</sup>/100, *black*, *south* (an indicator for residence in the southern region of the U.S.), and *urban* (an indicator for residence in a standard metropolitan statistical area). We drop observations for which *wage* is missing. The remaining sample has 3,010 observations. His data is the file Card1995 on the textbook website.

The point estimate obtained by least-squares suggests an 8% increase in earnings for each year of education.

Table 12.1: Instrumental Variable Wage Regressions

	OLS	IV(a)	IV(b)	2SLS(a)	2SLS(b)	LIML
education	0.074 (0.004)	0.132 (0.049)	0.133 (0.051)	0.161 (0.040)	0.160 (0.041)	0.164 (0.042)
experience	0.084 (0.007)	0.107 (0.021)	0.056 (0.026)	0.119 (0.018)	0.047 (0.025)	0.120 (0.019)
experience <sup>2</sup> /100	-0.224 (0.032)	-0.228 (0.035)	-0.080 (0.133)	-0.231 (0.037)	-0.032 (0.127)	-0.231 (0.037)
black	-0.190 (0.017)	-0.131 (0.051)	-0.103 (0.075)	-0.102 (0.044)	-0.064 (0.061)	-0.099 (0.045)
south	-0.125 (0.015)	-0.105 (0.023)	-0.098 (0.0284)	-0.095 (0.022)	-0.086 (0.026)	-0.094 (0.022)
urban	0.161 (0.015)	0.131 (0.030)	0.108 (0.049)	0.116 (0.026)	0.083 (0.041)	0.115 (0.027)
Sargan				0.82	0.52	0.82
p-value				0.37	0.47	0.37

Notes:

1. IV(a) uses *college* as an instrument for *education*.
2. IV(b) uses *college*, *age*, and *age*<sup>2</sup>/100 as instruments for *education*, *experience*, and *experience*<sup>2</sup>/100.
3. 2SLS(a) uses *public* and *private* as instruments for *education*.
4. 2SLS(b) uses *public*, *private*, *age*, and *age*<sup>2</sup> as instruments for *education*, *experience*, and *experience*<sup>2</sup>/100.
5. LIML uses *public* and *private* as instruments for *education*.

As discussed in the previous sections, it is reasonable to view years of education as a choice made by an individual, and thus is likely endogenous for the structural return to education. This means that least-squares is an estimate of a linear projection, but is inconsistent for coefficient of a structural equation representing the causal impact of years of education on expected wages. Labor economics predicts that ability, education, and wages will be positively correlated. This suggests that the population projection coefficient estimated by least-squares will be higher than the structural parameter (and hence upwards biased). However, the sign of the bias is uncertain since there are multiple regressors and there are other potential sources of endogeneity.

To instrument for the endogeneity of education, Card suggested that a reasonable instrument is a dummy variable indicating if the individual grew up near a college. We will consider three measures:

- college* Grew up in same county as a 4-year college
- public* Grew up in same county as a 4-year public college
- private* Grew up in same county as a 4-year private college.

## 12.6 Reduced Form

The reduced form is the relationship between the regressors  $\mathbf{x}_i$  and the instruments  $\mathbf{z}_i$ . A linear reduced form model for  $\mathbf{x}_i$  is

$$\mathbf{x}_i = \boldsymbol{\Gamma}' \mathbf{z}_i + \mathbf{u}_i. \quad (12.9)$$

This is a multivariate regression as introduced in Chapter 11. The  $\ell \times k$  coefficient matrix  $\Gamma$  can be defined by linear projection. Thus

$$\Gamma = \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i)^{-1} \mathbb{E}(\mathbf{z}_i \mathbf{x}'_i) \quad (12.10)$$

so that

$$\mathbb{E}(\mathbf{z}_i \mathbf{u}'_i) = \mathbf{0}.$$

In matrix notation, we can write (12.9) as

$$\mathbf{X} = \mathbf{Z}\Gamma + \mathbf{U}$$

where  $\mathbf{U}$  is  $n \times k$ . Notice that the projection coefficient (12.10) is well defined and unique under (12.6).

Since  $\mathbf{z}_i$  and  $\mathbf{x}_i$  have the common variables  $\mathbf{x}_{1i}$ , we can focus on the reduced form for the endogenous regressors  $\mathbf{x}_{2i}$ . Recalling the partitions (12.3) and (12.8) we can partition  $\Gamma$  conformably as

$$\begin{aligned} \Gamma &= \begin{bmatrix} k_1 & k_2 \\ \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} \quad \ell_1 \\ &= \begin{bmatrix} \mathbf{I}_{k_1} & \Gamma_{12} \\ \mathbf{0} & \Gamma_{22} \end{bmatrix} \end{aligned} \quad (12.11)$$

and similarly partition  $\mathbf{u}_i$ . Then (12.9) can be rewritten as two equation systems

$$\mathbf{x}_{1i} = \mathbf{z}_{1i} \quad (12.12)$$

$$\mathbf{x}_{2i} = \Gamma'_{12} \mathbf{z}_{1i} + \Gamma'_{22} \mathbf{z}_{2i} + \mathbf{u}_{2i}. \quad (12.13)$$

The first equation (12.12) is a tautology. The second equation (12.13) is the primary reduced form equation of interest. It is a multivariate linear regression for  $\mathbf{x}_{2i}$  as a function of the included and excluded exogenous variables  $\mathbf{z}_{1i}$  and  $\mathbf{z}_{2i}$ .

We can also construct a reduced form equation for  $y_i$ . Substituting (12.9) into (12.4), we find

$$\begin{aligned} y_i &= (\Gamma' \mathbf{z}_i + \mathbf{u}_i)' \boldsymbol{\beta} + e_i \\ &= \mathbf{z}'_i \boldsymbol{\lambda} + v_i \end{aligned} \quad (12.14)$$

where

$$\boldsymbol{\lambda} = \Gamma \boldsymbol{\beta} \quad (12.15)$$

and

$$v_i = \mathbf{u}'_i \boldsymbol{\beta} + e_i.$$

Observe that

$$\mathbb{E}(\mathbf{z}_i v_i) = \mathbb{E}(\mathbf{z}_i \mathbf{u}'_i) \boldsymbol{\beta} + \mathbb{E}(\mathbf{z}_i e_i) = \mathbf{0}.$$

Thus (12.14) is a projection equation. It is the reduced form for  $y_i$ , as it expresses  $y_i$  as a function of exogenous variables only. Since it is a projection equation we can write the reduced form coefficient as

$$\boldsymbol{\lambda} = \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i)^{-1} \mathbb{E}(\mathbf{z}_i y_i)$$

which is well defined under (12.6).

Alternatively, we can substitute (12.13) into (12.4) and use  $\mathbf{x}_{1i} = \mathbf{z}_{1i}$  to obtain

$$\begin{aligned} y_i &= \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + (\Gamma'_{12} \mathbf{z}_{1i} + \Gamma'_{22} \mathbf{z}_{2i} + \mathbf{u}_{2i})' \boldsymbol{\beta}_2 + e_i \\ &= \mathbf{z}'_{1i} \boldsymbol{\lambda}_1 + \mathbf{z}'_{2i} \boldsymbol{\lambda}_2 + v_i \end{aligned} \quad (12.16)$$

where

$$\boldsymbol{\lambda}_1 = \boldsymbol{\beta}_1 + \Gamma_{12} \boldsymbol{\beta}_2 \quad (12.17)$$

$$\boldsymbol{\lambda}_2 = \Gamma_{22} \boldsymbol{\beta}_2. \quad (12.18)$$

which is an alternative (and equivalent) expression of (12.15) given (12.11).

(12.9) and (12.14) together (or (12.13) and (12.16) together) are the **reduced form equations** for the system

$$\begin{aligned} y_i &= \mathbf{z}'_i \boldsymbol{\lambda} + v_i \\ \mathbf{x}_i &= \boldsymbol{\Gamma}' \mathbf{z}_i + \mathbf{u}_i. \end{aligned}$$

The relationships (12.15) and (12.17)-(12.18) are critically important for understanding the identification of the structural parameters  $\boldsymbol{\beta}_1$  and  $\boldsymbol{\beta}_2$ , as we discuss below. These equations show the tight relationship between the parameters of the structural equations ( $\boldsymbol{\beta}_1$  and  $\boldsymbol{\beta}_2$ ) and those of the reduced form equations ( $\boldsymbol{\lambda}_1$ ,  $\boldsymbol{\lambda}_2$ ,  $\boldsymbol{\Gamma}_{12}$  and  $\boldsymbol{\Gamma}_{22}$ ).

## 12.7 Reduced Form Estimation

The reduced form equations are projections, so the coefficient matrices may be estimated by least-squares (see Chapter 11). The least-squares estimate of (12.9) is

$$\hat{\boldsymbol{\Gamma}} = \left( \sum_{i=1}^n \mathbf{z}_i \mathbf{z}'_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{z}_i \mathbf{x}'_i \right). \quad (12.19)$$

The estimates of equation (12.9) can be written as

$$\mathbf{x}_i = \hat{\boldsymbol{\Gamma}}' \mathbf{z}_i + \hat{\mathbf{u}}_i. \quad (12.20)$$

In matrix notation, these can be written as

$$\hat{\boldsymbol{\Gamma}} = (\mathbf{Z}' \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{X})$$

and

$$\mathbf{X} = \mathbf{Z} \hat{\boldsymbol{\Gamma}} + \hat{\mathbf{U}}.$$

Since  $\mathbf{X}$  and  $\mathbf{Z}$  have a common sub-matrix, we have the partition

$$\hat{\boldsymbol{\Gamma}} = \begin{bmatrix} \mathbf{I}_{k_1} & \hat{\boldsymbol{\Gamma}}_{12} \\ \mathbf{0} & \hat{\boldsymbol{\Gamma}}_{22} \end{bmatrix}.$$

The reduced form estimates of equation (12.13) can be written as

$$\mathbf{x}_{2i} = \hat{\boldsymbol{\Gamma}}_{12}' \mathbf{z}_{1i} + \hat{\boldsymbol{\Gamma}}_{22}' \mathbf{z}_{2i} + \hat{\mathbf{u}}_{2i}$$

or in matrix notation as

$$\mathbf{X}_2 = \mathbf{Z}_1 \hat{\boldsymbol{\Gamma}}_{12} + \mathbf{Z}_2 \hat{\boldsymbol{\Gamma}}_{22} + \hat{\mathbf{U}}_2.$$

We can write the submatrix estimates as

$$\begin{bmatrix} \hat{\boldsymbol{\Gamma}}_{12} \\ \hat{\boldsymbol{\Gamma}}_{22} \end{bmatrix} = \left( \sum_{i=1}^n \mathbf{z}_i \mathbf{z}'_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{z}_i \mathbf{x}'_{2i} \right) = (\mathbf{Z}' \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{X}_2).$$

The reduced form estimate of equation (12.14) is

$$\begin{aligned} \hat{\boldsymbol{\lambda}} &= \left( \sum_{i=1}^n \mathbf{z}_i \mathbf{z}'_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{z}_i y_i \right) \\ y_i &= \mathbf{z}'_i \hat{\boldsymbol{\lambda}} + \hat{v}_i \\ &= \mathbf{z}'_{1i} \hat{\boldsymbol{\lambda}}_1 + \mathbf{z}'_{2i} \hat{\boldsymbol{\lambda}}_2 + \hat{v}_i \end{aligned}$$

or in matrix notation

$$\begin{aligned} \hat{\boldsymbol{\lambda}} &= (\mathbf{Z}' \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{y}) \\ \mathbf{y} &= \mathbf{Z} \hat{\boldsymbol{\lambda}} + \hat{\mathbf{v}} \\ &= \mathbf{Z}_1 \hat{\boldsymbol{\lambda}}_1 + \mathbf{Z}_2 \hat{\boldsymbol{\lambda}}_2 + \hat{\mathbf{v}}. \end{aligned}$$

## 12.8 Identification

A parameter is **identified** if it is a unique function of the probability distribution of the observables. One way to show that a parameter is identified is to write it as an explicit function of population moments. For example, the reduced form coefficient matrices  $\Gamma$  and  $\lambda$  are identified since they can be written as explicit functions of the moments of the observables  $(y_i, \mathbf{x}_i, \mathbf{z}_i)$ . That is,

$$\Gamma = \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i)^{-1} \mathbb{E}(\mathbf{z}_i \mathbf{x}'_i) \quad (12.21)$$

$$\lambda = \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i)^{-1} \mathbb{E}(\mathbf{z}_i y_i). \quad (12.22)$$

These are uniquely determined by the probability distribution of  $(y_i, \mathbf{x}_i, \mathbf{z}_i)$  if Definition 12.1 holds, since this includes the requirement that  $\mathbb{E}(\mathbf{z}_i \mathbf{z}'_i)$  is invertible.

We are interested in the structural parameter  $\beta$ . It relates to  $(\lambda, \Gamma)$  through (12.15), or

$$\lambda = \Gamma \beta. \quad (12.23)$$

It is identified if it uniquely determined by this relation. This is a set of  $\ell$  equations with  $k$  unknowns with  $\ell \geq k$ . From standard linear algebra we know that there is a unique solution if and only if  $\Gamma$  has full rank  $k$ .

$$\text{rank}(\Gamma) = k. \quad (12.24)$$

Under (12.24),  $\beta$  can be uniquely solved from the linear system  $\lambda = \Gamma \beta$ . On the other hand if  $\text{rank}(\Gamma) < k$  then  $\lambda = \Gamma \beta$  has fewer mutually independent linear equations than coefficients so there is not a unique solution.

From the definitions (12.21)-(12.22) the identification equation (12.23) is the same as

$$\mathbb{E}(\mathbf{z}_i y_i) = \mathbb{E}(\mathbf{z}_i \mathbf{x}'_i) \beta$$

which is again a set of  $\ell$  equations with  $k$  unknowns. This has a unique solution if (and only if)

$$\text{rank}(\mathbb{E}(\mathbf{z}_i \mathbf{x}'_i)) = k \quad (12.25)$$

which was listed in (12.7) as a conditions of Definition 12.1. (Indeed, this is why it was listed as part of the definition.) We can also see that (12.24) and (12.25) are equivalent ways of expressing the same requirement. If this condition fails then  $\beta$  will not be identified. The condition (12.24)-(12.25) is called the **relevance condition**.

It is useful to have explicit expressions for the solution  $\beta$ . The easiest case is when  $\ell = k$ . Then (12.24) implies  $\Gamma$  is invertible, so the structural parameter equals  $\beta = \Gamma^{-1} \lambda$ . It is a unique solution because  $\Gamma$  and  $\lambda$  are unique and  $\Gamma$  is invertible.

When  $\ell > k$  we can solve for  $\beta$  by applying least-squares to the system of equations  $\lambda = \Gamma \beta$ . This is  $\ell$  equations with  $k$  unknowns and no error. The least-squares solution is  $\beta = (\Gamma' \Gamma)^{-1} \Gamma' \lambda$ . Under (12.24) the matrix  $\Gamma' \Gamma$  is invertible so the solution is unique.

$\beta$  is identified if  $\text{rank}(\Gamma) = k$ , which is true if and only if  $\text{rank}(\Gamma_{22}) = k_2$  (by the upper-diagonal structure of  $\Gamma$ ). Thus the key to identification of the model rests on the  $\ell_2 \times k_2$  matrix  $\Gamma_{22}$  in (12.13). To see this, recall the reduced form relationships (12.17)-(12.18). We can see that  $\beta_2$  is identified from (12.18) alone, and the necessary and sufficient condition is  $\text{rank}(\Gamma_{22}) = k_2$ . If this is satisfied then the solution can be written as  $\beta_2 = (\Gamma'_{22} \Gamma_{22})^{-1} \Gamma'_{22} \lambda_2$ . Then  $\beta_1$  is identified from this and (12.17), with the explicit solution  $\beta_1 = \lambda_1 - \Gamma_{12} (\Gamma'_{22} \Gamma_{22})^{-1} \Gamma'_{22} \lambda_2$ . In the just-identified case ( $\ell_2 = k_2$ ) these equations simplify to take the form  $\beta_2 = \Gamma_{22}^{-1} \lambda_2$  and  $\beta_1 = \lambda_1 - \Gamma_{12} \Gamma_{22}^{-1} \lambda_2$ .

## 12.9 Instrumental Variables Estimator

In this section we consider the special case where the model is just-identified, so that  $\ell = k$ .

The assumption that  $\mathbf{z}_i$  is an instrumental variable implies that

$$\mathbb{E}(\mathbf{z}_i e_i) = \mathbf{0}.$$

Making the substitution  $e_i = y_i - \mathbf{x}'_i \boldsymbol{\beta}$  we find

$$\mathbb{E}(\mathbf{z}_i (y_i - \mathbf{x}'_i \boldsymbol{\beta})) = \mathbf{0}.$$

Expanding,

$$\mathbb{E}(\mathbf{z}_i y_i) - \mathbb{E}(\mathbf{z}_i \mathbf{x}'_i) \boldsymbol{\beta} = \mathbf{0}.$$

This is a system of  $\ell = k$  equations and  $k$  unknowns. Solving for  $\boldsymbol{\beta}$  we find

$$\boldsymbol{\beta} = (\mathbb{E}(\mathbf{z}_i \mathbf{x}'_i))^{-1} \mathbb{E}(\mathbf{z}_i y_i).$$

This solution assumes that the matrix  $\mathbb{E}(\mathbf{z}_i \mathbf{x}'_i)$  is invertible, which holds under (12.7) or equivalently (12.24).

The **instrumental variables** (IV) estimator  $\boldsymbol{\beta}$  replaces the population moments by their sample versions. We find

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{iv}} &= \left( \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{x}'_i \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i y_i \right) \\ &= \left( \sum_{i=1}^n \mathbf{z}_i \mathbf{x}'_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{z}_i y_i \right) \\ &= (\mathbf{Z}' \mathbf{X})^{-1} (\mathbf{Z}' \mathbf{y}).\end{aligned}\tag{12.26}$$

More generally, it is common to refer to any estimator of the form

$$\hat{\boldsymbol{\beta}}_{\text{iv}} = (\mathbf{W}' \mathbf{X})^{-1} (\mathbf{W}' \mathbf{y})$$

given an  $n \times k$  matrix  $\mathbf{W}$  as an IV estimator for  $\boldsymbol{\beta}$  using the instrument  $\mathbf{W}$ .

Alternatively, recall that when  $\ell = k$  the structural parameter can be written as a function of the reduced form parameters as  $\boldsymbol{\beta} = \boldsymbol{\Gamma}^{-1} \boldsymbol{\lambda}$ . Replacing  $\boldsymbol{\Gamma}$  and  $\boldsymbol{\lambda}$  by their least-squares estimates we can construct what is called the **Indirect Least Squares** (ILS) estimator:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{ils}} &= \hat{\boldsymbol{\Gamma}}^{-1} \hat{\boldsymbol{\lambda}} \\ &= ((\mathbf{Z}' \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{X}))^{-1} ((\mathbf{Z}' \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{y})) \\ &= (\mathbf{Z}' \mathbf{X})^{-1} (\mathbf{Z}' \mathbf{Z}) (\mathbf{Z}' \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{y}) \\ &= (\mathbf{Z}' \mathbf{X})^{-1} (\mathbf{Z}' \mathbf{y}).\end{aligned}$$

We see that this equals the IV estimator (12.26). Thus the ILS and IV estimators are identical.

Given the IV estimator we define the residual vector

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}_{\text{iv}}$$

which satisfies

$$\mathbf{Z}' \hat{\mathbf{e}} = \mathbf{Z}' \mathbf{y} - \mathbf{Z}' \mathbf{X} (\mathbf{Z}' \mathbf{X})^{-1} (\mathbf{Z}' \mathbf{y}) = \mathbf{0}.\tag{12.27}$$

Since  $\mathbf{Z}$  includes an intercept, this means that the residuals sum to zero, and are uncorrelated with the included and excluded instruments.

To illustrate, we estimate the reduced form equations corresponding to the college proximity example of Table 12.1, now treating *education* as endogenous and using *college* as an instrumental variable.

Table 12.2: Reduced Form Regressions

	log(wage)	education	education	experience	experience <sup>2</sup> /100	education
experience	0.053 (0.007)	-0.410 (0.032)				-0.413 (0.032)
experience <sup>2</sup> /100	-0.219 (0.033)	0.073 (0.170)				0.093 (0.171)
black	-0.264 (0.018)	-1.006 (0.088)	-1.468 (0.115)	1.468 (0.115)	0.282 (0.026)	-1.006 (0.088)
south	-0.143 (0.017)	-0.291 (0.078)	-0.460 (0.103)	0.460 (0.103)	0.112 (0.022)	-0.267 (0.079)
urban	0.185 (0.017)	0.404 (0.085)	0.835 (0.112)	-0.835 (0.112)	-0.176 (0.025)	0.400 (0.085)
college	0.045 (0.016)	0.337 (0.081)	0.347 (0.109)	-0.347 (0.109)	-0.073 (0.023)	
public						0.430 (0.086)
private						0.123 (0.101)
age			1.061 (0.296)	-0.061 (0.296)		-0.555 (0.065)
age <sup>2</sup> /100			-1.876 (0.516)	1.876 (0.516)		1.313 (0.116)
F	17.51	8.22	1581	1112		13.87

The reduced form equations for *log(wage)* and *education* are reported in the first and second columns of Table 12.2.

Of particular interest is the equation for the endogenous regressor (*education*), and the coefficients for the excluded instruments – in this case *college*. The estimated coefficient equals 0.347 with a small standard error. This implies that growing up near a 4-year college increases average educational attainment by 0.3 years. This seems to be a reasonable magnitude.

Since the structural equation is just-identified with one right-hand-side endogenous variable, we can calculate the ILS/IV estimate for the education coefficient as the ratio of the coefficient estimates for the instrument *college* in the two equations, e.g.  $0.045/0.347 = 0.13$ , implying a 13% return to each year of education. This is substantially greater than the 7% least-squares estimate from the first column of Table 12.1.

The IV estimates of the full equation are reported in the second column of Table 12.1.

Card (1995) also points out that if *education* is endogenous, then so is our measure of *experience*, since it is calculated by subtracting *education* from *age*. He suggests that we can use the variables *age* and *age<sup>2</sup>* as instruments for *experience* and *experience<sup>2</sup>*, as they are clearly exogenous and yet highly correlated with *experience* and *experience<sup>2</sup>*. Notice that this approach treats *experience<sup>2</sup>* as a variable separate from *experience*. Indeed, this is the correct approach.

Following this recommendation we now have three endogenous regressors and three instruments. We present the three reduced form equations for the three endogenous regressors in the third through fifth columns of Table 12.2. It is interesting to compare the equations for *education* and *experience*. The two sets of coefficients are simply the sign change of the other, with the exception of the coefficient on *age*. Indeed this must be the case, because the three variables are linearly related. Does this cause a problem for 2SLS? Fortunately, no. The fact that the coefficient on *age* is not simply a sign change means that the equations are not linearly singular. Hence Assumption (12.24) is not violated.

The IV estimates using the three instruments *college*, *age* and *age<sup>2</sup>* for the endogenous regressors *education*, *experience* and *experience<sup>2</sup>* is presented in the third column of Table 12.1. The estimate of

the returns to schooling is not affected by this change in the instrument set, but the estimated return to experience profile flattens (the quadratic effect diminishes).

The IV estimator may be calculated in Stata using the `ivregress 2sls` command.

## 12.10 Demeaned Representation

Does the well-known demeaned representation for linear regression (3.19) carry over to the IV estimator? To see this, write the linear projection equation in the format

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \alpha + e_i$$

where  $\alpha$  is the intercept and  $\mathbf{x}_i$  does not contain a constant. Similarly, partition the instrument as  $(1, \mathbf{z}_i)$  where  $\mathbf{z}_i$  does not contain an intercept. We can write the IV estimates as

$$y_i = \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{\text{iv}} + \hat{\alpha}_{\text{iv}} + \hat{e}_i.$$

The orthogonality (12.27) implies the two-equation system

$$\begin{aligned} \sum_{i=1}^n (y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{\text{iv}} - \hat{\alpha}_{\text{iv}}) &= 0 \\ \sum_{i=1}^n \mathbf{z}_i (y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{\text{iv}} - \hat{\alpha}_{\text{iv}}) &= \mathbf{0}. \end{aligned}$$

The first equation implies

$$\hat{\alpha}_{\text{iv}} = \bar{y} - \bar{\mathbf{x}}' \hat{\boldsymbol{\beta}}_{\text{iv}}.$$

Substituting into the second equation

$$\sum_{i=1}^n \mathbf{z}_i ((y_i - \bar{y}) - (\mathbf{x}_i - \bar{\mathbf{x}})' \hat{\boldsymbol{\beta}}_{\text{iv}})$$

and solving for  $\hat{\boldsymbol{\beta}}_{\text{iv}}$  we find

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{iv}} &= \left( \sum_{i=1}^n \mathbf{z}_i (\mathbf{x}_i - \bar{\mathbf{x}})' \right)^{-1} \left( \sum_{i=1}^n \mathbf{z}_i (y_i - \bar{y}) \right) \\ &= \left( \sum_{i=1}^n (\mathbf{z}_i - \bar{\mathbf{z}}) (\mathbf{x}_i - \bar{\mathbf{x}})' \right)^{-1} \left( \sum_{i=1}^n (\mathbf{z}_i - \bar{\mathbf{z}}) (y_i - \bar{y}) \right). \end{aligned} \quad (12.28)$$

Thus the demeaning equations for least-squares carry over to the IV estimator. The coefficient estimate  $\hat{\boldsymbol{\beta}}_{\text{iv}}$  is a function only of the demeaned data.

## 12.11 Wald Estimator

In many cases, including the Card proximity example, the excluded instrument is a binary (dummy) variable. Let's focus on that case, and suppose that the model has just one endogenous regressor and no other regressors beyond the intercept. Thus the model can be written as

$$\begin{aligned} y_i &= x_i \beta + \alpha + e_i \\ \mathbb{E}(e_i | z_i) &= 0 \end{aligned}$$

with  $z_i$  binary.

Notice that if we take expectations of the structural equation given  $z_i = 1$  and  $z_i = 0$ , respectively, we obtain

$$\begin{aligned}\mathbb{E}(y_i | z_i = 1) &= \mathbb{E}(x_i | z_i = 1)\beta + \alpha \\ \mathbb{E}(y_i | z_i = 0) &= \mathbb{E}(x_i | z_i = 0)\beta + \alpha.\end{aligned}$$

Subtracting and dividing, we obtain an expression for the slope coefficient  $\beta$

$$\beta = \frac{\mathbb{E}(y_i | z_i = 1) - \mathbb{E}(y_i | z_i = 0)}{\mathbb{E}(x_i | z_i = 1) - \mathbb{E}(x_i | z_i = 0)}. \quad (12.29)$$

The natural moment estimator for  $\beta$  replaces the expectations by the averages within the “grouped data” where  $z_i = 1$  and  $z_i = 0$ , respectively. That is, define the group means

$$\begin{aligned}\bar{y}_1 &= \frac{\sum_{i=1}^n z_i y_i}{\sum_{i=1}^n z_i}, & \bar{y}_0 &= \frac{\sum_{i=1}^n (1-z_i) y_i}{\sum_{i=1}^n (1-z_i)} \\ \bar{x}_1 &= \frac{\sum_{i=1}^n z_i x_i}{\sum_{i=1}^n z_i}, & \bar{x}_0 &= \frac{\sum_{i=1}^n (1-z_i) x_i}{\sum_{i=1}^n (1-z_i)}\end{aligned}$$

and the moment estimator

$$\hat{\beta} = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0}. \quad (12.30)$$

This is known as the “Wald estimator” as it was proposed by Wald (1940).

These expressions are rather insightful. (12.29) shows that the structural slope coefficient is the expected change in  $y_i$  due to changing the instrument divided by the expected change in  $x_i$  due to changing the instrument. Informally, it is the change in  $y$  (due to  $z$ ) over the change in  $x$  (due to  $z$ ). Equation (12.30) shows that the slope coefficient can be estimated by a simple ratio in means.

The expression (12.30) may appear like a distinct estimator from the IV estimator  $\hat{\beta}_{iv}$ , but it turns out that they are the same. That is,  $\hat{\beta} = \hat{\beta}_{iv}$ . To see this, use (12.28) to find

$$\begin{aligned}\hat{\beta}_{iv} &= \frac{\sum_{i=1}^n z_i (y_i - \bar{y})}{\sum_{i=1}^n z_i (x_i - \bar{x})} \\ &= \frac{\bar{y}_1 - \bar{y}}{\bar{x}_1 - \bar{x}}.\end{aligned}$$

Then notice

$$\bar{y}_1 - \bar{y} = \bar{y}_1 - \left( \frac{1}{n} \sum_{i=1}^n z_i \bar{y}_1 + \frac{1}{n} \sum_{i=1}^n (1-z_i) \bar{y}_0 \right) = \frac{1}{n} \sum_{i=1}^n (1-z_i) (\bar{y}_1 - \bar{y}_0)$$

and similarly

$$\bar{x}_1 - \bar{x} = \frac{1}{n} \sum_{i=1}^n (1-z_i) (\bar{x}_1 - \bar{x}_0)$$

and hence

$$\hat{\beta}_{iv} = \frac{\frac{1}{n} \sum_{i=1}^n (1-z_i) (\bar{y}_1 - \bar{y}_0)}{\frac{1}{n} \sum_{i=1}^n (1-z_i) (\bar{x}_1 - \bar{x}_0)} = \hat{\beta}$$

as defined in (12.30). Thus the Wald estimator equals the IV estimator.

We can illustrate using the Card proximity example. If we estimate a simple IV model with no covariates we obtain the estimate  $\hat{\beta}_{iv} = 0.19$ . If we estimate the group-mean log wages and education levels based on the instrument *college*, we find

	near college	not near college
log(wage)	6.311	6.156
education	13.527	12.698

Based on these estimates the Wald estimator of the slope coefficient is  $(6.311 - 6.156) / (13.527 - 12.698) = 0.19$ , the same as the IV estimator.

## 12.12 Two-Stage Least Squares

The IV estimator described in the previous section presumed  $\ell = k$ . Now we allow the general case of  $\ell \geq k$ . Examining the reduced-form equation (12.14) we see

$$\begin{aligned} y_i &= \mathbf{z}'_i \boldsymbol{\Gamma} \boldsymbol{\beta} + \nu_i \\ \mathbb{E}(\mathbf{z}_i \nu_i) &= \mathbf{0}. \end{aligned}$$

Defining  $\mathbf{w}_i = \boldsymbol{\Gamma}' \mathbf{z}_i$  we can write this as

$$\begin{aligned} y_i &= \mathbf{w}'_i \boldsymbol{\beta} + \nu_i \\ \mathbb{E}(\mathbf{w}_i \nu_i) &= \mathbf{0}. \end{aligned}$$

Suppose that  $\boldsymbol{\Gamma}$  were known. Then we would estimate  $\boldsymbol{\beta}$  by least-squares of  $y_i$  on  $\mathbf{w}_i = \boldsymbol{\Gamma}' \mathbf{z}_i$

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{W}' \mathbf{W})^{-1} (\mathbf{W}' \mathbf{y}) \\ &= (\boldsymbol{\Gamma}' \mathbf{Z}' \mathbf{Z} \boldsymbol{\Gamma})^{-1} (\boldsymbol{\Gamma}' \mathbf{Z}' \mathbf{y}). \end{aligned}$$

While this is infeasible, we can estimate  $\boldsymbol{\Gamma}$  from the reduced form regression. Replacing  $\boldsymbol{\Gamma}$  with its estimate  $\hat{\boldsymbol{\Gamma}} = (\mathbf{Z}' \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{X})$  we obtain

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{2sls}} &= (\hat{\boldsymbol{\Gamma}}' \mathbf{Z}' \mathbf{Z} \hat{\boldsymbol{\Gamma}})^{-1} (\hat{\boldsymbol{\Gamma}}' \mathbf{Z}' \mathbf{y}) \\ &= (\mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{y} \\ &= (\mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{y}. \end{aligned} \quad (12.31)$$

This is called the **two-stage-least squares** (2SLS) estimator. It was originally proposed by Theil (1953) and Basman (1957), and is a standard estimator for linear equations with instruments.

If the model is just-identified, so that  $k = \ell$ , then 2SLS simplifies to the IV estimator of the previous section. Since the matrices  $\mathbf{X}' \mathbf{Z}$  and  $\mathbf{Z}' \mathbf{X}$  are square, we can factor

$$\begin{aligned} (\mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X})^{-1} &= (\mathbf{Z}' \mathbf{X})^{-1} ((\mathbf{Z}' \mathbf{Z})^{-1})^{-1} (\mathbf{X}' \mathbf{Z})^{-1} \\ &= (\mathbf{Z}' \mathbf{X})^{-1} (\mathbf{Z}' \mathbf{Z}) (\mathbf{X}' \mathbf{Z})^{-1}. \end{aligned}$$

(Once again, this only works when  $k = \ell$ .) Then

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{2sls}} &= (\mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{y} \\ &= (\mathbf{Z}' \mathbf{X})^{-1} (\mathbf{Z}' \mathbf{Z}) (\mathbf{X}' \mathbf{Z})^{-1} \mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{y} \\ &= (\mathbf{Z}' \mathbf{X})^{-1} (\mathbf{Z}' \mathbf{Z}) (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{y} \\ &= (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \mathbf{y} \\ &= \hat{\boldsymbol{\beta}}_{\text{iv}} \end{aligned}$$

as claimed. This shows that the 2SLS estimator as defined in (12.31) is a generalization of the IV estimator defined in (12.26).

There are several alternative representations of the 2SLS estimator which we now describe. First, defining the projection matrix

$$\mathbf{P}_{\mathbf{Z}} = \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \quad (12.32)$$

we can write the 2SLS estimator more compactly as

$$\hat{\boldsymbol{\beta}}_{\text{2sls}} = (\mathbf{X}' \mathbf{P}_{\mathbf{Z}} \mathbf{X})^{-1} \mathbf{X}' \mathbf{P}_{\mathbf{Z}} \mathbf{y}. \quad (12.33)$$

This is useful for representation and derivations, but is not useful for computation as the  $n \times n$  matrix  $\mathbf{P}_Z$  is too large to compute when  $n$  is large.

Second, define the fitted values for  $\mathbf{X}$  from the reduced form

$$\hat{\mathbf{X}} = \mathbf{P}_Z \mathbf{X} = \mathbf{Z}\hat{\boldsymbol{\Gamma}}.$$

Then the 2SLS estimator can be written as

$$\hat{\boldsymbol{\beta}}_{2\text{sls}} = (\hat{\mathbf{X}}' \mathbf{X})^{-1} \hat{\mathbf{X}}' \mathbf{y}.$$

This is an IV estimator as defined in the previous section using  $\hat{\mathbf{X}}$  as the instrument.

Third, since  $\mathbf{P}_Z$  is idempotent, we can also write the 2SLS estimator as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{2\text{sls}} &= (\mathbf{X}' \mathbf{P}_Z \mathbf{P}_Z \mathbf{X})^{-1} \mathbf{X}' \mathbf{P}_Z \mathbf{y} \\ &= (\hat{\mathbf{X}}' \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}' \mathbf{y}\end{aligned}$$

which is the least-squares estimator obtained by regressing  $\mathbf{y}$  on the fitted values  $\hat{\mathbf{X}}$ .

This is the source of the “two-stage” name is since it can be computed as follows.

- First regress  $\mathbf{X}$  on  $\mathbf{Z}$ , vis.,  $\hat{\boldsymbol{\Gamma}} = (\mathbf{Z}' \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{X})$  and  $\hat{\mathbf{X}} = \mathbf{Z}\hat{\boldsymbol{\Gamma}} = \mathbf{P}_Z \mathbf{X}$ .
- Second, regress  $\mathbf{y}$  on  $\hat{\mathbf{X}}$ , vis.,  $\hat{\boldsymbol{\beta}}_{2\text{sls}} = (\hat{\mathbf{X}}' \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}' \mathbf{y}$ .

It is useful to scrutinize the projection  $\hat{\mathbf{X}}$ . Recall,  $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$  and  $\mathbf{Z} = [\mathbf{X}_1, \mathbf{Z}_2]$ . Notice  $\hat{\mathbf{X}}_1 = \mathbf{P}_Z \mathbf{X}_1 = \mathbf{X}_1$  since  $\mathbf{X}_1$  lies in the span of  $\mathbf{Z}$ . Then

$$\hat{\mathbf{X}} = [\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2] = [\mathbf{X}_1, \hat{\mathbf{X}}_2].$$

Thus in the second stage, we regress  $\mathbf{y}$  on  $\mathbf{X}_1$  and  $\hat{\mathbf{X}}_2$ . So only the endogenous variables  $\mathbf{X}_2$  are replaced by their fitted values:

$$\hat{\mathbf{X}}_2 = \mathbf{X}_1 \hat{\boldsymbol{\Gamma}}_{12} + \mathbf{Z}_2 \hat{\boldsymbol{\Gamma}}_{22}.$$

This least squares estimator can be written as

$$\mathbf{y} = \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 + \hat{\mathbf{X}}_2 \hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\epsilon}}.$$

A fourth representation of 2SLS can be obtained from the previous representation for  $\hat{\boldsymbol{\beta}}_2$ . Set  $\mathbf{P}_1 = \mathbf{X}_1 (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1'$ . Applying the FWL theorem we obtain

$$\begin{aligned}\hat{\boldsymbol{\beta}}_2 &= (\hat{\mathbf{X}}_2' (\mathbf{I}_n - \mathbf{P}_1) \hat{\mathbf{X}}_2)^{-1} (\hat{\mathbf{X}}_2' (\mathbf{I}_n - \mathbf{P}_1) \mathbf{y}) \\ &= (\mathbf{X}_2' \mathbf{P}_Z (\mathbf{I}_n - \mathbf{P}_1) \mathbf{P}_Z \mathbf{X}_2)^{-1} (\mathbf{X}_2' \mathbf{P}_Z (\mathbf{I}_n - \mathbf{P}_1) \mathbf{y}) \\ &= (\mathbf{X}_2' (\mathbf{P}_Z - \mathbf{P}_1) \mathbf{X}_2)^{-1} (\mathbf{X}_2' (\mathbf{P}_Z - \mathbf{P}_1) \mathbf{y})\end{aligned}$$

since  $\mathbf{P}_Z \mathbf{P}_1 = \mathbf{P}_1$ .

A fifth representation can be obtained by a further projection. The projection matrix  $\mathbf{P}_Z$  can be replaced by the projection onto the pair  $[\mathbf{X}_1, \tilde{\mathbf{Z}}_2]$  where  $\tilde{\mathbf{Z}}_2 = (\mathbf{I}_n - \mathbf{P}_1) \mathbf{Z}_2$  is  $\mathbf{Z}_2$  projected orthogonal to  $\mathbf{X}_1$ . Since  $\mathbf{X}_1$  and  $\tilde{\mathbf{Z}}_2$  are orthogonal,  $\mathbf{P}_Z = \mathbf{P}_1 + \mathbf{P}_2$  where  $\mathbf{P}_2 = \tilde{\mathbf{Z}}_2 (\tilde{\mathbf{Z}}_2' \tilde{\mathbf{Z}}_2)^{-1} \tilde{\mathbf{Z}}_2'$ . Thus  $\mathbf{P}_Z - \mathbf{P}_1 = \mathbf{P}_2$  and

$$\begin{aligned}\hat{\boldsymbol{\beta}}_2 &= (\mathbf{X}_2' \mathbf{P}_2 \mathbf{X}_2)^{-1} (\mathbf{X}_2' \mathbf{P}_2 \mathbf{y}) \\ &= \left( \mathbf{X}_2' \tilde{\mathbf{Z}}_2 (\tilde{\mathbf{Z}}_2' \tilde{\mathbf{Z}}_2)^{-1} \tilde{\mathbf{Z}}_2' \mathbf{X}_2 \right)^{-1} \left( \mathbf{X}_2' \tilde{\mathbf{Z}}_2 (\tilde{\mathbf{Z}}_2' \tilde{\mathbf{Z}}_2)^{-1} \tilde{\mathbf{Z}}_2' \mathbf{y} \right).\end{aligned}\tag{12.34}$$

Given the 2SLS estimator we define the residual vector

$$\hat{\boldsymbol{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{2\text{sls}},$$

When the model is overidentified, the instruments and residuals are not orthogonal. That is

$$\mathbf{Z}'\hat{\boldsymbol{e}} \neq \mathbf{0}.$$

It does, however, satisfy

$$\begin{aligned}\hat{\mathbf{X}}'\hat{\boldsymbol{e}} &= \hat{\boldsymbol{\Gamma}}'\mathbf{Z}'\hat{\boldsymbol{e}} \\ &= \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\boldsymbol{e}} \\ &= \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} - \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\hat{\boldsymbol{\beta}}_{2\text{sls}} \\ &= \mathbf{0}.\end{aligned}$$

Returning to Card's college proximity example, suppose that we treat experience as exogenous, but that instead of using the single instrument *college* (grew up near a 4-year college) we use the two instruments (*public*, *private*) (grew up near a public/private 4-year college, respectively). In this case we have one endogenous variable (*education*) and two instruments (*public*, *private*). The estimated reduced form equation for *education* is presented in the sixth column of Table 12.2. In this specification, the coefficient on *public* – growing up near a public 4-year college – is larger than that found for the variable *college* in the previous specification (column 2). Furthermore, the coefficient on *private* – growing up near a private 4-year college – is much smaller. This indicates that the key impact of proximity on education is via public colleges rather than private colleges.

The 2SLS estimates obtained using these two instruments are presented in the fourth column of Table 12.1. The coefficient on *education* increases to 0.161, indicating a 16% return to a year of education. This is roughly twice as large as the estimate obtained by least-squares in the first column.

Additionally, if we follow Card and treat *experience* as endogenous and use *age* as an instrument, we now have three endogenous variables (*education*, *experience*, *experience*<sup>2</sup>/100) and four instruments (*public*, *private*, *age*, *age*<sup>2</sup>). We present the 2SLS estimates using this specification in the fifth column of Table 12.1. The estimate of the return to education remains about 16%, but again the return to experience flattens.

You might wonder if we could use all three instruments – *college*, *public*, and *private*. The answer is no. This is because *college* = *public*+*private* so the three variables are colinear. Since the instruments are linearly related, the three together would violate the full-rank condition (12.6).

The 2SLS estimator may be calculated in Stata using the `ivregress 2sls` command.

## 12.13 Limited Information Maximum Likelihood

An alternative method to estimate the parameters of the structural equation is by maximum likelihood. Anderson and Rubin (1949) derived the maximum likelihood estimator for the joint distribution of  $(y_i, \mathbf{x}_{2i})$ . The estimator is known as **limited information maximum likelihood**, or LIML.

This estimator is called “limited information” because it is based on the structural equation for  $y_i$  combined with the reduced form equation for  $\mathbf{x}_{2i}$ . If maximum likelihood is derived based on a structural equation for  $\mathbf{x}_{2i}$  as well, then this leads to what is known as **full information maximum likelihood** (FIML). The advantage of the LIML approach relative to FIML is that the former does not require a structural model for  $\mathbf{x}_{2i}$ , and thus allows the researcher to focus on the structural equation of interest – that for  $y_i$ . We do not describe the FIML estimator here as it is not commonly used in applied econometric practice.

While the LIML estimator is less widely used among economists than 2SLS, it has received a resurgence of attention from econometric theorists.

To derive the LIML estimator, start by writing the joint reduced form equations (12.16) and (12.13) as

$$\begin{aligned}\mathbf{y}_i &= \begin{pmatrix} y_i \\ \mathbf{x}_{2i} \end{pmatrix} \\ &= \begin{bmatrix} \boldsymbol{\lambda}'_1 & \boldsymbol{\lambda}'_2 \\ \boldsymbol{\Gamma}'_{12} & \boldsymbol{\Gamma}'_{22} \end{bmatrix} \begin{pmatrix} \mathbf{z}_{1i} \\ \mathbf{z}_{2i} \end{pmatrix} + \begin{pmatrix} v_i \\ \mathbf{u}_{2i} \end{pmatrix} \\ &= \boldsymbol{\Pi}'_1 \mathbf{z}_{1i} + \boldsymbol{\Pi}'_2 \mathbf{z}_{2i} + \boldsymbol{a}_i\end{aligned}\quad (12.35)$$

where  $\boldsymbol{\Pi}_1 = [\boldsymbol{\lambda}_1 \ \boldsymbol{\Gamma}_{12}]$ ,  $\boldsymbol{\Pi}_2 = [\boldsymbol{\lambda}_2 \ \boldsymbol{\Gamma}_{22}]$  and  $\boldsymbol{a}'_i = [v_i \ \mathbf{u}'_{2i}]$ . The LIML estimator is derived under the assumption that  $\boldsymbol{a}_i$  is multivariate normal.

Define  $\boldsymbol{\gamma}' = [1 \ -\boldsymbol{\beta}'_2]$ . From (12.18) we find

$$\boldsymbol{\Pi}_2 \boldsymbol{\gamma} = \boldsymbol{\lambda}_2 - \boldsymbol{\Gamma}_{22} \boldsymbol{\beta}_2 = \mathbf{0}.$$

Thus the  $\ell_2 \times (k_2 + 1)$  coefficient matrix  $\boldsymbol{\Pi}_2$  in (12.35) has deficient rank. Indeed, its rank must be  $k_2$ , since  $\boldsymbol{\Gamma}_{22}$  has full rank.

This means that the model (12.35) is precisely the reduced rank regression model of Section 11.11. Theorem 11.7 presents the maximum likelihood estimators for the reduced rank parameters. In particular, the MLE for  $\boldsymbol{\gamma}$  is

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \frac{\boldsymbol{\gamma}' \mathbf{Y}' \mathbf{M}_1 \mathbf{Y} \boldsymbol{\gamma}}{\boldsymbol{\gamma}' \mathbf{Y}' \mathbf{M}_Z \mathbf{Y} \boldsymbol{\gamma}} \quad (12.36)$$

where  $\mathbf{Y} = [\mathbf{y}, \mathbf{X}_2]$  is the  $n \times (1 + k_2)$  matrix of the stacked endogenous variables  $\mathbf{y}'_i = (y_i \ \mathbf{x}'_{2i})$ ,  $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{Z}_1 (\mathbf{Z}'_1 \mathbf{Z}_1)^{-1} \mathbf{Z}'_1$  and  $\mathbf{M}_Z = \mathbf{I}_n - \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}'$ . The minimization (12.36) is sometimes called the “least variance ratio” problem.

The minimization problem (12.36) is invariant to the scale of  $\boldsymbol{\gamma}$  (that is,  $\hat{\boldsymbol{\gamma}}c$  is equivalently the argmin for any  $c$ ) so a normalization is required. For estimation of the structural parameters a convenient normalization is  $\boldsymbol{\gamma}' = [1 \ -\boldsymbol{\beta}'_2]$ . Another is to set  $\boldsymbol{\gamma}' \mathbf{Y}' \mathbf{M}_Z \mathbf{Y} \boldsymbol{\gamma} = 1$ . Using the second normalization and the theory of the minimum of quadratic forms (Section A.15)  $\hat{\boldsymbol{\gamma}}$  is the generalized eigenvector of  $\mathbf{Y}' \mathbf{M}_1 \mathbf{Y}$  with respect to  $\mathbf{Y}' \mathbf{M}_Z \mathbf{Y}$  associated with the smallest generalized eigenvalue. (See Section A.14 for the definition of generalized eigenvalues and eigenvectors.) Computationally this is straightforward. For example, in MATLAB, the generalized eigenvalues and eigenvectors of the matrix  $\mathbf{A}$  with respect to  $\mathbf{B}$  is found by the command `eig(A, B)`. Once this  $\hat{\boldsymbol{\gamma}}$  is found, any other normalization can be obtained by rescaling. For example, to obtain the MLE for  $\boldsymbol{\beta}_2$  make the partition  $\hat{\boldsymbol{\gamma}}' = [\hat{\gamma}_1 \ \hat{\gamma}'_2]$  and set  $\hat{\boldsymbol{\beta}}_2 = -\hat{\gamma}_2/\hat{\gamma}_1$ .

To obtain the MLE for  $\boldsymbol{\beta}_1$ , recall the structural equation  $y_i = \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + e_i$ . Replacing  $\boldsymbol{\beta}_2$  with the MLE  $\hat{\boldsymbol{\beta}}_2$  and then apply regression. Thus

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 (\mathbf{y} - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2). \quad (12.37)$$

These solutions are the MLE (known as the LIML estimator) for the structural parameters  $\boldsymbol{\beta}_1$  and  $\boldsymbol{\beta}_2$ .

Many previous econometrics textbooks do not present a derivation of the LIML estimator as the original derivation by Anderson and Rubin (1949) is lengthy and not particularly insightful. In contrast, the derivation given here based on reduced rank regression is relatively simple.

There is an alternative (and traditional) expression for the LIML estimator. Define the minimum obtained in (12.36)

$$\hat{\kappa} = \underset{\boldsymbol{\gamma}}{\min} \frac{\boldsymbol{\gamma}' \mathbf{Y}' \mathbf{M}_1 \mathbf{Y} \boldsymbol{\gamma}}{\boldsymbol{\gamma}' \mathbf{Y}' \mathbf{M}_Z \mathbf{Y} \boldsymbol{\gamma}} \quad (12.38)$$

which is the smallest generalized eigenvalue of  $\mathbf{Y}' \mathbf{M}_1 \mathbf{Y}$  with respect to  $\mathbf{Y}' \mathbf{M}_Z \mathbf{Y}$ . The LIML estimator then can be written as

$$\hat{\boldsymbol{\beta}}_{\text{lml}} = (\mathbf{X}' (\mathbf{I}_n - \hat{\kappa} \mathbf{M}_Z) \mathbf{X})^{-1} (\mathbf{X}' (\mathbf{I}_n - \hat{\kappa} \mathbf{M}_Z) \mathbf{y}). \quad (12.39)$$

We defer the derivation of (12.39) until the end of this section. Expression (12.39) does not simplify computation (since  $\hat{\kappa}$  requires solving the same eigenvector problem that yields  $\hat{\boldsymbol{\beta}}_2$ ). However (12.39)

is important for the distribution theory and to reveal the algebraic connection between LIML, least-squares, and 2SLS.

The estimator (12.39) with arbitrary  $\kappa$  is known as a  $k$  class estimator of  $\beta$ . While the LIML estimator obtains by setting  $\kappa = \hat{\kappa}$ , the least-squares estimator is obtained by setting  $\kappa = 0$  and 2SLS is obtained by setting  $\kappa = 1$ . It is worth observing that the LIML solution to (12.38) satisfies  $\hat{\kappa} \geq 1$ .

When the model is just-identified, the LIML estimator is identical to the IV and 2SLS estimators. They are only different in the over-identified setting. (One corollary is that under just-identification the IV estimator is MLE under normality.)

For inference, it is useful to observe that (12.39) shows that  $\hat{\beta}_{\text{liml}}$  can be written as an IV estimator

$$\hat{\beta}_{\text{liml}} = (\tilde{\mathbf{X}}' \mathbf{X})^{-1} (\tilde{\mathbf{X}}' \mathbf{y})$$

using the instrument

$$\tilde{\mathbf{X}} = (\mathbf{I}_n - \hat{\kappa} \mathbf{M}_Z) \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 - \hat{\kappa} \hat{\mathbf{U}}_2 \end{pmatrix}$$

where  $\hat{\mathbf{U}}_2 = \mathbf{M}_Z \mathbf{X}_2$  are the (reduced-form) residuals from the multivariate regression of the endogenous regressors  $\mathbf{x}_{2i}$  on the instruments  $\mathbf{z}_i$ . Expressing LIML using this IV formula is useful for variance estimation.

Asymptotically the LIML estimator has the same distribution as 2SLS. However, they can have quite different behaviors in finite samples. There is considerable evidence that the LIML estimator has superior finite sample bias relative to 2SLS when there are many instruments or the reduced form is weak. (We review these cases in the following sections.) However, on the other hand LIML has wider finite sample dispersion.

We now derive the expression (12.39). Use the normalization  $\gamma' = [1 \ -\boldsymbol{\beta}_2']$  to write (12.36) as

$$\hat{\boldsymbol{\beta}}_2 = \underset{\boldsymbol{\beta}_2}{\operatorname{argmin}} \frac{(\mathbf{y} - \mathbf{X}_2 \boldsymbol{\beta}_2)' \mathbf{M}_1 (\mathbf{y} - \mathbf{X}_2 \boldsymbol{\beta}_2)}{(\mathbf{y} - \mathbf{X}_2 \boldsymbol{\beta}_2)' \mathbf{M}_Z (\mathbf{y} - \mathbf{X}_2 \boldsymbol{\beta}_2)}.$$

The first-order-condition for minimization is  $2/(\mathbf{y} - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2)' \mathbf{M}_Z (\mathbf{y} - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2)$  times

$$\begin{aligned} \mathbf{0} &= \mathbf{X}_2' \mathbf{M}_1 (\mathbf{y} - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2) - \frac{(\mathbf{y} - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2)' \mathbf{M}_1 (\mathbf{y} - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2)}{(\mathbf{y} - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2)' \mathbf{M}_Z (\mathbf{y} - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2)} \mathbf{X}_2' \mathbf{M}_Z (\mathbf{y} - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2) \\ &= \mathbf{X}_2' \mathbf{M}_1 (\mathbf{y} - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2) - \hat{\kappa} \mathbf{X}_2' \mathbf{M}_Z (\mathbf{y} - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2) \end{aligned}$$

using definition (12.38). Rewriting,

$$\mathbf{X}_2' (\mathbf{M}_1 - \hat{\kappa} \mathbf{M}_Z) \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2' (\mathbf{M}_1 - \hat{\kappa} \mathbf{M}_Z) \mathbf{y}. \quad (12.40)$$

Equation (12.39) is the same as the two equation system

$$\begin{aligned} \mathbf{X}_1' \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 + \mathbf{X}_1' \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2 &= \mathbf{X}_1' \mathbf{y} \\ \mathbf{X}_2' \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 + (\mathbf{X}_2' (\mathbf{I}_n - \hat{\kappa} \mathbf{M}_Z) \mathbf{X}_2) \hat{\boldsymbol{\beta}}_2 &= \mathbf{X}_2' (\mathbf{I}_n - \hat{\kappa} \mathbf{M}_Z) \mathbf{y}. \end{aligned}$$

The first equation is (12.37). Using (12.37), the second is

$$\mathbf{X}_2' \mathbf{X}_1 (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' (\mathbf{y} - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2) + (\mathbf{X}_2' (\mathbf{I}_n - \hat{\kappa} \mathbf{M}_Z) \mathbf{X}_2) \hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2' (\mathbf{I}_n - \hat{\kappa} \mathbf{M}_Z) \mathbf{y}$$

which is (12.40) when rearranged. We have thus shown that (12.39) is equivalent to (12.37) and (12.40) and is thus a valid expression for the LIML estimator.

Returning to the Card college proximity example, we now present the LIML estimates of the equation with the two instruments (*public*, *private*). They are reported in the final column of Table 12.1. They are quite similar to the 2SLS estimates in this application.

The LIML estimator may be calculated in Stata using the `ivregress liml` command.

### Theodore Anderson

Theodore (Ted) Anderson (1918-2016) was a American statistician and econometrician, who made fundamental contributions to multivariate statistical theory. Important contributions include the Anderson-Darling distribution test, the Anderson-Rubin statistic, the method of reduced rank regression, and his most famous econometrics contribution – the LIML estimator. He continued working throughout his long life, even publishing theoretical work at the age of 97!

## 12.14 JIVE

The ideal instrument for estimation of  $\beta$  is  $w_i = \Gamma' z_i$ . We can write this ideal estimator as

$$\hat{\beta}_{\text{ideal}} = \left( \sum_{i=1}^n w_i x'_i \right)^{-1} \left( \sum_{i=1}^n w_i y_i \right).$$

This estimator is not feasible since  $\Gamma$  is unknown. The 2SLS estimator replaces  $\Gamma$  with the multivariate least-squares estimator  $\hat{\Gamma}$  and  $w_i$  with  $\hat{w}_i = \hat{\Gamma}' z_i$  leading to the following representation for 2SLS

$$\hat{\beta}_{\text{2sls}} = \left( \sum_{i=1}^n \hat{w}_i x'_i \right)^{-1} \left( \sum_{i=1}^n \hat{w}_i y_i \right).$$

Since  $\hat{\Gamma}$  is estimated on the full sample including observation  $i$  it is a function of the reduced form error  $u_i$  which is correlated with the structural error  $e_i$ . It follows that  $\hat{w}_i$  and  $e_i$  are correlated, which means that  $\hat{\beta}_{\text{2sls}}$  is biased for  $\beta$ . This correlation and bias disappears asymptotically but it can be important in applications.

A solution to this problem is to replace  $\hat{w}_i$  with a predicted value which is uncorrelated with the error  $e_i$ . This can be obtained by a standard leave-one-out estimator for  $\Gamma$ . Specifically, let

$$\hat{\Gamma}_{(-i)} = (\mathbf{Z}' \mathbf{Z} - \mathbf{z}_i \mathbf{z}'_i)^{-1} (\mathbf{Z}' \mathbf{X} - \mathbf{z}_i \mathbf{x}'_i)$$

be the least-squares leave-one-out estimator of the reduced form matrix  $\Gamma$ , and let  $\tilde{w}_i = \hat{\Gamma}'_{(-i)} \mathbf{z}_i$  be the reduced form predicted values. Using  $\tilde{w}_i$  as an instrument we obtain the estimator

$$\begin{aligned} \hat{\beta}_{\text{jive1}} &= \left( \sum_{i=1}^n \tilde{w}_i x'_i \right)^{-1} \left( \sum_{i=1}^n \tilde{w}_i y_i \right) \\ &= \left( \sum_{i=1}^n \hat{\Gamma}'_{(-i)} \mathbf{z}_i \mathbf{x}'_i \right)^{-1} \left( \sum_{i=1}^n \hat{\Gamma}'_{(-i)} \mathbf{z}_i y_i \right). \end{aligned}$$

This was called the jackknife instrumental variables (JIVE1) estimator by Angrist, Imbens, and Krueger (1999). It first appeared in Phillips and Hale (1977).

Angrist, Imbens, and Krueger (1999) pointed out that a somewhat simpler adjustment also removes the correlation and bias. Define the estimator and predicted value

$$\begin{aligned} \bar{\Gamma}_{(-i)} &= (\mathbf{Z}' \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{X} - \mathbf{z}_i \mathbf{x}'_i) \\ \bar{w}_i &= \bar{\Gamma}'_{(-i)} \mathbf{z}_i \end{aligned}$$

which only adjusts the  $\mathbf{Z}' \mathbf{X}$  component. Their JIVE2 estimator is

$$\begin{aligned} \hat{\beta}_{\text{jive2}} &= \left( \sum_{i=1}^n \bar{w}_i x'_i \right)^{-1} \left( \sum_{i=1}^n \bar{w}_i y_i \right) \\ &= \left( \sum_{i=1}^n \bar{\Gamma}'_{(-i)} \mathbf{z}_i \mathbf{x}'_i \right)^{-1} \left( \sum_{i=1}^n \bar{\Gamma}'_{(-i)} \mathbf{z}_i y_i \right). \end{aligned}$$

Using the formula for leave-one-out estimators (Theorem 3.7), the JIVE1 and JIVE2 estimators use two linear operations: the first to create the predicted values  $\tilde{\mathbf{w}}_i$  or  $\bar{\mathbf{w}}_i$ , and the second to calculate the IV estimator. Thus the estimators do not require significantly more computation than 2SLS.

An asymptotic distribution theory for the JIVE1 and JIVE2 estimators was developed by Chao, Swanson, Hausman, Newey, and Woutersen (2012).

The JIVE1 and JIVE2 estimators may be calculated in Stata using the `jive` command. It is not a part of the standard package but can be easily added.

## 12.15 Consistency of 2SLS

We now present a demonstration of the consistency of the 2SLS estimate for the structural parameter. The following is a set of regularity conditions.

### Assumption 12.1

1. The observations  $(y_i, \mathbf{x}_i, \mathbf{z}_i)$ ,  $i = 1, \dots, n$ , are independent and identically distributed.
2.  $\mathbb{E}(y^2) < \infty$ .
3.  $\mathbb{E}\|\mathbf{x}\|^2 < \infty$ .
4.  $\mathbb{E}\|\mathbf{z}\|^2 < \infty$ .
5.  $\mathbb{E}(\mathbf{z}\mathbf{z}')$  is positive definite.
6.  $\mathbb{E}(\mathbf{z}\mathbf{x}')$  has full rank  $k$ .
7.  $\mathbb{E}(\mathbf{z}\mathbf{e}) = 0$ .

Assumptions 12.1.2-4 state that all variables have finite variances. Assumption 12.1.5 states that the instrument vector has an invertible design matrix, which is identical to the core assumption about regressors in the linear regression model. This excludes linearly redundant instruments. Assumptions 12.1.6 and 12.1.7 are the key identification conditions for instrumental variables. Assumption 12.1.6 states that the instruments and regressors have a full-rank cross-moment matrix. This is often called the relevance condition. Assumption 12.1.7 states that the instrumental variables and structural error are uncorrelated. Assumptions 12.1.5-7 are identical to Definition 12.1.

**Theorem 12.1** Under Assumption 12.1,  $\hat{\boldsymbol{\beta}}_{\text{2sls}} \xrightarrow{P} \boldsymbol{\beta}$  as  $n \rightarrow \infty$ .

The proof of the theorem is provided below.

This theorem shows that the 2SLS estimator is consistent for the structural coefficient  $\boldsymbol{\beta}$  under similar moment conditions as the least-squares estimator. The key differences are the instrumental variables assumption  $\mathbb{E}(\mathbf{z}\mathbf{e}) = 0$  and the identification assumption  $\text{rank}(\mathbb{E}(\mathbf{z}\mathbf{x}')) = k$ .

The result includes the IV estimator (when  $\ell = k$ ) as a special case.

The proof of this consistency result is similar to that for the least-squares estimator. Take the structural equation  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$  in matrix format and substitute it into the expression for the estimator. We obtain

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{2\text{sls}} &= \left( \mathbf{Z}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X} \right)^{-1} \mathbf{Z}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' (\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) \\ &= \boldsymbol{\beta} + \left( \mathbf{Z}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X} \right)^{-1} \mathbf{Z}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{e}.\end{aligned}\quad (12.41)$$

This separates out the stochastic component. Re-writing and applying the WLLN and CMT

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{2\text{sls}} - \boldsymbol{\beta} &= \left( \left( \frac{1}{n} \mathbf{X}' \mathbf{Z} \right) \left( \frac{1}{n} \mathbf{Z}' \mathbf{Z} \right)^{-1} \left( \frac{1}{n} \mathbf{Z}' \mathbf{X} \right) \right)^{-1} \\ &\quad \cdot \left( \frac{1}{n} \mathbf{X}' \mathbf{Z} \right) \left( \frac{1}{n} \mathbf{Z}' \mathbf{Z} \right)^{-1} \left( \frac{1}{n} \mathbf{Z}' \mathbf{e} \right) \\ &\xrightarrow{p} (\mathbf{Q}_{xz} \mathbf{Q}_{zz}^{-1} \mathbf{Q}_{zx})^{-1} \mathbf{Q}_{xz} \mathbf{Q}_{zz}^{-1} \mathbb{E}(\mathbf{z}_i \mathbf{e}_i) = 0\end{aligned}$$

where

$$\begin{aligned}\mathbf{Q}_{xz} &= \mathbb{E}(\mathbf{x}_i \mathbf{z}'_i) \\ \mathbf{Q}_{zz} &= \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i) \\ \mathbf{Q}_{zx} &= \mathbb{E}(\mathbf{z}_i \mathbf{x}'_i).\end{aligned}$$

The WLLN holds under the i.i.d. Assumption 12.1.1 and the finite second moment Assumptions 12.1.2-4. The continuous mapping theorem applies if the matrices  $\mathbf{Q}_{zz}$  and  $\mathbf{Q}_{xz} \mathbf{Q}_{zz}^{-1} \mathbf{Q}_{zx}$  are invertible, which hold under the identification Assumptions 12.1.5 and 12.1.6. The final equality uses Assumption 12.1.7.

## 12.16 Asymptotic Distribution of 2SLS

We now show that the 2SLS estimator satisfies a central limit theorem. We first state a set of sufficient regularity conditions.

**Assumption 12.2** In addition to Assumption 12.1,

1.  $\mathbb{E}(y^4) < \infty$ .
2.  $\mathbb{E}\|\mathbf{z}\|^4 < \infty$ .
3.  $\boldsymbol{\Omega} = \mathbb{E}(\mathbf{z}\mathbf{z}' e^2)$  is positive definite.

Assumption 12.2 strengthens Assumption 12.1 by requiring that the dependent variable and instruments have finite fourth moments. This is used to establish the central limit theorem.

**Theorem 12.2** Under Assumption 12.2, as  $n \rightarrow \infty$ .

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{2\text{sls}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, V_{\boldsymbol{\beta}})$$

where

$$V_{\boldsymbol{\beta}} = (\mathbf{Q}_{xz} \mathbf{Q}_{zz}^{-1} \mathbf{Q}_{zx})^{-1} (\mathbf{Q}_{xz} \mathbf{Q}_{zz}^{-1} \boldsymbol{\Omega} \mathbf{Q}_{zz}^{-1} \mathbf{Q}_{zx}) (\mathbf{Q}_{xz} \mathbf{Q}_{zz}^{-1} \mathbf{Q}_{zx})^{-1}.$$

This shows that the 2SLS estimator converges at a  $\sqrt{n}$  rate to a normal random vector. It shows as well the form of the covariance matrix. The latter takes a substantially more complicated form than the least-squares estimator.

As in the case of least-squares estimation, the asymptotic variance simplifies under a conditional homoskedasticity condition. For 2SLS the simplification occurs when  $\mathbb{E}(e_i^2 | \mathbf{z}_i) = \sigma^2$ . This holds when  $\mathbf{z}_i$  and  $e_i$  are independent. It may be reasonable in some contexts to conceive that the error  $e_i$  is independent of the excluded instruments  $\mathbf{z}_{2i}$ , since by assumption the impact of  $\mathbf{z}_{2i}$  on  $y_i$  is only through  $\mathbf{x}_i$ , but there is no reason to expect  $e_i$  to be independent of the included exogenous variables  $\mathbf{x}_{1i}$ . Hence heteroskedasticity should be equally expected in 2SLS and least-squares regression. Nevertheless, under the homoskedasticity condition then we have the simplifications  $\boldsymbol{\Omega} = \mathbf{Q}_{zz}\sigma^2$  and  $V_{\beta} = V_{\beta}^0 \stackrel{def}{=} (\mathbf{Q}_{xz}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zx})^{-1}\sigma^2$ .

The derivation of the asymptotic distribution builds on the proof of consistency. Using equation (12.41) we have

$$\begin{aligned} \sqrt{n}(\hat{\beta}_{2\text{SLS}} - \beta) &= \left( \left( \frac{1}{n} \mathbf{X}' \mathbf{Z} \right) \left( \frac{1}{n} \mathbf{Z}' \mathbf{Z} \right)^{-1} \left( \frac{1}{n} \mathbf{Z}' \mathbf{X} \right) \right)^{-1} \\ &\quad \cdot \left( \frac{1}{n} \mathbf{X}' \mathbf{Z} \right) \left( \frac{1}{n} \mathbf{Z}' \mathbf{Z} \right)^{-1} \left( \frac{1}{\sqrt{n}} \mathbf{Z}' \mathbf{e} \right). \end{aligned}$$

We apply the WLLN and CMT for the moment matrices involving  $\mathbf{X}$  and  $\mathbf{Z}$  the same as in the proof of consistency. In addition, by the CLT for i.i.d. observations

$$\frac{1}{\sqrt{n}} \mathbf{Z}' \mathbf{e} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{z}_i e_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$$

because the vector  $\mathbf{z}_i e_i$  is i.i.d. and mean zero under Assumptions 12.1.1 and 12.1.7, and has a finite second moment as we verify below.

We obtain

$$\begin{aligned} \sqrt{n}(\hat{\beta}_{2\text{SLS}} - \beta) &= \left( \left( \frac{1}{n} \mathbf{X}' \mathbf{Z} \right) \left( \frac{1}{n} \mathbf{Z}' \mathbf{Z} \right)^{-1} \left( \frac{1}{n} \mathbf{Z}' \mathbf{X} \right) \right)^{-1} \\ &\quad \cdot \left( \frac{1}{n} \mathbf{X}' \mathbf{Z} \right) \left( \frac{1}{n} \mathbf{Z}' \mathbf{Z} \right)^{-1} \left( \frac{1}{\sqrt{n}} \mathbf{Z}' \mathbf{e} \right) \\ &\xrightarrow{d} (\mathbf{Q}_{xz}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zx})^{-1} \mathbf{Q}_{xz}\mathbf{Q}_{zz}^{-1} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}) = \mathcal{N}(\mathbf{0}, V_{\beta}) \end{aligned}$$

as stated.

For completeness, we demonstrate that  $\mathbf{z}_i e_i$  has a finite second moment under Assumption 12.2. To see this, note that by Minkowski's inequality (B.33)

$$\begin{aligned} (\mathbb{E}(e^4))^{1/4} &= \left( \mathbb{E}((y - \mathbf{x}'\beta)^4) \right)^{1/4} \\ &\leq (\mathbb{E}(y^4))^{1/4} + \|\beta\| (\mathbb{E}\|\mathbf{x}\|^4)^{1/4} < \infty \end{aligned}$$

under Assumptions 12.2.1 and 12.2.2. Then by the Cauchy-Schwarz inequality (B.31)

$$\mathbb{E}\|\mathbf{z}e\|^2 \leq (\mathbb{E}\|\mathbf{z}\|^4)^{1/2} (\mathbb{E}(e^4))^{1/2} < \infty$$

using Assumptions 12.2.3.

## 12.17 Determinants of 2SLS Variance

It is instructive to examine the asymptotic variance of the 2SLS estimator to understand the factors which determine the precision (or lack thereof) of the estimator. As in the least-squares case, it is more

transparent to examine the variance under the assumption of homoskedasticity. In this case the asymptotic variance takes the form

$$\begin{aligned} V_{\beta}^0 &= (\mathbf{Q}_{xz}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zx})^{-1}\sigma^2 \\ &= \left(\mathbb{E}(\mathbf{x}_i\mathbf{z}'_i)(\mathbb{E}(\mathbf{z}_i\mathbf{z}'_i))^{-1}\mathbb{E}(\mathbf{z}_i\mathbf{x}'_i)\right)^{-1}\mathbb{E}(e_i^2). \end{aligned}$$

As in the least-squares case, we can see that the variance is increasing in the variance of the error  $e_i$ , and decreasing in the variance of  $\mathbf{x}_i$ . What is different is that the variance is decreasing in the (matrix-valued) correlation between  $\mathbf{x}_i$  and  $\mathbf{z}_i$ .

It is also useful to observe that the variance expression is not affected by the variance structure of  $\mathbf{z}_i$ . Indeed,  $V_{\beta}^0$  is invariant to rotations of  $\mathbf{z}_i$  (if you replace  $\mathbf{z}_i$  with  $\mathbf{C}\mathbf{z}_i$  for invertible  $\mathbf{C}$  the expression does not change). This means that the variance expression is not affected by the scaling of  $\mathbf{z}_i$ , and is not directly affected by correlation among the  $\mathbf{z}_i$ .

We can also use this expression to examine the impact of increasing the instrument set. Suppose we partition  $\mathbf{z}_i = (\mathbf{z}_{ai}, \mathbf{z}_{bi})$  where  $\dim(\mathbf{z}_{ai}) \geq k$  so we can construct the 2SLS estimator using  $\mathbf{z}_{ai}$ . Let  $\hat{\beta}_a$  and  $\hat{\beta}$  denote the 2SLS estimators constructed using the instrument sets  $\mathbf{z}_{ai}$  and  $(\mathbf{z}_{ai}, \mathbf{z}_{bi})$ , respectively. Without loss of generality we can assume that  $\mathbf{z}_{ai}$  and  $\mathbf{z}_{bi}$  are uncorrelated (if not, replace  $\mathbf{z}_{bi}$  with the projection error after projecting onto  $\mathbf{z}_{ai}$ ). In this case both  $\mathbb{E}(\mathbf{z}_i\mathbf{z}'_i)$  and  $(\mathbb{E}(\mathbf{z}_i\mathbf{z}'_i))^{-1}$  are block diagonal, so

$$\begin{aligned} \text{avar}(\hat{\beta}) &= \left(\mathbb{E}(\mathbf{x}_i\mathbf{z}'_i)(\mathbb{E}(\mathbf{z}_i\mathbf{z}'_i))^{-1}\mathbb{E}(\mathbf{z}_i\mathbf{x}'_i)\right)^{-1}\sigma^2 \\ &= \left(\mathbb{E}(\mathbf{x}_i\mathbf{z}'_{ai})(\mathbb{E}(\mathbf{z}_{ai}\mathbf{z}'_{ai}))^{-1}\mathbb{E}(\mathbf{z}_{ai}\mathbf{x}'_i) + \mathbb{E}(\mathbf{x}_i\mathbf{z}'_{bi})(\mathbb{E}(\mathbf{z}_{bi}\mathbf{z}'_{bi}))^{-1}\mathbb{E}(\mathbf{z}_{bi}\mathbf{x}'_i)\right)^{-1}\sigma^2 \\ &\leq \left(\mathbb{E}(\mathbf{x}_i\mathbf{z}'_{ai})(\mathbb{E}(\mathbf{z}_{ai}\mathbf{z}'_{ai}))^{-1}\mathbb{E}(\mathbf{z}_{ai}\mathbf{x}'_i)\right)^{-1}\sigma^2 \\ &= \text{avar}(\hat{\beta}_a) \end{aligned}$$

with strict inequality if  $\mathbb{E}(\mathbf{x}_i\mathbf{z}'_{bi}) \neq \mathbf{0}$ . Thus the 2SLS estimator with the full instrument set has a smaller asymptotic variance than the estimator with the smaller instrument set.

What we have shown is that the asymptotic variance of the 2SLS estimator is decreasing as the number of instruments increases. From the viewpoint of asymptotic efficiency, this means that it is better to use more instruments (when they are available and are all known to be valid instruments) rather than less.

Unfortunately, there is always a catch. In this case it turns out that the finite sample bias of the 2SLS estimator (which cannot be calculated exactly, but can be approximated using asymptotic expansions) is generically increasing linearly as the number of instruments increases. We will see some calculations illustrating this phenomenon in Section 12.37. Thus the choice of instruments in practice induces a trade-off between bias and variance.

## 12.18 Covariance Matrix Estimation

Estimation of the asymptotic variance matrix  $V_{\beta}$  is done using similar techniques as for least-squares estimation. The estimator is constructed by replacing the population moment matrices by sample counterparts. Thus

$$\hat{V}_{\beta} = \left(\hat{\mathbf{Q}}_{xz}\hat{\mathbf{Q}}_{zz}^{-1}\hat{\mathbf{Q}}_{zx}\right)^{-1}\left(\hat{\mathbf{Q}}_{xz}\hat{\mathbf{Q}}_{zz}^{-1}\hat{\Omega}\hat{\mathbf{Q}}_{zz}^{-1}\hat{\mathbf{Q}}_{zx}\right)\left(\hat{\mathbf{Q}}_{xz}\hat{\mathbf{Q}}_{zz}^{-1}\hat{\mathbf{Q}}_{zx}\right)^{-1} \quad (12.42)$$

where

$$\begin{aligned}\hat{\mathbf{Q}}_{zz} &= \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{z}'_i = \frac{1}{n} \mathbf{Z}' \mathbf{Z} \\ \hat{\mathbf{Q}}_{xz} &= \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{z}'_i = \frac{1}{n} \mathbf{X}' \mathbf{Z} \\ \hat{\boldsymbol{\Omega}} &= \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{z}'_i \hat{e}_i^2 \\ \hat{e}_i &= y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{2\text{sls}}.\end{aligned}$$

The homoskedastic variance matrix can be estimated by

$$\begin{aligned}\hat{V}_{\boldsymbol{\beta}}^0 &= (\hat{\mathbf{Q}}_{xz} \hat{\mathbf{Q}}_{zz}^{-1} \hat{\mathbf{Q}}_{zx})^{-1} \hat{\sigma}^2 \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n \hat{e}_i^2.\end{aligned}$$

Standard errors for the coefficients are obtained as the square roots of the diagonal elements of  $n^{-1} \hat{V}_{\boldsymbol{\beta}}$ . Confidence intervals, t-tests, and Wald tests may all be constructed from the coefficient estimates and covariance matrix estimate exactly as for least-squares regression.

In Stata, the `ivregress` command by default calculates the covariance matrix estimator using the homoskedastic variance matrix. To obtain covariance matrix estimation and standard errors with the robust estimator  $\hat{V}_{\boldsymbol{\beta}}$ , use the “, r” option.

**Theorem 12.3** Under Assumption 12.2, as  $n \rightarrow \infty$ ,

$$\hat{V}_{\boldsymbol{\beta}}^0 \xrightarrow{p} V_{\boldsymbol{\beta}}^0$$

$$\hat{V}_{\boldsymbol{\beta}} \xrightarrow{p} V_{\boldsymbol{\beta}}.$$

To prove Theorem 12.3 the key is to show  $\hat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$  as the other convergence results were established in the proof of consistency. We defer this to Exercise 12.6.

It is important that the covariance matrix be constructed using the correct residual formula  $\hat{e}_i = y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{2\text{sls}}$ . This is different than what would be obtained if the “two-stage” computation method is used. To see this, let’s walk through the two-stage method. First, we estimate the reduced form

$$\mathbf{x}_i = \hat{\boldsymbol{\Gamma}}' \mathbf{z}_i + \hat{\boldsymbol{u}}_i$$

to obtain the predicted values  $\hat{\mathbf{x}}_i = \hat{\boldsymbol{\Gamma}}' \mathbf{z}_i$ . Second, we regress  $y_i$  on  $\hat{\mathbf{x}}_i$  to obtain the 2SLS estimator  $\hat{\boldsymbol{\beta}}_{2\text{sls}}$ . This latter regression takes the form

$$y_i = \hat{\mathbf{x}}'_i \hat{\boldsymbol{\beta}}_{2\text{sls}} + \hat{v}_i \tag{12.43}$$

where  $\hat{v}_i$  are least-squares residuals. The covariance matrix (and standard errors) reported by this regression are constructed using the residual  $\hat{v}_i$ . For example, the homoskedastic formula is

$$\begin{aligned}\hat{V}_{\boldsymbol{\beta}} &= \left( \frac{1}{n} \hat{\mathbf{X}}' \hat{\mathbf{X}} \right)^{-1} \hat{\sigma}_v^2 = (\hat{\mathbf{Q}}_{xz} \hat{\mathbf{Q}}_{zz}^{-1} \hat{\mathbf{Q}}_{zx})^{-1} \hat{\sigma}_v^2 \\ \hat{\sigma}_v^2 &= \frac{1}{n} \sum_{i=1}^n \hat{v}_i^2\end{aligned}$$

which is proportional to the variance estimate  $\hat{\sigma}_v^2$  rather than  $\hat{\sigma}^2$ . This is important because the residual  $\hat{v}_i$  differs from  $\hat{e}_i$ . We can see this because the regression (12.43) uses the regressor  $\hat{x}_i$  rather than  $x_i$ . Indeed, we can calculate that

$$\begin{aligned}\hat{v}_i &= y_i - \mathbf{x}'_i \hat{\beta}_{\text{2sls}} + (\mathbf{x}_i - \hat{\mathbf{x}}_i)' \hat{\beta}_{\text{2sls}} \\ &= \hat{e}_i + \hat{\mathbf{u}}'_i \hat{\beta}_{\text{2sls}} \\ &\neq \hat{e}_i.\end{aligned}$$

This means that standard errors reported by the regression (12.43) will be incorrect.

This problem is avoided if the 2SLS estimator is constructed directly and the standard errors calculated with the correct formula rather than taking the “two-step” shortcut.

## 12.19 LIML Asymptotic Distribution

In this section we show that the LIML estimator is asymptotically equivalent to the 2SLS estimator. We recommend, however, a different covariance matrix estimator based on the IV representation.

We start by deriving the asymptotic distribution. Recall that the LIML estimator has several representations, including

$$\hat{\beta}_{\text{liml}} = (\mathbf{X}'(\mathbf{I}_n - \hat{\kappa} \mathbf{M}_Z) \mathbf{X})^{-1} (\mathbf{X}'(\mathbf{I}_n - \hat{\kappa} \mathbf{M}_Z) \mathbf{y})$$

where

$$\hat{\kappa} = \min_{\gamma} \frac{\gamma' Y' M_1 Y \gamma}{\gamma' Y' M_Z Y \gamma}.$$

For the distribution theory, it is useful to rewrite this as

$$\hat{\beta}_{\text{liml}} = (\mathbf{X}' P_Z \mathbf{X} - \hat{\mu} \mathbf{X}' M_Z \mathbf{X})^{-1} (\mathbf{X}' P_Z \mathbf{y} - \hat{\mu} \mathbf{X}' M_Z \mathbf{y})$$

where

$$\hat{\mu} = \hat{\kappa} - 1 = \min_{\gamma} \frac{\gamma' Y' M_1 Z_2 (Z_2' M_1 Z_2)^{-1} Z_2' M_1 Y \gamma}{\gamma' Y' M_Z Y \gamma}.$$

This second equality holds since the span of  $Z = [Z_1, Z_2]$  equals the span of  $[Z_1, M_1 Z_2]$ . This implies

$$\begin{aligned}P_Z &= Z (Z' Z)^{-1} Z' \\ &= Z_1 (Z_1' Z_1)^{-1} Z_1' + M_1 Z_2 (Z_2' M_1 Z_2)^{-1} Z_2' M_1.\end{aligned}$$

We now show that  $n\hat{\mu} = O_p(1)$ . The reduced form (12.35) implies that

$$Y = Z_1 \Pi_1 + Z_2 \Pi_2 + e.$$

It will be important to note that

$$\Pi_2 = [\lambda_2, \Gamma_{22}] = [\Gamma_{22} \beta_2, \Gamma_{22}]$$

using (12.18). It follows that  $\Pi_2 \bar{\gamma} = 0$  for  $\bar{\gamma} = (1, -\beta_2')'$ . Note  $u \bar{\gamma} = e$ . Then  $M_Z Y \bar{\gamma} = M_Z e$  and  $M_1 Y \bar{\gamma} = M_1 e$ . Hence

$$\begin{aligned}n\hat{\mu} &= \min_{\gamma} \frac{\gamma' Y' M_1 Z_2 (Z_2' M_1 Z_2)^{-1} Z_2' M_1 Y \gamma}{\gamma' \frac{1}{n} Y' M_Z Y \gamma} \\ &\leq \frac{\left(\frac{1}{\sqrt{n}} e' M_1 Z_2\right) \left(\frac{1}{n} Z_2' M_1 Z_2\right)^{-1} \left(\frac{1}{\sqrt{n}} Z_2' M_1 e\right)}{\frac{1}{n} e' M_Z e} \\ &= O_p(1).\end{aligned}$$

It follows that

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\text{liml}} - \boldsymbol{\beta}) &= \left( \frac{1}{n} \mathbf{X}' \mathbf{P}_Z \mathbf{X} - \widehat{\mu} \frac{1}{n} \mathbf{X}' \mathbf{M}_Z \mathbf{X} \right)^{-1} \left( \frac{1}{\sqrt{n}} \mathbf{X}' \mathbf{P}_Z \mathbf{e} - \sqrt{n} \widehat{\mu} \frac{1}{n} \mathbf{X}' \mathbf{M}_Z \mathbf{e} \right) \\ &= \left( \frac{1}{n} \mathbf{X}' \mathbf{P}_Z \mathbf{X} - o_p(1) \right)^{-1} \left( \frac{1}{\sqrt{n}} \mathbf{X}' \mathbf{P}_Z \mathbf{e} - o_p(1) \right) \\ &= \sqrt{n}(\widehat{\boldsymbol{\beta}}_{\text{2sls}} - \boldsymbol{\beta}) + o_p(1)\end{aligned}$$

which means that LIML and 2SLS have the same asymptotic distribution. This holds under the same assumptions as for 2SLS, and in particular does not require normality of the errors.

Consequently, one method to obtain an asymptotically valid covariance estimate for LIML is to use the same formula as for 2SLS. However, this is not the best choice. Rather, consider the IV representation for LIML

$$\widehat{\boldsymbol{\beta}}_{\text{liml}} = (\tilde{\mathbf{X}}' \tilde{\mathbf{X}})^{-1} (\tilde{\mathbf{X}}' \mathbf{y})$$

where

$$\tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 - \widehat{\kappa} \widehat{\mathbf{U}}_2 \end{pmatrix}$$

and  $\widehat{\mathbf{U}}_2 = \mathbf{M}_Z \mathbf{X}_2$ . The asymptotic covariance matrix formula for an IV estimator is

$$\widehat{V}_{\boldsymbol{\beta}} = \left( \frac{1}{n} \tilde{\mathbf{X}}' \tilde{\mathbf{X}} \right)^{-1} \widehat{\boldsymbol{\Omega}} \left( \frac{1}{n} \mathbf{X}' \tilde{\mathbf{X}} \right)^{-1} \quad (12.44)$$

where

$$\begin{aligned}\widehat{\boldsymbol{\Omega}} &= \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i \tilde{e}_i^2 \\ \tilde{e}_i &= y_i - \mathbf{x}'_i \widehat{\boldsymbol{\beta}}_{\text{liml}}.\end{aligned}$$

This simplifies to the 2SLS formula when  $\widehat{\kappa} = 1$  but otherwise differs. The estimator (12.44) is a better choice than the 2SLS formula for covariance matrix estimation as it takes advantage of the LIML estimator structure.

## 12.20 Functions of Parameters

Given the distribution theory in Theorems 12.2 and 12.3 it is straightforward to derive the asymptotic distribution of smooth nonlinear functions of the coefficients.

Specifically, given a function  $\mathbf{r}(\boldsymbol{\beta}) : \mathbb{R}^k \rightarrow \Theta \subset \mathbb{R}^q$  we define the parameter

$$\boldsymbol{\theta} = \mathbf{r}(\boldsymbol{\beta}).$$

Given  $\widehat{\boldsymbol{\beta}}_{\text{2sls}}$  a natural estimator of  $\boldsymbol{\theta}$  is  $\widehat{\boldsymbol{\theta}}_{\text{2sls}} = \mathbf{r}(\widehat{\boldsymbol{\beta}}_{\text{2sls}})$ .

Consistency follows from Theorem 12.1 and the continuous mapping theorem.

**Theorem 12.4** Under Assumptions 12.1 and 7.3, as  $n \rightarrow \infty$ ,  $\widehat{\boldsymbol{\theta}}_{\text{2sls}} \xrightarrow{p} \boldsymbol{\theta}$ .

If  $\mathbf{r}(\boldsymbol{\beta})$  is differentiable then an estimator of the asymptotic covariance matrix for  $\widehat{\boldsymbol{\theta}}$  is

$$\begin{aligned}\widehat{V}_{\boldsymbol{\theta}} &= \widehat{\mathbf{R}}' \widehat{V}_{\boldsymbol{\beta}} \widehat{\mathbf{R}} \\ \widehat{\mathbf{R}} &= \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{r}(\widehat{\boldsymbol{\beta}}_{\text{2sls}})'.\end{aligned}$$

We similarly define the homoskedastic variance estimator as

$$\widehat{V}_{\boldsymbol{\theta}}^0 = \widehat{\mathbf{R}}' \widehat{V}_{\boldsymbol{\beta}}^0 \widehat{\mathbf{R}}.$$

The asymptotic distribution theory follows from Theorems 12.2 and 12.3 and the delta method.

**Theorem 12.5** Under Assumptions 12.2 and 7.3, as  $n \rightarrow \infty$ ,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{2\text{sls}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, V_{\boldsymbol{\theta}})$$

where

$$V_{\boldsymbol{\theta}} = \mathbf{R}' V_{\boldsymbol{\beta}} \mathbf{R}$$

$$\mathbf{R} = \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{r}(\boldsymbol{\beta})'$$

and

$$\widehat{V}_{\boldsymbol{\theta}} \xrightarrow{p} V_{\boldsymbol{\theta}}.$$

When  $q = 1$ , a standard error for  $\widehat{\boldsymbol{\theta}}_{2\text{sls}}$  is  $s(\widehat{\boldsymbol{\theta}}_{2\text{sls}}) = \sqrt{n^{-1} \widehat{V}_{\boldsymbol{\theta}}}$ .

For example, let's take the parameter estimates from the fifth column of Table 12.1, which are the 2SLS estimates with three endogenous regressors and four excluded instruments. Suppose we are interested in the return to experience, which depends on the level of experience. The estimated return at  $experience = 10$  is  $0.047 - 0.032 * 2 * 10/100 = 0.041$  and its standard error is 0.003. This implies a 4% increase in wages per year of experience and is precisely estimated. Or suppose we are interested in the level of experience at which the function maximizes. The estimate is  $50 * 0.047 / 0.032 = 73$ . This has a standard error of 249. The large standard error implies that the estimate (73 years of experience) is without precision and is thus uninformative.

## 12.21 Hypothesis Tests

As in the previous section, for a given function  $\mathbf{r}(\boldsymbol{\beta}) : \mathbb{R}^k \rightarrow \Theta \subset \mathbb{R}^q$  we define the parameter  $\boldsymbol{\theta} = \mathbf{r}(\boldsymbol{\beta})$  and consider tests of hypotheses of the form

$$\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$$

against

$$\mathbb{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

The Wald statistic for  $\mathbb{H}_0$  is

$$W = n(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \widehat{V}_{\boldsymbol{\theta}}^{-1} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

From Theorem 12.5 we deduce that  $W$  is asymptotically chi-square distributed. Let  $G_q(u)$  denote the  $\chi_q^2$  distribution function.

**Theorem 12.6** Under Assumptions 12.2 and 7.3 and  $\mathbb{H}_0$  holds, then as  $n \rightarrow \infty$ ,

$$W \xrightarrow{d} \chi_q^2.$$

For  $c$  satisfying  $\alpha = 1 - G_q(c)$ ,

$$\mathbb{P}(W > c | \mathbb{H}_0) \longrightarrow \alpha$$

so the test “Reject  $\mathbb{H}_0$  if  $W > c$ ” has asymptotic size  $\alpha$ .

In linear regression we often report the  $F$  version of the Wald statistic (by dividing by degrees of freedom) and use the  $F$  distribution for inference, as this is justified in the normal sampling model. For 2SLS estimation, however, this is not done as there is no finite sample  $F$  justification for the  $F$  version of the Wald statistic.

To illustrate, once again let’s take the parameter estimates from the fifth column of Table 12.1 and again consider the return to experience which is determined by the coefficients on *experience* and *experience*<sup>2</sup>/100. Neither coefficient is statistically significant at the 5% level and it is unclear if the overall effect is statistically significant. We can assess this by testing the joint hypothesis that both coefficients are zero. The Wald statistic for this hypothesis is  $W = 244$ , which is highly significant with an asymptotic p-value of 0.0000. Thus by examining the joint test in contrast to the individual tests is quite clear that experience has a non-zero effect.

## 12.22 Finite Sample Theory

In Chapter 5 we reviewed the rich exact distribution available for the linear regression model under the assumption of normal innovations. There was a similarly rich literature in econometrics which developed a distribution theory for IV, 2SLS and LIML estimators. An excellent review of the theory, mostly developed in the 1970s and early 1980s, is reviewed by Peter Phillips (1983).

This theory was developed under the assumption that the structural error vector  $\mathbf{e}$  and reduced form error  $\mathbf{u}_2$  are multivariate normally distributed. Even though the errors are normal, IV-type estimators are non-linear functions of these errors and are thus the estimators non-normally distributed. Formulae for the exact distributions have been derived, but are unfortunately functions of model parameters and hence are not directly useful for finite sample inference.

One important implication of this literature is that it is quite clear that even in this optimal context of exact normal innovations, the finite sample distributions of the IV estimators are non-normal and the finite sample distributions of test statistics are not chi-squared. The normal and chi-squared approximations hold asymptotically, but there is no reason to expect these approximations to be accurate in finite samples.

A second important result is that under the assumption of normal errors, most of the estimators do not have finite moments in any finite sample. A clean statement concerning the existence of moments for the 2SLS estimator was obtained by Kinal (1980) for the case of joint normality. Let  $\hat{\boldsymbol{\beta}}_{2\text{sls},2}$  be the 2SLS estimators of the coefficients on the endogenous regressors.

**Theorem 12.7** If  $(y_i, \mathbf{x}_i, \mathbf{z}_i)$  are jointly normal, then for any  $r$ ,  $\mathbb{E} \|\hat{\boldsymbol{\beta}}_{2\text{sls},2}\|^r < \infty$  if and only if  $r < \ell_2 - k_2 + 1$ .

This result states that in the just-identified case the IV estimator does not have any finite order integer moments. In the over-identified case the number of finite moments corresponds to the number of

overidentifying restrictions ( $\ell_2 - k_2$ ). Thus if there is one over-identifying restriction the 2SLS estimator has a finite mean, and if there are two over-identifying restrictions then the 2SLS estimator has a finite variance.

The LIML estimator has a more severe moment problem, as it has no finite integer moments (Mariano, 1982) regardless of the number of over-identifying restrictions. Due to this lack of moments, Fuller (1977) proposed the following modification of LIML. Instead of (12.39), Fuller's estimator is

$$\begin{aligned}\hat{\beta}_{\text{Fuller}} &= (\mathbf{X}'(\mathbf{I}_n - K\mathbf{M}_Z)\mathbf{X})^{-1}(\mathbf{X}'(\mathbf{I}_n - K\mathbf{M}_Z)\mathbf{y}) \\ K &= \hat{\kappa} - \frac{C}{n-k}\end{aligned}$$

for some  $C \geq 1$ . Fuller showed that his estimator has all moments finite under suitable conditions.

Hausman, Newey, Woutersen, Chao and Swanson (2012) propose an estimator they call HFUL which combines the ideas of JIVE and Fuller which has excellent finite sample properties.

## 12.23 Bootstrap for 2SLS

The standard bootstrap algorithm for IV, 2SLS and GMM generates bootstrap samples by sampling the triplets  $(y_i^*, \mathbf{x}_i^*, \mathbf{z}_i^*)$  independently and with replacement from the original sample  $\{(y_i, \mathbf{x}_i, \mathbf{z}_i) : i = 1, \dots, n\}$ . Sampling  $n$  such observations and stacking into observation matrices  $(\mathbf{y}^*, \mathbf{X}^*, \mathbf{Z}^*)$ , the bootstrap 2SLS estimator is

$$\hat{\beta}_{\text{2sls}}^* = (\mathbf{X}^{*'} \mathbf{Z}^* (\mathbf{Z}^{*'} \mathbf{Z}^*)^{-1} \mathbf{Z}^{*'} \mathbf{X}^*)^{-1} \mathbf{X}^{*'} \mathbf{Z}^* (\mathbf{Z}^{*'} \mathbf{Z}^*)^{-1} \mathbf{Z}^{*'} \mathbf{y}^*.$$

This is repeated  $B$  times to create a sample of  $B$  bootstrap draws. Given these draws, bootstrap statistics can be calculated. This includes the bootstrap estimate of variance, standard errors, and confidence intervals, including percentile, BC percentile,  $\text{BC}_a$  and percentile-t.

We now show that the bootstrap estimator has the same asymptotic distribution as the sample estimator. For overidentified cases this demonstration requires a bit of extra care. This was first shown by Hahn (1996).

The sample observations satisfy the model

$$\begin{aligned}y_i &= \mathbf{x}_i' \boldsymbol{\beta} + e_i \\ \mathbb{E}(\mathbf{z}_i e_i) &= 0.\end{aligned}$$

The true value of  $\boldsymbol{\beta}$  in the population can be written as

$$\boldsymbol{\beta} = \left( \mathbb{E}(\mathbf{x}_i \mathbf{z}_i') \mathbb{E}(\mathbf{z}_i \mathbf{z}_i')^{-1} \mathbb{E}(\mathbf{z}_i \mathbf{x}_i') \right)^{-1} \mathbb{E}(\mathbf{x}_i \mathbf{z}_i') \mathbb{E}(\mathbf{z}_i \mathbf{z}_i')^{-1} \mathbb{E}(\mathbf{z}_i y_i).$$

The true value in the bootstrap universe is obtained by replacing the population moments by the sample moments, which equals the 2SLS estimator

$$\begin{aligned}&\left( \mathbb{E}^*(\mathbf{x}_i^* \mathbf{z}_i^{*'}) \mathbb{E}^*(\mathbf{z}_i^* \mathbf{z}_i^{*'})^{-1} \mathbb{E}^*(\mathbf{z}_i^* \mathbf{x}_i^{*'}) \right)^{-1} \mathbb{E}^*(\mathbf{x}_i^* \mathbf{z}_i^{*'}) \mathbb{E}^*(\mathbf{z}_i^* \mathbf{z}_i^{*'})^{-1} \mathbb{E}^*(\mathbf{z}_i^* y_i^*) \\&= \left( \left( \frac{1}{n} \mathbf{X}' \mathbf{Z} \right) \left( \frac{1}{n} \mathbf{Z}' \mathbf{Z} \right)^{-1} \left( \frac{1}{n} \mathbf{Z}' \mathbf{X} \right) \right)^{-1} \left( \frac{1}{n} \mathbf{X}' \mathbf{Z} \right) \left( \frac{1}{n} \mathbf{Z}' \mathbf{Z} \right)^{-1} \left[ \frac{1}{n} \mathbf{Z}' \mathbf{y} \right] \\&= \hat{\beta}_{\text{2sls}}.\end{aligned}$$

The bootstrap observations thus satisfy the equation

$$y_i^* = \mathbf{x}_i^{*'} \hat{\beta}_{\text{2sls}} + e_i^*.$$

In matrix notation

$$\mathbf{y}^* = \mathbf{X}^{*'} \hat{\beta}_{\text{2sls}} + \mathbf{e}^*. \quad (12.45)$$

Given a bootstrap triple  $(y_i^*, \mathbf{x}_i^*, \mathbf{z}_i^*) = (y_j, \mathbf{x}_j, \mathbf{z}_j)$  for some observation  $j$ , the true bootstrap error is

$$e_i^* = y_j - \mathbf{x}_j' \hat{\beta}_{2\text{sls}} = \hat{e}_j.$$

It follows that

$$\mathbb{E}^*(\mathbf{z}_i^* e_i^*) = n^{-1} \mathbf{Z}' \hat{\mathbf{e}}. \quad (12.46)$$

This is generally not equal to zero in the over-identified case.

This is an important complication. In over-identified models the true observations satisfy the population condition  $\mathbb{E}(\mathbf{z}_i e_i) = 0$  but in the bootstrap sample  $\mathbb{E}^*(\mathbf{z}_i^* e_i^*) \neq 0$ . This means that to apply the central limit theorem to the bootstrap estimator we will first have to recenter the moment condition. That is, (12.46) and the bootstrap CLT imply

$$\frac{1}{\sqrt{n}} (\mathbf{Z}^{*\prime} \mathbf{e}^* - \mathbf{Z}' \hat{\mathbf{e}}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{z}_i^* e_i^* - \mathbb{E}^*(\mathbf{z}_i^* e_i^*)) \xrightarrow{d^*} N(\mathbf{0}, \Omega) \quad (12.47)$$

where

$$\Omega = \mathbb{E}(\mathbf{z}_i \mathbf{z}_i' e_i^2).$$

Using (12.45) we can normalize the bootstrap estimator as

$$\begin{aligned} \sqrt{n} (\hat{\beta}_{2\text{sls}}^* - \hat{\beta}_{2\text{sls}}) &= \sqrt{n} \left( \mathbf{X}^{*\prime} \mathbf{Z}^* (\mathbf{Z}^{*\prime} \mathbf{Z}^*)^{-1} \mathbf{Z}^{*\prime} \mathbf{X}^* \right)^{-1} \mathbf{X}^{*\prime} \mathbf{Z}^* (\mathbf{Z}^{*\prime} \mathbf{Z}^*)^{-1} \mathbf{Z}^{*\prime} \mathbf{e}^* \\ &= \left( \left( \frac{1}{n} \mathbf{X}^{*\prime} \mathbf{Z}^* \right) \left( \frac{1}{n} \mathbf{Z}^{*\prime} \mathbf{Z}^* \right)^{-1} \left( \frac{1}{n} \mathbf{Z}^{*\prime} \mathbf{X}^* \right) \right)^{-1} \\ &\quad \cdot \left( \frac{1}{n} \mathbf{X}^{*\prime} \mathbf{Z}^* \right) \left( \frac{1}{n} \mathbf{Z}^{*\prime} \mathbf{Z}^* \right)^{-1} \frac{1}{\sqrt{n}} (\mathbf{Z}^{*\prime} \mathbf{e}^* - \mathbf{Z}' \hat{\mathbf{e}}) \end{aligned} \quad (12.48)$$

$$\begin{aligned} &+ \left( \left( \frac{1}{n} \mathbf{X}^{*\prime} \mathbf{Z}^* \right) \left( \frac{1}{n} \mathbf{Z}^{*\prime} \mathbf{Z}^* \right)^{-1} \left( \frac{1}{n} \mathbf{Z}^{*\prime} \mathbf{X}^* \right) \right)^{-1} \\ &\quad \cdot \left( \frac{1}{n} \mathbf{X}^{*\prime} \mathbf{Z}^* \right) \left( \frac{1}{n} \mathbf{Z}^{*\prime} \mathbf{Z}^* \right)^{-1} \left( \frac{1}{\sqrt{n}} \mathbf{Z}' \hat{\mathbf{e}} \right). \end{aligned} \quad (12.49)$$

Using the bootstrap WLLN,

$$\begin{aligned} \frac{1}{n} \mathbf{X}^{*\prime} \mathbf{Z}^* &= \frac{1}{n} \mathbf{X}' \mathbf{Z} + o_p(1) \\ \frac{1}{n} \mathbf{Z}^{*\prime} \mathbf{Z}^* &= \frac{1}{n} \mathbf{Z}' \mathbf{Z} + o_p(1). \end{aligned}$$

This implies (12.49) is equal to

$$\sqrt{n} \left( \mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{X}) \right)^{-1} \mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \hat{\mathbf{e}} + o_p(1) = 0 + o_p(1).$$

The equality holds because the 2SLS first-order condition implies  $\mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \hat{\mathbf{e}} = 0$ . Also, combined with (12.47) we see that (12.48) converges in bootstrap distribution to

$$(\mathbf{Q}_{xz} \mathbf{Q}_{zz}^{-1} \mathbf{Q}_{zx})^{-1} \mathbf{Q}_{xz} \mathbf{Q}_{zz}^{-1} N(\mathbf{0}, \Omega) = N(\mathbf{0}, V_\beta)$$

where  $V_\beta$  is the 2SLS asymptotic variance from Theorem 12.2. This is the asymptotic distribution of  $\sqrt{n} (\hat{\beta}_{2\text{sls}}^* - \hat{\beta}_{2\text{sls}})$ .

By standard calculations we can also show that bootstrap t-ratios are asymptotically normal.

**Theorem 12.8** Under Assumption 12.2, as  $n \rightarrow \infty$

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{2\text{SLS}}^* - \widehat{\boldsymbol{\beta}}_{2\text{SLS}}) \xrightarrow{d^*} N(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}})$$

where  $\mathbf{V}_{\boldsymbol{\beta}}$  is the 2SLS asymptotic variance from Theorem 12.2. Furthermore,

$$T^* = \frac{\sqrt{n}(\widehat{\boldsymbol{\beta}}_{2\text{SLS}}^* - \widehat{\boldsymbol{\beta}}_{2\text{SLS}})}{s(\widehat{\boldsymbol{\beta}}_{2\text{SLS}}^*)} \xrightarrow{d^*} N(0, 1).$$

This shows that percentile-type and percentile-t confidence intervals are asymptotically valid.

One might expect that the asymptotic refinement arguments extend to the BC<sub>a</sub> and percentile-t methods, but this does not appear to be the case. While  $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{2\text{SLS}}^* - \widehat{\boldsymbol{\beta}}_{2\text{SLS}})$  and  $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{2\text{SLS}} - \boldsymbol{\beta})$  have the same asymptotic distribution, they differ in finite samples by an  $O_p(n^{-1/2})$  term. This means that they have distinct Edgeworth expansions. Consequently, unadjusted bootstrap methods will not achieve an asymptotic refinement.

An alternative suggested by Hall and Horowitz (1996) is to recenter the bootstrap 2SLS estimator so that it satisfies the correct orthogonality condition. Define

$$\widehat{\boldsymbol{\beta}}_{2\text{SLS}}^{**} = \left( \mathbf{X}^{*'} \mathbf{Z}^* (\mathbf{Z}^{*'} \mathbf{Z}^*)^{-1} \mathbf{Z}^{*'} \mathbf{X}^* \right)^{-1} \mathbf{X}^{*'} \mathbf{Z}^* (\mathbf{Z}^{*'} \mathbf{Z}^*)^{-1} (\mathbf{Z}^{*'} \mathbf{y}^* - \mathbf{Z}' \widehat{\mathbf{e}}).$$

We can see that

$$\begin{aligned} \sqrt{n}(\widehat{\boldsymbol{\beta}}_{2\text{SLS}}^{**} - \widehat{\boldsymbol{\beta}}_{2\text{SLS}}) &= \left( \frac{1}{n} \mathbf{X}^{*'} \mathbf{Z}^* \left( \frac{1}{n} \mathbf{Z}^{*'} \mathbf{Z}^* \right)^{-1} \frac{1}{n} \mathbf{Z}^{*'} \mathbf{X}^* \right)^{-1} \\ &\quad \cdot \left( \frac{1}{n} \mathbf{X}^{*'} \mathbf{Z}^* \right) \left( \frac{1}{n} \mathbf{Z}^{*'} \mathbf{Z}^* \right)^{-1} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{z}_i^* \mathbf{e}_i^* - \mathbb{E}^*(\mathbf{z}_i^* \mathbf{e}_i^*)) \right) \end{aligned}$$

which directly converges to the  $N(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}})$  distribution without special handling. Hall and Horowitz (1996) show that percentile-t methods applied to  $\widehat{\boldsymbol{\beta}}_{2\text{SLS}}^{**}$  achieve an asymptotic refinement and are thus preferred to the unadjusted bootstrap estimator.

This recentered estimator, however, is not the standard implementation of the bootstrap for 2SLS as used in empirical practice.

## 12.24 The Peril of Bootstrap 2SLS Standard Errors

It is tempting to use the bootstrap algorithm to estimate variance matrices and standard errors for the 2SLS estimator. In fact this is one of the most common use of bootstrap methods in current econometric practice. Unfortunately this is an unjustified and ill-conceived idea and should not be done. In finite samples the 2SLS estimator may not have a finite second moment, meaning that bootstrap variance estimates are unstable and unreliable.

Theorem 12.7 shows that under jointly normality the 2SLS estimator will have a finite variance if and only if the number of overidentifying restrictions is two or larger. Thus for just-identified IV, and 2SLS with one degree of overidentification, the finite sample variance is infinite. The bootstrap will be attempting to estimate this value – infinity – and will yield nonsensical answers. When the observations are not jointly normal there is no finite sample theory (so it is possible that the finite sample variance is actually finite) but this is unknown and unverifiable.

In overidentified settings when the number of overidentifying restrictions is two or larger the bootstrap can be applied for standard error estimation. However this is not the most common application of IV methods in econometric practice and thus should be viewed as the exception rather than the norm.

To understand what is going on, consider the simplest case of a just-identified model with a single endogenous regressor and no included exogeneous regressors. In this case the estimator can be written as a ratio of means

$$\hat{\beta}_{\text{iv}} - \beta = \frac{\sum_{i=1}^n z_i e_i}{\sum_{i=1}^n z_i x_i}.$$

Under joint normality of  $(e_i, x_i)$ , this has a Cauchy-like distribution which does not possess any finite integer moments. The trouble is that the denominator can be either positive or negative, and arbitrarily close to zero. This means that the ratio can take arbitrarily large values.

To illustrate let us return to the basic Card IV wage regression from column 2 of Table 12.1 which uses college as an *instrument* for *education*. Estimate this equation for the subsample of black men, which has  $n = 703$  observations. We focus on the coefficient for the return to education. The coefficient estimate is reported in Table 12.3, along with asymptotic, jackknife, and two bootstrap standard errors each calculated with 10,000 bootstrap replications.

Table 12.3: Instrumental Variable Return to Education for Black Men

Estimate	0.11
Asymptotic s.e.	(0.11)
Jackknife s.e.	(0.11)
Bootstrap s.e. (standard)	(1.42)
Bootstrap s.e. (repeat)	(4.79)

The bootstrap standard errors are an order of magnitude larger than the asymptotic standard errors, and vary substantially across the bootstrap runs despite using 10,000 bootstrap replications. This indicates moment failure and unreliability of the bootstrap standard errors.

This is a strong message that **bootstrap standard errors should not be computed for IV estimators**. Instead, report percentile-type confidence intervals.

## 12.25 Clustered Dependence

In Section 4.21 we introduced clustered dependence. We can also use the methods of clustered dependence for 2SLS estimation. Recall, the  $g^{th}$  cluster has the observations  $\mathbf{y}_g = (y_{1g}, \dots, y_{n_g g})'$ ,  $\mathbf{X}_g = (\mathbf{x}_{1g}, \dots, \mathbf{x}_{n_g g})'$ , and  $\mathbf{Z}_g = (\mathbf{z}_{1g}, \dots, \mathbf{z}_{n_g g})'$ . The structural equation for the  $g^{th}$  cluster can be written as the matrix system

$$\mathbf{y}_g = \mathbf{X}_g \boldsymbol{\beta} + \mathbf{e}_g.$$

Using this notation the centered 2SLS estimator can be written as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{2sls}} - \boldsymbol{\beta} &= \left( \mathbf{Z}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X} \right)^{-1} \mathbf{Z}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{e} \\ &= \left( \mathbf{Z}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X} \right)^{-1} \mathbf{Z}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \left( \sum_{g=1}^G \mathbf{Z}'_g \mathbf{e}_g \right).\end{aligned}$$

The cluster-robust covariance matrix estimator for  $\hat{\boldsymbol{\beta}}_{\text{2sls}}$  thus takes the form

$$\hat{\mathbf{V}}_{\boldsymbol{\beta}} = \left( \mathbf{Z}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X} \right)^{-1} \mathbf{Z}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \hat{\mathbf{S}} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X} \left( \mathbf{Z}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X} \right)^{-1}$$

with

$$\hat{\mathbf{S}} = \sum_{g=1}^G \mathbf{Z}'_g \hat{\mathbf{e}}_g \hat{\mathbf{e}}'_g \mathbf{Z}_g$$

and the clustered residuals

$$\hat{\mathbf{e}}_g = \mathbf{y}_g - \mathbf{X}_g \hat{\boldsymbol{\beta}}_{\text{2sls}}.$$

The difference between the heteroskedasticity-robust estimator and the cluster-robust estimator is the covariance estimator  $\hat{\mathbf{S}}$ .

## 12.26 Generated Regressors

The “two-stage” form of the 2SLS estimator is an example of what is called “estimation with generated regressors”. We say a regressor is a **generated** if it is an estimate of an idealized regressor, or if it is a function of estimated parameters. Typically, a generated regressor  $\hat{\mathbf{w}}_i$  is an estimate of an unobserved ideal regressor  $\mathbf{w}_i$ . As an estimate,  $\hat{\mathbf{w}}_i$  is a function of the sample, not just observation  $i$ . Hence it is not “i.i.d.” as it is dependent across observations, which invalidates the conventional regression assumptions. Consequently, the sampling distribution of regression estimates is affected. Unless this is incorporated into our inference methods, covariance matrix estimates and standard errors will be incorrect.

The econometric theory of generated regressors was developed by Pagan (1984) for linear models, and extended to non-linear models and more general two-step estimators by Pagan (1986). Independently, similar results were obtained by Murphy and Topel (1985). Here we focus on the linear model:

$$\begin{aligned} y_i &= \mathbf{w}'_i \boldsymbol{\beta} + \nu_i \\ \mathbf{w}_i &= \mathbf{A}' \mathbf{z}_i \\ \mathbb{E}(\mathbf{z}_i \nu_i) &= \mathbf{0}. \end{aligned} \tag{12.50}$$

The observables are  $(y_i, \mathbf{z}_i)$ . We also have an estimate  $\hat{\mathbf{A}}$  of  $\mathbf{A}$ .

Given  $\hat{\mathbf{A}}$  we construct the estimate  $\hat{\mathbf{w}}_i = \hat{\mathbf{A}}' \mathbf{z}_i$  of  $\mathbf{w}_i$ , replace  $\mathbf{w}_i$  in (12.50) with  $\hat{\mathbf{w}}_i$ , and then estimate  $\boldsymbol{\beta}$  by least-squares, resulting in the estimator

$$\hat{\boldsymbol{\beta}} = \left( \sum_{i=1}^n \hat{\mathbf{w}}_i \hat{\mathbf{w}}'_i \right)^{-1} \left( \sum_{i=1}^n \hat{\mathbf{w}}_i y_i \right). \tag{12.51}$$

The regressors  $\hat{\mathbf{w}}_i$  are called **generated regressors**. The properties of  $\hat{\boldsymbol{\beta}}$  are different than least-squares with i.i.d. observations, since the generated regressors are themselves estimates.

This framework includes the 2SLS estimator as well as other common estimators. The 2SLS model can be written as (12.50) by looking at the reduced form equation (12.14), with  $\mathbf{w}_i = \boldsymbol{\Gamma}' \mathbf{z}_i$ ,  $\mathbf{A} = \boldsymbol{\Gamma}$ , and  $\hat{\mathbf{A}} = \hat{\boldsymbol{\Gamma}}$  is (12.19).

The examples which motivated Pagan (1984) and Murphy and Topel (1985) emerged from the macroeconomics literature, in particular the work of Barro (1977) which examined the impact of inflation expectations and expectation errors on economic output. For example, let  $\pi_i$  denote realized inflation and  $\mathbf{z}_i$  be the information available to economic agents. A model of inflation expectations sets  $w_i = \mathbb{E}(\pi_i | \mathbf{z}_i) = \boldsymbol{\gamma}' \mathbf{z}_i$  and a model of expectation error sets  $w_i = \pi_i - \mathbb{E}(\pi_i | \mathbf{z}_i) = \pi_i - \boldsymbol{\gamma}' \mathbf{z}_i$ . Since expectations and errors are not observed they are replaced in applications with the fitted values  $\hat{\mathbf{w}}_i = \hat{\boldsymbol{\gamma}}' \mathbf{z}_i$  or residuals  $\hat{\mathbf{w}}_i = \pi_i - \hat{\boldsymbol{\gamma}}' \mathbf{z}_i$  where  $\hat{\boldsymbol{\gamma}}$  is a coefficient estimate from a regression of  $\pi_i$  on  $\mathbf{z}_i$ .

The generated regressor framework includes all of these examples.

The goal is to obtain a distributional approximation for  $\hat{\boldsymbol{\beta}}$  in order to construct standard errors, confidence intervals and conduct tests. Start by substituting equation (12.50) into (12.51). We obtain

$$\hat{\boldsymbol{\beta}} = \left( \sum_{i=1}^n \hat{\mathbf{w}}_i \hat{\mathbf{w}}'_i \right)^{-1} \left( \sum_{i=1}^n \hat{\mathbf{w}}_i (\mathbf{w}'_i \boldsymbol{\beta} + \nu_i) \right).$$

Next, substitute  $\mathbf{w}'_i \boldsymbol{\beta} = \hat{\mathbf{w}}'_i \boldsymbol{\beta} + (\mathbf{w}_i - \hat{\mathbf{w}}_i)' \boldsymbol{\beta}$ . We obtain

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left( \sum_{i=1}^n \hat{\mathbf{w}}_i \hat{\mathbf{w}}'_i \right)^{-1} \left( \sum_{i=1}^n \hat{\mathbf{w}}_i ((\mathbf{w}_i - \hat{\mathbf{w}}_i)' \boldsymbol{\beta} + \nu_i) \right). \tag{12.52}$$

Effectively, this shows that the distribution of  $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$  has two random components, one due to the conventional regression component  $\hat{\mathbf{w}}_i \nu_i$ , and the second due to the generated regressor  $(\mathbf{w}_i - \hat{\mathbf{w}}_i)' \boldsymbol{\beta}$ . Conventional variance estimators do not address this second component and thus will be biased.

Interestingly, the distribution in (12.52) dramatically simplifies in the special case that the “generated regressor term”  $(\mathbf{w}_i - \hat{\mathbf{w}}_i)' \boldsymbol{\beta}$  disappears. This occurs when the slope coefficients on the generated regressors are zero. To be specific, partition  $\mathbf{w}_i = (\mathbf{w}_{1i}, \mathbf{w}_{2i})$ ,  $\hat{\mathbf{w}}_i = (\mathbf{w}_{1i}, \hat{\mathbf{w}}_{2i})$ , and  $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$  so that  $\mathbf{w}_{1i}$  are the conventional observed regressors and  $\hat{\mathbf{w}}_{2i}$  are the generated regressors. Then  $(\mathbf{w}_i - \hat{\mathbf{w}}_i)' \boldsymbol{\beta} = (\mathbf{w}_{2i} - \hat{\mathbf{w}}_{2i})' \boldsymbol{\beta}_2$ . Thus if  $\boldsymbol{\beta}_2 = \mathbf{0}$  this term disappears. In this case (12.52) equals

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left( \sum_{i=1}^n \hat{\mathbf{w}}_i \hat{\mathbf{w}}_i' \right)^{-1} \left( \sum_{i=1}^n \hat{\mathbf{w}}_i v_i \right).$$

This is a dramatic simplification.

Furthermore, since  $\hat{\mathbf{w}}_i = \hat{\mathbf{A}}' \mathbf{z}_i$  we can write the estimator as a function of sample moments:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left( \hat{\mathbf{A}}' \left( \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i' \right) \hat{\mathbf{A}} \right)^{-1} \hat{\mathbf{A}}' \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{z}_i v_i \right).$$

If  $\hat{\mathbf{A}} \xrightarrow{p} \mathbf{A}$  we find from standard manipulations that

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, V_{\boldsymbol{\beta}})$$

where

$$V_{\boldsymbol{\beta}} = (\mathbf{A}' \mathbb{E}(\mathbf{z}_i \mathbf{z}_i') \mathbf{A})^{-1} (\mathbf{A}' \mathbb{E}(\mathbf{z}_i \mathbf{z}_i' v_i^2) \mathbf{A}) (\mathbf{A}' \mathbb{E}(\mathbf{z}_i \mathbf{z}_i') \mathbf{A})^{-1}. \quad (12.53)$$

The conventional asymptotic covariance matrix estimator for  $\hat{\boldsymbol{\beta}}$  takes the form

$$\hat{V}_{\boldsymbol{\beta}} = \left( \frac{1}{n} \sum_{i=1}^n \hat{\mathbf{w}}_i \hat{\mathbf{w}}_i' \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \hat{\mathbf{w}}_i \hat{\mathbf{w}}_i' \hat{v}_i^2 \right) \left( \frac{1}{n} \sum_{i=1}^n \hat{\mathbf{w}}_i \hat{\mathbf{w}}_i' \right)^{-1} \quad (12.54)$$

where  $\hat{v}_i = y_i - \hat{\mathbf{w}}_i' \hat{\boldsymbol{\beta}}$ . Under the given assumptions,  $\hat{V}_{\boldsymbol{\beta}} \xrightarrow{p} V_{\boldsymbol{\beta}}$ . Thus inference using  $\hat{V}_{\boldsymbol{\beta}}$  is asymptotically valid. This is useful when we are interested in tests of  $\boldsymbol{\beta}_2 = \mathbf{0}$ . Often this is of major interest in applications.

To test  $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$  we partition  $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$  and construct a conventional Wald statistic

$$W = n \hat{\boldsymbol{\beta}}_2' ([\hat{V}_{\boldsymbol{\beta}}]_{22})^{-1} \hat{\boldsymbol{\beta}}_2.$$

**Theorem 12.9** Take model (12.50) with  $\mathbb{E}(y_i^4) < \infty$ ,  $\mathbb{E}\|\mathbf{z}_i\|^4 < \infty$ ,  $\mathbf{A}' \mathbb{E}(\mathbf{z}_i \mathbf{z}_i') \mathbf{A} > 0$ ,  $\hat{\mathbf{A}} \xrightarrow{p} \mathbf{A}$  and  $\hat{\mathbf{w}}_i = (\mathbf{w}_{1i}, \hat{\mathbf{w}}_{2i})$ . Under  $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ , then as  $n \rightarrow \infty$ ,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, V_{\boldsymbol{\beta}})$$

where  $V_{\boldsymbol{\beta}}$  is given in (12.53). For  $\hat{V}_{\boldsymbol{\beta}}$  given in (12.54),

$$\hat{V}_{\boldsymbol{\beta}} \xrightarrow{p} V_{\boldsymbol{\beta}}.$$

Furthermore,

$$W \xrightarrow{d} \chi_q^2$$

where  $q = \dim(\boldsymbol{\beta}_2)$ . For  $c$  satisfying  $\alpha = 1 - G_q(c)$

$$\mathbb{P}(W > c | H_0) \longrightarrow \alpha$$

so the test “Reject  $H_0$  if  $W > c$ ” has asymptotic size  $\alpha$ .

In the special case that  $\widehat{\mathbf{A}} = \mathbf{A}(\mathbf{X}, \mathbf{Z})$  and  $v_i | \mathbf{x}_i, \mathbf{z}_i \sim N(0, \sigma^2)$  then there is a finite sample version of the previous result. Let  $W^0$  be the Wald statistic constructed with a homoskedastic variance matrix estimator, and let

$$F = W/q \quad (12.55)$$

be the the  $F$  statistic, where  $q = \dim(\boldsymbol{\beta}_2)$ .

**Theorem 12.10** Take model (12.50) with  $\widehat{\mathbf{A}} = \mathbf{A}(\mathbf{X}, \mathbf{Z})$ ,  $v_i | \mathbf{x}_i, \mathbf{z}_i \sim N(0, \sigma^2)$  and  $\widehat{\mathbf{w}}_i = (\mathbf{w}_{1i}, \widehat{\mathbf{w}}_{2i})$ . Under  $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ , t-statistics have exact  $N(0, 1)$  distributions, and the  $F$  statistic (12.55) has an exact  $F_{q, n-k}$  distribution, where  $q = \dim(\boldsymbol{\beta}_2)$  and  $k = \dim(\boldsymbol{\beta})$ .

To summarize, in the model  $y_i = \mathbf{w}'_1 \boldsymbol{\beta}_1 + \mathbf{w}'_2 \boldsymbol{\beta}_2 + v_i$  where  $\mathbf{w}_2$  is not observed but replaced with an estimate  $\widehat{\mathbf{w}}_2$ , conventional significance tests for  $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$  are asymptotically valid without adjustment.

While this theory allows tests of  $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ , it unfortunately does just justify conventional standard errors or confidence intervals. For this, we need to work out the distribution without imposing the simplification  $\boldsymbol{\beta}_2 = \mathbf{0}$ . This often needs to be worked out case-by-case, or by using methods based on the generalized method of moments to be introduced in Chapter 13. However, in some important set of examples it is straightforward to work out the asymptotic distribution.

For the remainder of this section we examine the setting where the estimators  $\widehat{\mathbf{A}}$  take a least-squares form, so for some  $\mathbf{X}$  can be written as  $\widehat{\mathbf{A}} = (\mathbf{Z}' \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{X})$ . Such estimators correspond to the multivariate projection model

$$\begin{aligned} \mathbf{x}_i &= \mathbf{A}' \mathbf{z}_i + \mathbf{u}_i \\ \mathbb{E}(\mathbf{z}_i \mathbf{u}'_i) &= \mathbf{0}. \end{aligned} \quad (12.56)$$

This class of estimators directly includes 2SLS and the expectation model described above. We can write the matrix of generated regressors as  $\widehat{\mathbf{W}} = \mathbf{Z} \widehat{\mathbf{A}}$  and then (12.52) as

$$\begin{aligned} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= (\widehat{\mathbf{W}}' \widehat{\mathbf{W}})^{-1} (\widehat{\mathbf{W}}' ((\mathbf{W} - \widehat{\mathbf{W}}) \boldsymbol{\beta} + \mathbf{v})) \\ &= (\widehat{\mathbf{A}}' \mathbf{Z}' \mathbf{Z} \widehat{\mathbf{A}})^{-1} (\widehat{\mathbf{A}}' \mathbf{Z}' (-\mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{U}) \boldsymbol{\beta} + \mathbf{v})) \\ &= (\widehat{\mathbf{A}}' \mathbf{Z}' \mathbf{Z} \widehat{\mathbf{A}})^{-1} (\widehat{\mathbf{A}}' \mathbf{Z}' (-\mathbf{U} \boldsymbol{\beta} + \mathbf{v})) \\ &= (\widehat{\mathbf{A}}' \mathbf{Z}' \mathbf{Z} \widehat{\mathbf{A}})^{-1} (\widehat{\mathbf{A}}' \mathbf{Z}' \mathbf{e}) \end{aligned}$$

where

$$e_i = v_i - \mathbf{u}'_i \boldsymbol{\beta} = y_i - \mathbf{x}'_i \boldsymbol{\beta}. \quad (12.57)$$

This estimator has the asymptotic distribution

$$\sqrt{n} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}})$$

where

$$\mathbf{V}_{\boldsymbol{\beta}} = (\mathbf{A}' \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i) \mathbf{A})^{-1} (\mathbf{A}' \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i e_i^2) \mathbf{A}) (\mathbf{A}' \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i) \mathbf{A})^{-1}. \quad (12.58)$$

Under conditional homoskedasticity the covariance matrix simplifies to

$$\mathbf{V}_{\boldsymbol{\beta}} = (\mathbf{A}' \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i) \mathbf{A})^{-1} \mathbb{E}(e_i^2).$$

An appropriate estimator of  $V_\beta$  is

$$\begin{aligned}\widehat{V}_\beta &= \left( \frac{1}{n} \widehat{\mathbf{W}}' \widehat{\mathbf{W}} \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \widehat{\mathbf{w}}_i \widehat{\mathbf{w}}_i' \widehat{e}_i^2 \right) \left( \frac{1}{n} \widehat{\mathbf{W}}' \widehat{\mathbf{W}} \right)^{-1} \\ \widehat{e}_i &= y_i - \mathbf{x}'_i \widehat{\beta}.\end{aligned}\quad (12.59)$$

Under the assumption of conditional homoskedasticity this can be simplified as usual.

This appears to be the usual covariance matrix estimator, but it is not, because the least-squares residuals  $\widehat{v}_i = y_i - \widehat{\mathbf{w}}'_i \widehat{\beta}$  have been replaced with  $\widehat{e}_i = y_i - \mathbf{x}'_i \widehat{\beta}$ . This is exactly the substitution made by the 2SLS covariance matrix formula. Indeed, the covariance matrix estimator  $\widehat{V}_\beta$  precisely equals the estimator (12.42).

**Theorem 12.11** Take model (12.50) and (12.56) with  $\mathbb{E}(y_i^4) < \infty$ ,  $\mathbb{E}\|\mathbf{z}_i\|^4 < \infty$ ,  $\mathbf{A}'\mathbb{E}(\mathbf{z}_i \mathbf{z}'_i) \mathbf{A} > 0$ , and  $\widehat{\mathbf{A}} = (\mathbf{Z}' \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{X})$ . As  $n \rightarrow \infty$ ,

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} N(\mathbf{0}, V_\beta)$$

where  $V_\beta$  is given in (12.58) with  $e_i$  defined in (12.57). For  $\widehat{V}_\beta$  given in (12.59),

$$\widehat{V}_\beta \xrightarrow{P} V_\beta.$$

Since the parameter estimates are asymptotically normal and the covariance matrix is consistently estimated, standard errors and test statistics constructed from  $\widehat{V}_\beta$  are asymptotically valid with conventional interpretations.

We now summarize the results of this section. In general, care needs to be exercised when estimating models with generated regressors. As a general rule, generated regressors and two-step estimation affects sampling distributions and variance matrices. An important simplification occurs for tests that the generated regressors have zero slopes. In this case conventional tests have conventional distributions, both asymptotically and in finite samples. Another important special case occurs when the generated regressors are least-squares fitted values. In this case the asymptotic distribution takes a conventional form, but the conventional residual needs to be replaced by one constructed with the forecasted variable. With this one modification asymptotic inference using the generated regressors is conventional.

## 12.27 Regression with Expectation Errors

In this section we examine a generated regressor model which includes expectation errors in the regression. This is an important class of generated regressor models, and is relatively straightforward to characterize.

The model is

$$\begin{aligned}y_i &= \mathbf{w}'_i \beta + \mathbf{u}'_i \alpha + v_i \\ \mathbf{w}_i &= \mathbf{A}' \mathbf{z}_i \\ \mathbf{x}_i &= \mathbf{w}_i + \mathbf{u}_i \\ \mathbb{E}(\mathbf{z}_i \varepsilon_i) &= \mathbf{0} \\ \mathbb{E}(\mathbf{u}_i \varepsilon_i) &= \mathbf{0} \\ \mathbb{E}(\mathbf{z}_i \mathbf{u}'_i) &= \mathbf{0}.\end{aligned}$$

The observables are  $(y_i, \mathbf{x}_i, \mathbf{z}_i)$ . This model states that  $\mathbf{w}_i$  is the expectation of  $\mathbf{x}_i$  (or more generally, the projection of  $\mathbf{x}_i$  on  $\mathbf{z}_i$ ) and  $\mathbf{u}_i$  is its expectation error. The model allows for exogenous regressors as in the standard IV model if they are listed in  $\mathbf{w}_i$ ,  $\mathbf{x}_i$  and  $\mathbf{z}_i$ . This model is used, for example, to decompose the effect of expectations from expectation errors. In some cases it is desired to include only the expectation error  $\mathbf{u}_i$ , not the expectation  $\mathbf{w}_i$ . This does not change the results described here.

The model is estimated as follows. First,  $\mathbf{A}$  is estimated by multivariate least-squares of  $\mathbf{x}_i$  on  $\mathbf{z}_i$ ,  $\hat{\mathbf{A}} = (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X})$ , which yields as by-products the fitted values  $\hat{\mathbf{W}} = \mathbf{Z}\hat{\mathbf{A}}$  and residuals  $\hat{\mathbf{U}} = \hat{\mathbf{X}} - \hat{\mathbf{W}}$ . Second, the coefficients are estimated by least-squares of  $y_i$  on the fitted values  $\hat{\mathbf{w}}_i$  and residuals  $\hat{\mathbf{u}}_i$

$$y_i = \hat{\mathbf{w}}_i'\hat{\beta} + \hat{\mathbf{u}}_i'\hat{\alpha} + \hat{v}_i.$$

We now examine the asymptotic distributions of these estimates.

By the first-step regression  $\mathbf{Z}'\hat{\mathbf{U}} = \mathbf{0}$ ,  $\hat{\mathbf{W}}'\hat{\mathbf{U}} = \mathbf{0}$  and  $\mathbf{W}'\hat{\mathbf{U}} = \mathbf{0}$ . This means that  $\hat{\beta}$  and  $\hat{\alpha}$  can be computed separately. Notice that

$$\hat{\beta} = (\hat{\mathbf{W}}'\hat{\mathbf{W}})^{-1}\hat{\mathbf{W}}'y$$

and

$$y = \hat{\mathbf{W}}\beta + \mathbf{U}\alpha + (\mathbf{W} - \hat{\mathbf{W}})\beta + \nu.$$

Substituting, using  $\hat{\mathbf{W}}'\hat{\mathbf{U}} = \mathbf{0}$  and  $\mathbf{W} - \hat{\mathbf{W}} = -\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{U}$  we find

$$\begin{aligned}\hat{\beta} - \beta &= (\hat{\mathbf{W}}'\hat{\mathbf{W}})^{-1}\hat{\mathbf{W}}'(\mathbf{U}\alpha + (\mathbf{W} - \hat{\mathbf{W}})\beta + \nu) \\ &= (\hat{\mathbf{A}}'\mathbf{Z}'\mathbf{Z}\hat{\mathbf{A}})^{-1}\hat{\mathbf{A}}'\mathbf{Z}'(\mathbf{U}\alpha - \mathbf{U}\beta + \nu) \\ &= (\hat{\mathbf{A}}'\mathbf{Z}'\mathbf{Z}\hat{\mathbf{A}})^{-1}\hat{\mathbf{A}}'\mathbf{Z}'e\end{aligned}$$

where

$$e_i = v_i + \mathbf{u}_i'(\alpha - \beta) = y_i - \mathbf{x}_i'\beta.$$

We also find

$$\hat{\alpha} = (\hat{\mathbf{U}}'\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}'y.$$

Since  $\hat{\mathbf{U}}'\mathbf{W} = \mathbf{0}$ ,  $\mathbf{U} - \hat{\mathbf{U}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{U}$  and  $\hat{\mathbf{U}}'\mathbf{Z} = \mathbf{0}$  then

$$\begin{aligned}\hat{\alpha} - \alpha &= (\hat{\mathbf{U}}'\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}'(\mathbf{W}\beta + (\mathbf{U} - \hat{\mathbf{U}})\alpha + \nu) \\ &= (\hat{\mathbf{U}}'\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}'\nu.\end{aligned}$$

Together, we establish the following distributional result.

**Theorem 12.12** For the model and estimates described in this section, with  $\mathbb{E}(y_i^4) < \infty$ ,  $\mathbb{E}\|\mathbf{z}_i\|^4 < \infty$ ,  $\mathbb{E}\|\mathbf{x}_i\|^4 < \infty$ ,  $\mathbf{A}'\mathbb{E}(\mathbf{z}_i\mathbf{z}_i')\mathbf{A} > 0$ , and  $\mathbb{E}(\mathbf{u}_i\mathbf{u}_i') > 0$ , as  $n \rightarrow \infty$

$$\sqrt{n}\begin{pmatrix} \hat{\beta} - \beta \\ \hat{\alpha} - \alpha \end{pmatrix} \xrightarrow{d} N(\mathbf{0}, V) \quad (12.60)$$

where

$$V = \begin{pmatrix} V_{\beta\beta} & V_{\beta\alpha} \\ V_{\alpha\beta} & V_{\alpha\alpha} \end{pmatrix}$$

and

$$\begin{aligned}V_{\beta\beta} &= (\mathbf{A}'\mathbb{E}(\mathbf{z}_i\mathbf{z}_i')\mathbf{A})^{-1}(\mathbf{A}'\mathbb{E}(\mathbf{z}_i\mathbf{z}_i'e_i^2)\mathbf{A})(\mathbf{A}'\mathbb{E}(\mathbf{z}_i\mathbf{z}_i')\mathbf{A})^{-1} \\ V_{\alpha\beta} &= (\mathbb{E}(\mathbf{u}_i\mathbf{u}_i'))^{-1}(\mathbb{E}(\mathbf{u}_i\mathbf{z}_i'e_i v_i)\mathbf{A})(\mathbf{A}'\mathbb{E}(\mathbf{z}_i\mathbf{z}_i')\mathbf{A})^{-1} \\ V_{\alpha\alpha} &= (\mathbb{E}(\mathbf{u}_i\mathbf{u}_i'))^{-1}\mathbb{E}(\mathbf{u}_i\mathbf{u}_i'v_i^2)(\mathbb{E}(\mathbf{u}_i\mathbf{u}_i'))^{-1}.\end{aligned}$$

The asymptotic covariance matrix is estimated by

$$\begin{aligned}\widehat{\mathbf{V}}_{\beta\beta} &= \left( \frac{1}{n} \widehat{\mathbf{W}}' \widehat{\mathbf{W}} \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \widehat{\mathbf{w}}_i \widehat{\mathbf{w}}_i' \widehat{e}_i^2 \right) \left( \frac{1}{n} \widehat{\mathbf{W}}' \widehat{\mathbf{W}} \right)^{-1} \\ \widehat{\mathbf{V}}_{\alpha\beta} &= \left( \frac{1}{n} \widehat{\mathbf{U}}' \widehat{\mathbf{U}} \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \widehat{\mathbf{u}}_i \widehat{\mathbf{w}}_i' \widehat{e}_i \widehat{v}_i \right) \left( \frac{1}{n} \widehat{\mathbf{W}}' \widehat{\mathbf{W}} \right)^{-1} \\ \widehat{\mathbf{V}}_{\alpha\alpha} &= \left( \frac{1}{n} \widehat{\mathbf{U}}' \widehat{\mathbf{U}} \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}_i' \widehat{v}_i^2 \right) \left( \frac{1}{n} \widehat{\mathbf{U}}' \widehat{\mathbf{U}} \right)^{-1}\end{aligned}$$

where

$$\begin{aligned}\widehat{\mathbf{w}}_i &= \widehat{\mathbf{A}}' \mathbf{z}_i \\ \widehat{\mathbf{u}}_i &= \widehat{\mathbf{x}}_i - \widehat{\mathbf{w}}_i \\ \widehat{e}_i &= y_i - \mathbf{x}'_i \widehat{\beta} \\ \widehat{v}_i &= y_i - \widehat{\mathbf{w}}'_i \widehat{\beta} - \widehat{\mathbf{u}}'_i \widehat{\alpha}.\end{aligned}$$

Under conditional homoskedasticity, specifically

$$\mathbb{E} \left( \begin{pmatrix} e_i^2 & e_i v_i \\ e_i v_i & v_i^2 \end{pmatrix} | \mathbf{z}_i \right) = \mathbf{C}$$

then  $\mathbf{V}_{\alpha\beta} = 0$  and the coefficient estimates  $\widehat{\beta}$  and  $\widehat{\alpha}$  are asymptotically independent. The variance components also simplify to

$$\begin{aligned}\mathbf{V}_{\beta\beta} &= (\mathbf{A}' \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i) \mathbf{A})^{-1} \mathbb{E}(e_i^2) \\ \mathbf{V}_{\alpha\alpha} &= (\mathbb{E}(\mathbf{u}_i \mathbf{u}'_i))^{-1} \mathbb{E}(v_i^2).\end{aligned}$$

In this case we have the covariance matrix estimators

$$\begin{aligned}\widehat{\mathbf{V}}_{\beta\beta}^0 &= \left( \frac{1}{n} \widehat{\mathbf{W}}' \widehat{\mathbf{W}} \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \widehat{e}_i^2 \right) \\ \widehat{\mathbf{V}}_{\alpha\alpha}^0 &= \left( \frac{1}{n} \widehat{\mathbf{U}}' \widehat{\mathbf{U}} \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \widehat{v}_i^2 \right)\end{aligned}$$

and  $\widehat{\mathbf{V}}_{\alpha\beta}^0 = \mathbf{0}$ .

## 12.28 Control Function Regression

In this section we present an alternative way of computing the 2SLS estimator by least squares. It is useful in more complicated nonlinear contexts, and also in the linear model to construct tests for endogeneity.

The structural and reduced form equations for the standard IV model are

$$\begin{aligned}y_i &= \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + e_i \\ \mathbf{x}_{2i} &= \boldsymbol{\Gamma}'_{12} \mathbf{z}_{1i} + \boldsymbol{\Gamma}'_{22} \mathbf{z}_{2i} + \mathbf{u}_{2i}.\end{aligned}$$

Since the instrumental variable assumption specifies that  $\mathbb{E}(\mathbf{z}_i e_i) = \mathbf{0}$ ,  $\mathbf{x}_{2i}$  is endogenous (correlated with  $e_i$ ) if and only if  $\mathbf{u}_{2i}$  and  $e_i$  are correlated. We can therefore consider the linear projection of  $e_i$  on  $\mathbf{u}_{2i}$

$$\begin{aligned}e_i &= \mathbf{u}'_{2i} \boldsymbol{\alpha} + \varepsilon_i \\ \boldsymbol{\alpha} &= (\mathbb{E}(\mathbf{u}_{2i} \mathbf{u}'_{2i}))^{-1} \mathbb{E}(\mathbf{u}_{2i} e_i) \\ \mathbb{E}(\mathbf{u}_{2i} \varepsilon_i) &= \mathbf{0}.\end{aligned}$$

Substituting this into the structural form equation we find

$$\begin{aligned} y_i &= \mathbf{x}'_1 \boldsymbol{\beta}_1 + \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + \mathbf{u}'_{2i} \boldsymbol{\alpha} + \varepsilon_i \\ \mathbb{E}(\mathbf{x}_1 \varepsilon_i) &= \mathbf{0} \\ \mathbb{E}(\mathbf{x}_{2i} \varepsilon_i) &= \mathbf{0} \\ \mathbb{E}(\mathbf{u}_{2i} \varepsilon_i) &= \mathbf{0}. \end{aligned} \tag{12.61}$$

Notice that  $\mathbf{x}_{2i}$  is uncorrelated with  $\varepsilon_i$ . This is because  $\mathbf{x}_{2i}$  is correlated with  $e_i$  only through  $\mathbf{u}_{2i}$ , and  $\varepsilon_i$  is the error after  $e_i$  has been projected orthogonal to  $\mathbf{u}_{2i}$ .

If  $\mathbf{u}_{2i}$  were observed we could then estimate (12.61) by least-squares. While it is not observed, we can estimate  $\mathbf{u}_{2i}$  by the reduced-form residual

$$\hat{\mathbf{u}}_{2i} = \mathbf{x}_{2i} - \hat{\Gamma}'_{12} \mathbf{z}_{1i} - \hat{\Gamma}'_{22} \mathbf{z}_{2i}$$

as defined in (12.20). Then the coefficients  $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\alpha})$  can be estimated by least-squares of  $y_i$  on  $(\mathbf{x}_1, \mathbf{x}_{2i}, \hat{\mathbf{u}}_{2i})$ . We can write this as

$$y_i = \mathbf{x}'_i \hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}'_{2i} \hat{\boldsymbol{\alpha}} + \hat{\varepsilon}_i \tag{12.62}$$

or in matrix notation as

$$\mathbf{y} = \mathbf{X} \hat{\boldsymbol{\beta}} + \hat{\mathbf{U}}_2 \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\varepsilon}}.$$

This turns out to be an alternative algebraic expression for the 2SLS estimator.

Indeed, we now show that  $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{2\text{sls}}$ . First, note that the reduced form residual can be written as

$$\hat{\mathbf{U}}_2 = (\mathbf{I}_n - \mathbf{P}_Z) \mathbf{X}_2$$

where  $\mathbf{P}_Z$  is defined in (12.32). By the FWL representation

$$\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}' \tilde{\mathbf{X}})^{-1} (\tilde{\mathbf{X}}' \mathbf{y}) \tag{12.63}$$

where  $\tilde{\mathbf{X}} = [\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2]$ , with

$$\tilde{\mathbf{X}}_1 = \mathbf{X}_1 - \hat{\mathbf{U}}_2 (\hat{\mathbf{U}}'_2 \hat{\mathbf{U}}_2)^{-1} \hat{\mathbf{U}}'_2 \mathbf{X}_1 = \mathbf{X}_1$$

(since  $\hat{\mathbf{U}}'_2 \mathbf{X}_1 = 0$ ) and

$$\begin{aligned} \tilde{\mathbf{X}}_2 &= \mathbf{X}_2 - \hat{\mathbf{U}}_2 (\hat{\mathbf{U}}'_2 \hat{\mathbf{U}}_2)^{-1} \hat{\mathbf{U}}'_2 \mathbf{X}_2 \\ &= \mathbf{X}_2 - \hat{\mathbf{U}}_2 (\mathbf{X}'_2 (\mathbf{I}_n - \mathbf{P}_Z) \mathbf{X}_2)^{-1} \mathbf{X}'_2 (\mathbf{I}_n - \mathbf{P}_Z) \mathbf{X}_2 \\ &= \mathbf{X}_2 - \hat{\mathbf{U}}_2 \\ &= \mathbf{P}_Z \mathbf{X}_2. \end{aligned}$$

Thus  $\tilde{\mathbf{X}} = [\mathbf{X}_1, \mathbf{P}_Z \mathbf{X}_2] = \mathbf{P}_Z \mathbf{X}$ . Substituted into (12.63) we find

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{P}_Z \mathbf{X})^{-1} (\mathbf{X}' \mathbf{P}_Z \mathbf{y}) = \hat{\boldsymbol{\beta}}_{2\text{sls}}$$

which is (12.33) as claimed.

Again, what we have found is that OLS estimation of equation (12.62) yields algebraically the 2SLS estimator  $\hat{\boldsymbol{\beta}}_{2\text{sls}}$ .

We now consider the distribution of the control function estimates. It is a generated regression model, and in fact is covered by the model examined in Section 12.27 after a slight reparametrization. Let  $\mathbf{w}_i = \boldsymbol{\Gamma}' \mathbf{z}_i$  and  $\mathbf{u}_i = \mathbf{x}_i - \boldsymbol{\Gamma}' \mathbf{z}_i = (\mathbf{0}', \mathbf{u}'_{2i})'$ . Then the main equation (12.61) can be written as

$$y_i = \mathbf{w}'_i \boldsymbol{\beta} + \mathbf{u}'_{2i} \boldsymbol{\gamma} + \varepsilon_i$$

where  $\boldsymbol{\gamma} = \boldsymbol{\alpha} + \boldsymbol{\beta}_2$ . This is the model in Section 12.27.

Set  $\hat{\gamma} = \hat{\alpha} + \hat{\beta}_2$ . It follows from (12.60) that as  $n \rightarrow \infty$  we have the joint distribution

$$\sqrt{n} \begin{pmatrix} \hat{\beta}_2 - \beta_2 \\ \hat{\gamma} - \gamma \end{pmatrix} \xrightarrow{d} N(\mathbf{0}, V)$$

where

$$V = \begin{pmatrix} V_{22} & V_{2\gamma} \\ V_{\gamma 2} & V_{\gamma\gamma} \end{pmatrix}$$

$$\begin{aligned} V_{22} &= \left[ (\Gamma' \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i) \Gamma)^{-1} (\Gamma' \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i e_i^2 \Gamma)) (\Gamma' \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i) \Gamma)^{-1} \right]_{22} \\ V_{\gamma 2} &= \left[ (\mathbb{E}(\mathbf{u}_{2i} \mathbf{u}'_{2i}))^{-1} (\mathbb{E}(\mathbf{u}_{2i} \mathbf{z}'_i e_i \varepsilon_i) \Gamma) (\Gamma' \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i) \Gamma)^{-1} \right]_{.2} \\ V_{\gamma\gamma} &= (\mathbb{E}(\mathbf{u}_{2i} \mathbf{u}'_{2i}))^{-1} \mathbb{E}(\mathbf{u}_{2i} \mathbf{u}'_{2i} e_i^2) (\mathbb{E}(\mathbf{u}_{2i} \mathbf{u}'_{2i}))^{-1} \\ e_i &= y_i - \mathbf{x}'_i \beta. \end{aligned}$$

The asymptotic distribution of  $\hat{\gamma} = \hat{\alpha} - \hat{\beta}_2$  can then be deduced.

**Theorem 12.13** If  $\mathbb{E}(y_i^4) < \infty$ ,  $\mathbb{E}\|\mathbf{z}_i\|^4 < \infty$ ,  $\mathbb{E}\|\mathbf{x}_i\|^4 < \infty$ ,  $\mathbf{A}' \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i) \mathbf{A} > 0$ , and  $\mathbb{E}(\mathbf{u}_i \mathbf{u}'_i) > 0$ , as  $n \rightarrow \infty$

$$\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{d} N(\mathbf{0}, V_\alpha)$$

where

$$V_\alpha = V_{22} + V_{\gamma\gamma} - V_{\gamma 2} - V'_{\gamma 2}.$$

Under conditional homoskedasticity we have the important simplifications

$$\begin{aligned} V_{22} &= \left[ (\Gamma' \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i) \Gamma)^{-1} \right]_{22} \mathbb{E}(e_i^2) \\ V_{\gamma\gamma} &= (\mathbb{E}(\mathbf{u}_{2i} \mathbf{u}'_{2i}))^{-1} \mathbb{E}(\varepsilon_i^2) \\ V_{\gamma 2} &= \mathbf{0} \\ V_\alpha &= V_{22} + V_{\gamma\gamma}. \end{aligned}$$

An estimator for  $V_\alpha$  in the general case is

$$\hat{V}_\alpha = \hat{V}_{22} + \hat{V}_{\gamma\gamma} - \hat{V}_{\gamma 2} - \hat{V}'_{\gamma 2} \quad (12.64)$$

where

$$\begin{aligned} \hat{V}_{22} &= \left[ \frac{1}{n} (\mathbf{X}' \mathbf{P}_Z \mathbf{X})^{-1} \mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \left( \sum_{i=1}^n \mathbf{z}_i \mathbf{z}'_i \hat{e}_i^2 \right) (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X} (\mathbf{X}' \mathbf{P}_Z \mathbf{X})^{-1} \right]_{22} \\ \hat{V}_{\gamma 2} &= \left[ \frac{1}{n} (\hat{\mathbf{U}}' \hat{\mathbf{U}})^{-1} \left( \sum_{i=1}^n \hat{\mathbf{u}}_i \hat{\mathbf{w}}'_i \hat{e}_i \hat{\varepsilon}_i \right) (\mathbf{X}' \mathbf{P}_Z \mathbf{X})^{-1} \right]_{.2} \\ \hat{e}_i &= y_i - \mathbf{x}'_i \hat{\beta} \\ \hat{\varepsilon}_i &= y_i - \mathbf{x}'_i \hat{\beta} - \hat{\mathbf{u}}'_{2i} \hat{\alpha}. \end{aligned}$$

Under the assumption of conditional homoskedasticity we have the estimator

$$\begin{aligned} \hat{V}_\alpha^0 &= \hat{V}_{\beta\beta}^0 + \hat{V}_{\gamma\gamma}^0 \\ \hat{V}_{\beta\beta} &= \left[ (\mathbf{X}' \mathbf{P}_Z \mathbf{X})^{-1} \right]_{22} \left( \sum_{i=1}^n \hat{e}_i^2 \right) \\ \hat{V}_{\gamma\gamma} &= (\hat{\mathbf{U}}' \hat{\mathbf{U}})^{-1} \left( \sum_{i=1}^n \hat{\varepsilon}_i^2 \right). \end{aligned}$$

## 12.29 Endogeneity Tests

The 2SLS estimator allows the regressor  $\mathbf{x}_{2i}$  to be endogenous, meaning that  $\mathbf{x}_{2i}$  is correlated with the structural error  $e_i$ . If this correlation is zero, then  $\mathbf{x}_{2i}$  is exogenous and the structural equation can be estimated by least-squares. This is a testable restriction. Effectively, the null hypothesis is

$$\mathbb{H}_0 : \mathbb{E}(\mathbf{x}_{2i} e_i) = \mathbf{0}$$

with the alternative

$$\mathbb{H}_1 : \mathbb{E}(\mathbf{x}_{2i} e_i) \neq \mathbf{0}.$$

The maintained hypothesis is  $\mathbb{E}(\mathbf{z}_i e_i) = \mathbf{0}$ . Since  $\mathbf{x}_{1i}$  is a component of  $\mathbf{z}_i$ , this implies  $\mathbb{E}(\mathbf{x}_{1i} e_i) = \mathbf{0}$ . Consequently we could alternatively write the null as  $\mathbb{H}_0 : \mathbb{E}(\mathbf{x}_i e_i) = \mathbf{0}$  (and some authors do so).

Recall the control function regression (12.61)

$$\begin{aligned} y_i &= \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + \mathbf{u}'_{2i} \boldsymbol{\alpha} + \varepsilon_i \\ \boldsymbol{\alpha} &= (\mathbb{E}(\mathbf{u}_{2i} \mathbf{u}'_{2i}))^{-1} \mathbb{E}(\mathbf{u}_{2i} e_i). \end{aligned}$$

Notice that  $\mathbb{E}(\mathbf{x}_{2i} e_i) = \mathbf{0}$  if and only if  $\mathbb{E}(\mathbf{u}_{2i} e_i) = \mathbf{0}$ , so the hypothesis can be restated as  $\mathbb{H}_0 : \boldsymbol{\alpha} = \mathbf{0}$  against  $\mathbb{H}_1 : \boldsymbol{\alpha} \neq \mathbf{0}$ . Thus a natural test is based on the Wald statistic  $W$  for  $\boldsymbol{\alpha} = \mathbf{0}$  in the control function regression (12.28). Under Theorem 12.9 and Theorem 12.10, under  $\mathbb{H}_0$ ,  $W$  is asymptotically chi-square with  $k_2$  degrees of freedom. In addition, under the normal regression assumptions the  $F$  statistic has an exact  $F(k_2, n - k_1 - 2k_2)$  distribution. We accept the null hypothesis that  $\mathbf{x}_{2i}$  is exogenous if  $W$  (or  $F$ ) is smaller than the critical value, and reject in favor of the hypothesis that  $\mathbf{x}_{2i}$  is endogenous if the statistic is larger than the critical value.

Specifically, estimate the reduced form by least squares

$$\mathbf{x}_{2i} = \widehat{\boldsymbol{\Gamma}}'_{12} \mathbf{z}_{1i} + \widehat{\boldsymbol{\Gamma}}'_{22} \mathbf{z}_{2i} + \widehat{\mathbf{u}}_{2i}$$

to obtain the residuals. Then estimate the control function by least squares

$$y_i = \mathbf{x}'_i \widehat{\boldsymbol{\beta}} + \widehat{\mathbf{u}}'_i \widehat{\boldsymbol{\alpha}} + \widehat{\varepsilon}_i. \quad (12.65)$$

Let  $W$ ,  $W^0$  and  $F = W^0/k_2$  denote the Wald statistic, homoskedastic Wald statistic, and  $F$  statistic for  $\boldsymbol{\alpha} = \mathbf{0}$ .

**Theorem 12.14** Under  $\mathbb{H}_0$ ,  $W \xrightarrow{d} \chi^2_{k_2}$ . Let  $c_{1-\alpha}$  solve  $\mathbb{P}\left(\chi^2_{k_2} \leq c_{1-\alpha}\right) = 1 - \alpha$ . The test “Reject  $\mathbb{H}_0$  if  $W > c_{1-\alpha}$ ” has asymptotic size  $\alpha$ .

**Theorem 12.15** Suppose  $e_i | \mathbf{x}_i, \mathbf{z}_i \sim N(0, \sigma^2)$ . Under  $\mathbb{H}_0$ ,  $F \sim F(k_2, n - k_1 - 2k_2)$ . Let  $c_{1-\alpha}$  solve  $\mathbb{P}(F(k_2, n - k_1 - 2k_2) \leq c_{1-\alpha}) = 1 - \alpha$ . The test “Reject  $\mathbb{H}_0$  if  $F > c_{1-\alpha}$ ” has exact size  $\alpha$ .

Since in general we do not want to impose homoskedasticity, these results suggest that the most appropriate test is the Wald statistic constructed with the robust heteroskedastic covariance matrix. This can be computed in Stata using the command `estat endogenous after ivregress` when the latter uses a robust covariance option. Stata reports the Wald statistic in  $F$  form (and thus uses the  $F$  distribution to

calculate the p-value) as “Robust regression F”. Using the  $F$  rather than the  $\chi^2$  distribution is not formally justified but is a reasonable finite sample adjustment. If the command `estat endogenous` is applied after `ivregress` without a robust covariance option, Stata reports the  $F$  statistic as “Wu-Hausman F”.

There is an alternative (and traditional) way to derive a test for endogeneity. Under  $H_0$ , both OLS and 2SLS are consistent estimators. But under  $H_1$ , they converge to different values. Thus the difference between the OLS and 2SLS estimators is a valid test statistic for endogeneity. It also measures what we often care most about – the impact of endogeneity on the parameter estimates. This literature was developed under the assumption of conditional homoskedasticity (and it is important for these results) so we assume this condition for the development of the statistics.

Let  $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$  be the OLS estimator and let  $\tilde{\beta} = (\tilde{\beta}_1, \tilde{\beta}_2)$  be the 2SLS estimator. Under  $H_0$  (and homoskedasticity) the OLS estimator is Gauss-Markov efficient, so by the Hausman equality

$$\begin{aligned}\text{var}(\hat{\beta}_2 - \tilde{\beta}_2) &= \text{var}(\tilde{\beta}_2) - \text{var}(\hat{\beta}_2) \\ &= \left( (\mathbf{X}'_2 (\mathbf{P}_Z - \mathbf{P}_1) \mathbf{X}_2)^{-1} - (\mathbf{X}'_2 \mathbf{M}_1 \mathbf{X}_2)^{-1} \right) \sigma^2\end{aligned}$$

where  $\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ ,  $\mathbf{P}_1 = \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1$ , and  $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{P}_1$ . Thus a valid test statistic for  $H_0$  is

$$T = \frac{(\hat{\beta}_2 - \tilde{\beta}_2)' \left( (\mathbf{X}'_2 (\mathbf{P}_Z - \mathbf{P}_1) \mathbf{X}_2)^{-1} - (\mathbf{X}'_2 \mathbf{M}_1 \mathbf{X}_2)^{-1} \right)^{-1} (\hat{\beta}_2 - \tilde{\beta}_2)}{\hat{\sigma}^2} \quad (12.66)$$

for some estimate  $\hat{\sigma}^2$  of  $\sigma^2$ . Durbin (1954) first proposed  $T$  as a test for endogeneity in the context of IV estimation, setting  $\hat{\sigma}^2$  to be the least-squares estimate of  $\sigma^2$ . Wu (1973) proposed  $T$  as a test for endogeneity in the context of 2SLS estimation, considering a set of possible estimates  $\hat{\sigma}^2$ , including the regression estimate from (12.65). Hausman (1978) proposed a version of  $T$  based on the full contrast  $\hat{\beta} - \tilde{\beta}$ , and observed that it equals the regression Wald statistic  $W^0$  described earlier. In fact, when  $\hat{\sigma}^2$  is the regression estimate from (12.65), the statistic (12.66) algebraically equals both  $W^0$  and the version of (12.66) based on the full contrast  $\hat{\beta} - \tilde{\beta}$ . We show these equalities below. Thus these three approaches yield exactly the same statistic except for possible differences regarding the choice of  $\hat{\sigma}^2$ . Since the regression  $F$  test described earlier has an exact  $F$  distribution in the normal sampling model, and thus can exactly control test size, this is the preferred version of the test. The general class of tests are called **Durbin-Wu-Hausman** tests, **Wu-Hausman** tests, or **Hausman** tests, depending on the author.

When  $k_2 = 1$  (there is one right-hand-side endogenous variable) which is quite common in applications, the endogeneity test can be equivalently expressed at the t-statistic for  $\hat{\alpha}$  in the estimated control function. Thus it is sufficient to estimate the control function regression and check the t-statistic for  $\hat{\alpha}$ . If  $|\hat{\alpha}| > 2$  then we can reject the hypothesis that  $x_{2i}$  is exogenous for  $\beta$ .

We illustrate using the Card proximity example using the two instruments *public* and *private*. We first estimate the reduced form for *education*, obtain the residual, and then estimate the control function regression. The residual has a coefficient  $-0.088$  with a standard error of  $0.037$  and a t-statistic of  $2.4$ . Since the latter exceeds the  $5\%$  critical value (its p-value is  $0.017$ ) we reject exogeneity. This means that the 2SLS estimates are statistically different from the least-squares estimates of the structural equation and supports our decision to treat education as an endogenous variable. (Alternatively, the  $F$  statistic is  $2.4^2 = 5.7$  with the same p-value).

We now show the equality of the various statistics.

We first show that the statistic (12.66) is not altered if based on the full contrast  $\hat{\beta} - \tilde{\beta}$ . Indeed,  $\hat{\beta}_1 - \tilde{\beta}_1$  is a linear function of  $\hat{\beta}_2 - \tilde{\beta}_2$ , so there is no extra information in the full contrast. To see this, observe that given  $\hat{\beta}_2$ , we can solve by least-squares to find

$$\hat{\beta}_1 = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} (\mathbf{X}'_1 (\mathbf{y} - \mathbf{X}_2 \hat{\beta}_2))$$

and similarly

$$\begin{aligned}\tilde{\beta}_1 &= (\mathbf{X}'_1 \mathbf{X}_1)^{-1} (\mathbf{X}'_1 (\mathbf{y} - \mathbf{P}_Z \mathbf{X}_2 \tilde{\beta})) \\ &= (\mathbf{X}'_1 \mathbf{X}_1)^{-1} (\mathbf{X}'_1 (\mathbf{y} - \mathbf{X}_2 \tilde{\beta}))\end{aligned}$$

the second equality since  $\mathbf{P}_Z \mathbf{X}_1 = \mathbf{X}_1$ . Thus

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_1 - \widetilde{\boldsymbol{\beta}}_1 &= (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 (\mathbf{y} - \mathbf{X}_2 \widehat{\boldsymbol{\beta}}_2) - (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 (\mathbf{y} - \mathbf{P}_Z \mathbf{X}_2 \widetilde{\boldsymbol{\beta}}) \\ &= (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{X}_2 (\widetilde{\boldsymbol{\beta}}_2 - \widehat{\boldsymbol{\beta}}_2)\end{aligned}$$

as claimed.

We next show that  $T$  in (12.66) equals the homoskedastic Wald statistic  $W^0$  for  $\widehat{\boldsymbol{\alpha}}$  from the regression (12.65). Consider the latter regression. Since  $\mathbf{X}_2$  is contained in  $\mathbf{X}$ , the coefficient estimate  $\widehat{\boldsymbol{\alpha}}$  is invariant to replacing  $\widehat{\mathbf{U}}_2 = \mathbf{X}_2 - \widehat{\mathbf{X}}_2$  with  $-\widehat{\mathbf{X}}_2 = -\mathbf{P}_Z \mathbf{X}_2$ . By the FWL representation, setting  $\mathbf{M}_X = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

$$\begin{aligned}\widehat{\boldsymbol{\alpha}} &= -\left(\widehat{\mathbf{X}}'_2 \mathbf{M}_X \widehat{\mathbf{X}}_2\right)^{-1} \widehat{\mathbf{X}}'_2 \mathbf{M}_X \mathbf{y} \\ &= -\left(\mathbf{X}'_2 \mathbf{P}_Z \mathbf{M}_X \mathbf{P}_Z \mathbf{X}_2\right)^{-1} \mathbf{X}'_2 \mathbf{P}_Z \mathbf{M}_X \mathbf{y}.\end{aligned}$$

It follows that

$$W^0 = \frac{\mathbf{y}' \mathbf{M}_X \mathbf{P}_Z \mathbf{X}_2 (\mathbf{X}'_2 \mathbf{P}_Z \mathbf{M}_X \mathbf{P}_Z \mathbf{X}_2)^{-1} \mathbf{X}'_2 \mathbf{P}_Z \mathbf{M}_X \mathbf{y}}{\widehat{\sigma}^2}.$$

Our goal is to show that  $T = W^0$ . Define  $\tilde{\mathbf{X}}_2 = (\mathbf{I}_n - \mathbf{P}_1) \mathbf{X}_2$  so  $\widehat{\boldsymbol{\beta}}_2 = (\tilde{\mathbf{X}}'_2 \tilde{\mathbf{X}}_2)^{-1} \tilde{\mathbf{X}}'_2 \mathbf{y}$ . Then defining using  $(\mathbf{P}_Z - \mathbf{P}_1)(\mathbf{I}_n - \mathbf{P}_1) = (\mathbf{P}_Z - \mathbf{P}_1)$  and defining  $\mathbf{Q} = \tilde{\mathbf{X}}_2 (\tilde{\mathbf{X}}'_2 \tilde{\mathbf{X}}_2)^{-1} \tilde{\mathbf{X}}'_2$

$$\begin{aligned}\Delta &\stackrel{def}{=} (\mathbf{X}'_2 (\mathbf{P}_Z - \mathbf{P}_1) \mathbf{X}_2) (\widetilde{\boldsymbol{\beta}}_2 - \widehat{\boldsymbol{\beta}}_2) \\ &= \mathbf{X}'_2 (\mathbf{P}_Z - \mathbf{P}_1) \mathbf{y} - (\mathbf{X}'_2 (\mathbf{P}_Z - \mathbf{P}_1) \mathbf{X}_2) (\tilde{\mathbf{X}}'_2 \tilde{\mathbf{X}}_2)^{-1} \tilde{\mathbf{X}}'_2 \mathbf{y} \\ &= \mathbf{X}'_2 (\mathbf{P}_Z - \mathbf{P}_1) (\mathbf{I}_n - \mathbf{Q}) \mathbf{y} \\ &= \mathbf{X}'_2 (\mathbf{P}_Z - \mathbf{P}_1 - \mathbf{P}_Z \mathbf{Q}) \mathbf{y} \\ &= \mathbf{X}'_2 \mathbf{P}_Z (\mathbf{I}_n - \mathbf{P}_1 - \mathbf{Q}) \mathbf{y} \\ &= \mathbf{X}'_2 \mathbf{P}_Z \mathbf{M}_X \mathbf{y}.\end{aligned}$$

The third-to-last equality is  $\mathbf{P}_1 \mathbf{Q} = \mathbf{0}$  and the final uses  $\mathbf{M}_X = \mathbf{I}_n - \mathbf{P}_1 - \mathbf{Q}$ . We also calculate that

$$\begin{aligned}\mathbf{Q}^* &\stackrel{def}{=} (\mathbf{X}'_2 (\mathbf{P}_Z - \mathbf{P}_1) \mathbf{X}_2) \left( (\mathbf{X}'_2 (\mathbf{P}_Z - \mathbf{P}_1) \mathbf{X}_2)^{-1} - (\mathbf{X}'_2 \mathbf{M}_1 \mathbf{X}_2)^{-1} \right) \\ &\quad \cdot (\mathbf{X}'_2 (\mathbf{P}_Z - \mathbf{P}_1) \mathbf{X}_2) \\ &= \mathbf{X}'_2 (\mathbf{P}_Z - \mathbf{P}_1 - (\mathbf{P}_Z - \mathbf{P}_1) \mathbf{Q} (\mathbf{P}_Z - \mathbf{P}_1)) \mathbf{X}_2 \\ &= \mathbf{X}'_2 (\mathbf{P}_Z - \mathbf{P}_1 - \mathbf{P}_Z \mathbf{Q} \mathbf{P}_Z) \mathbf{X}_2 \\ &= \mathbf{X}'_2 \mathbf{P}_Z \mathbf{M}_X \mathbf{P}_Z \mathbf{X}_2.\end{aligned}$$

Thus

$$\begin{aligned}T &= \frac{\Delta' \mathbf{Q}^{*-1} \Delta}{\widehat{\sigma}^2} \\ &= \frac{\mathbf{y}' \mathbf{M}_X \mathbf{P}_Z \mathbf{X}_2 (\mathbf{X}'_2 \mathbf{P}_Z \mathbf{M}_X \mathbf{P}_Z \mathbf{X}_2)^{-1} \mathbf{X}'_2 \mathbf{P}_Z \mathbf{M}_X \mathbf{y}}{\widehat{\sigma}^2} \\ &= W^0\end{aligned}$$

as claimed.

## 12.30 Subset Endogeneity Tests

In some cases we may only wish to test the endogeneity of a subset of the variables. In the Card proximity example, we may wish test the exogeneity of *education* separately from *experience* and its square. To execute a subset endogeneity test it is useful to partition the regressors into three groups, so that the structural model is

$$\begin{aligned} y_i &= \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + \mathbf{x}'_{3i} \boldsymbol{\beta}_3 + e_i \\ \mathbb{E}(\mathbf{z}_i e_i) &= \mathbf{0}. \end{aligned}$$

As before, the instrument vector  $\mathbf{z}_i$  includes  $\mathbf{x}_{1i}$ . The variables  $\mathbf{x}_{3i}$  is treated as endogenous, and  $\mathbf{x}_{2i}$  is treated as potentially endogenous. The hypothesis to test is that  $\mathbf{x}_{2i}$  is exogenous, or

$$\mathbb{H}_0 : \mathbb{E}(\mathbf{x}_{2i} e_i) = \mathbf{0}$$

against

$$\mathbb{H}_1 : \mathbb{E}(\mathbf{x}_{2i} e_i) \neq \mathbf{0}.$$

Under homoskedasticity, a straightforward test can be constructed by the Durbin-Wu-Hausman principle. Under  $\mathbb{H}_0$ , the appropriate estimator is 2SLS using the instruments  $(\mathbf{z}_i, \mathbf{x}_{2i})$ . Let this estimator of  $\boldsymbol{\beta}_2$  be denoted  $\hat{\boldsymbol{\beta}}_2$ . Under  $\mathbb{H}_1$ , the appropriate estimator is 2SLS using the smaller instrument set  $\mathbf{z}_i$ . Let this estimator of  $\boldsymbol{\beta}_2$  be denoted  $\tilde{\boldsymbol{\beta}}_2$ . A Durbin-Wu-Hausman-type test of  $\mathbb{H}_0$  against  $\mathbb{H}_1$  is

$$T = (\hat{\boldsymbol{\beta}}_2 - \tilde{\boldsymbol{\beta}}_2)' (\widehat{\text{var}}(\tilde{\boldsymbol{\beta}}_2) - \widehat{\text{var}}(\hat{\boldsymbol{\beta}}_2))^{-1} (\hat{\boldsymbol{\beta}}_2 - \tilde{\boldsymbol{\beta}}_2).$$

The asymptotic distribution under  $\mathbb{H}_0$  is  $\chi^2_{k_2}$  where  $k_2 = \dim(\mathbf{x}_{2i})$ , so we reject the hypothesis that the variables  $\mathbf{x}_{2i}$  are exogenous if  $T$  exceeds an upper critical value from the  $\chi^2_{k_2}$  distribution.

Instead of using the Wald statistic, one could use the  $F$  version of the test by dividing by  $k_2$  and using the  $F$  distribution for critical values. There is no finite sample justification for this modification, however, since  $\mathbf{x}_{3i}$  is endogenous under the null hypothesis.

In Stata, the command `estat endogenous` (adding the variable name to specify which variable to test for exogeneity) after `ivregress` without a robust covariance option reports the  $F$  version of this statistic as “Wu-Hausman F”. For example, in the Card proximity example using the four instruments *public*, *private*, *age* and *age*<sup>2</sup>, if we estimate the equation by 2SLS with a non-robust covariance matrix, and then compute the endogeneity test for education, we find  $F = 272$  with a p-value of 0.0000, but if we compute the test for experience and its square we find  $F = 2.98$  with a p-value of 0.051. In this equation, education is clearly endogenous but the experience variables are unclear.

A heteroskedasticity or cluster-robust test cannot be constructed easily by the Durbin-Wu-Hausman approach, since the covariance matrix does not take a simple form. Instead, we can use the regression approach if we account for the generated regressor problem. The ideal control function regression takes the form

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \mathbf{u}'_{2i} \boldsymbol{\alpha}_2 + \mathbf{u}'_{3i} \boldsymbol{\alpha}_3 + \varepsilon_i$$

where  $\mathbf{u}_{2i}$  and  $\mathbf{u}_{3i}$  are the reduced-form errors from the projections of  $\mathbf{x}_{2i}$  and  $\mathbf{x}_{3i}$  on the instruments  $\mathbf{z}_i$ . The coefficients  $\boldsymbol{\alpha}_2$  and  $\boldsymbol{\alpha}_3$  solve the equations

$$\begin{pmatrix} \mathbb{E}(\mathbf{u}_{2i} \mathbf{u}'_{2i}) & \mathbb{E}(\mathbf{u}_{2i} \mathbf{u}'_{3i}) \\ \mathbb{E}(\mathbf{u}_{3i} \mathbf{u}'_{2i}) & \mathbb{E}(\mathbf{u}_{3i} \mathbf{u}'_{3i}) \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_2 \\ \boldsymbol{\alpha}_3 \end{pmatrix} = \begin{pmatrix} \mathbb{E}(\mathbf{u}_{2i} e_i) \\ \mathbb{E}(\mathbf{u}_{3i} e_i) \end{pmatrix}.$$

The null hypothesis  $\mathbb{E}(\mathbf{x}_{2i} e_i) = \mathbf{0}$  is equivalent to  $\mathbb{E}(\mathbf{u}_{2i} e_i) = \mathbf{0}$ . This implies

$$\boldsymbol{\Psi}' \begin{pmatrix} \boldsymbol{\alpha}_2 \\ \boldsymbol{\alpha}_3 \end{pmatrix} = \mathbf{0} \tag{12.67}$$

where

$$\boldsymbol{\Psi} = \begin{pmatrix} \mathbb{E}(\mathbf{u}_{2i} \mathbf{u}'_{2i}) \\ \mathbb{E}(\mathbf{u}_{3i} \mathbf{u}'_{2i}) \end{pmatrix}.$$

This suggests that an appropriate regression-based test of  $H_0$  versus  $H_1$  is to construct a Wald statistic for the restriction (12.67) in the control function regression

$$y_i = \mathbf{x}'_i \hat{\beta} + \hat{\mathbf{u}}'_{2i} \hat{\alpha}_2 + \hat{\mathbf{u}}'_{3i} \hat{\alpha}_3 + \hat{\varepsilon}_i \quad (12.68)$$

where  $\hat{\mathbf{u}}_{2i}$  and  $\hat{\mathbf{u}}_{3i}$  are the least-squares residuals from the regressions of  $\mathbf{x}_{2i}$  and  $\mathbf{x}_{3i}$  on the instruments  $\mathbf{z}_i$ , respectively, and  $\Psi$  is estimated by

$$\hat{\Psi} = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n \hat{\mathbf{u}}_{2i} \hat{\mathbf{u}}'_{2i} \\ \frac{1}{n} \sum_{i=1}^n \hat{\mathbf{u}}_{3i} \hat{\mathbf{u}}'_{2i} \end{pmatrix}.$$

A complication is that the regression (12.68) has generated regressors which have non-zero coefficients under  $H_0$ . The solution is to use the control-function-robust covariance matrix estimator (12.64) for  $(\hat{\alpha}_2, \hat{\alpha}_3)$ . This yields a valid Wald statistic for  $H_0$  versus  $H_1$ . The asymptotic distribution of the statistic under  $H_0$  is  $\chi^2_{k_2}$  where  $k_2 = \dim(\mathbf{x}_{2i})$ , so the null hypothesis that  $\mathbf{x}_{2i}$  is exogenous is rejected if the Wald statistic exceeds the upper critical value from the  $\chi^2_{k_2}$  distribution.

Heteroskedasticity-robust and cluster-robust subset endogeneity tests are not currently implemented in Stata.

### 12.31 OverIdentification Tests

When  $\ell > k$  the model is **overidentified** meaning that there are more moments than free parameters. This is a restriction and is testable. Such tests are called **overidentification tests**.

The instrumental variables model specifies that

$$\mathbb{E}(\mathbf{z}_i e_i) = \mathbf{0}.$$

Equivalently, since  $e_i = y_i - \mathbf{x}'_i \beta$ , this is the same as

$$\mathbb{E}(\mathbf{z}_i y_i) - \mathbb{E}(\mathbf{z}_i \mathbf{x}'_i) \beta = \mathbf{0}.$$

This is an  $\ell \times 1$  vector of restrictions on the moment matrices  $\mathbb{E}(\mathbf{z}_i y_i)$  and  $\mathbb{E}(\mathbf{z}_i \mathbf{x}'_i)$ . Yet since  $\beta$  is of dimension  $k$  which is less than  $\ell$ , it is not certain if indeed such a  $\beta$  exists.

To make things a bit more concrete, suppose there is a single endogenous regressor  $x_{2i}$ , no  $x_{1i}$ , and two instruments  $z_{1i}$  and  $z_{2i}$ . Then the model specifies that

$$\mathbb{E}(z_{1i} y_i) = \mathbb{E}(z_{1i} x_{2i}) \beta$$

and

$$\mathbb{E}(z_{2i} y_i) = \mathbb{E}(z_{2i} x_{2i}) \beta.$$

Thus  $\beta$  solves both equations. This is rather special.

Another way of thinking about this is that in this context we could solve for  $\beta$  using either one equation or the other. In terms of estimation, this is equivalent to estimating by IV using just the instrument  $z_1$  or instead just using the instrument  $z_2$ . These two estimators (in finite samples) will be different. But if the overidentification hypothesis is correct, both are estimating the same parameter, and both are consistent for  $\beta$  (if the instruments are relevant). In contrast, if the overidentification hypothesis is false, then the two estimators will converge to different probability limits and it is unclear if either probability limit is interesting.

For example, take the 2SLS estimates in the fourth column of Table 12.1, which use *public* and *private* as instruments for *education*. Suppose we instead estimate by IV, using just *public* as an instrument, and then repeat using *private*. The IV coefficient for *education* in the first case is 0.16, and in the second case 0.27. These appear to be quite different. However, the second estimate has quite a large standard error

(0.16) so perhaps the difference is sampling variation. An overidentification test addresses this question formally.

For a general overidentification test, the null and alternative hypotheses are

$$\begin{aligned}\mathbb{H}_0 : \mathbb{E}(\mathbf{z}_i e_i) &= \mathbf{0} \\ \mathbb{H}_1 : \mathbb{E}(\mathbf{z}_i e_i) &\neq \mathbf{0}.\end{aligned}$$

We will also add the conditional homoskedasticity assumption

$$\mathbb{E}(e_i^2 | \mathbf{z}_i) = \sigma^2. \quad (12.69)$$

To avoid imposing (12.69), it is best to take a GMM approach, which we defer until Chapter 13.

To implement a test of  $\mathbb{H}_0$ , consider a linear regression of the error  $e_i$  on the instruments  $\mathbf{z}_i$

$$e_i = \mathbf{z}'_i \boldsymbol{\alpha} + \varepsilon_i \quad (12.70)$$

with

$$\boldsymbol{\alpha} = (\mathbb{E}(\mathbf{z}_i \mathbf{z}'_i))^{-1} \mathbb{E}(\mathbf{z}_i e_i).$$

We can rewrite  $\mathbb{H}_0$  as  $\boldsymbol{\alpha} = \mathbf{0}$ . While  $e_i$  is not observed we can replace it with the 2SLS residual  $\hat{e}_i$ , and estimate  $\boldsymbol{\alpha}$  by least-squares regression

$$\hat{\boldsymbol{\alpha}} = (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \hat{\mathbf{e}}.$$

Sargan (1958) proposed testing  $\mathbb{H}_0$  via a score test, which takes the form

$$S = \hat{\boldsymbol{\alpha}}' (\widehat{\text{var}}(\hat{\boldsymbol{\alpha}}))^{-1} \hat{\boldsymbol{\alpha}} = \frac{\hat{\mathbf{e}}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \hat{\mathbf{e}}}{\hat{\sigma}^2}. \quad (12.71)$$

where  $\hat{\sigma}^2 = \frac{1}{n} \hat{\mathbf{e}}' \hat{\mathbf{e}}$ . Basmann (1960) independently proposed a Wald statistic for  $\mathbb{H}_0$ , which is  $S$  with  $\hat{\sigma}^2$  replaced with  $\tilde{\sigma}^2 = n^{-1} \hat{\mathbf{e}}' \hat{\mathbf{e}}$  where  $\hat{\mathbf{e}} = \hat{\mathbf{e}} - \mathbf{Z} \hat{\boldsymbol{\alpha}}$ . By the equivalence of homoskedastic score and Wald tests (see Section 9.16), Basmann's statistic is a monotonic function of Sargan's statistic and hence they yield equivalent tests. Sargan's version is more typically reported.

The Sargan test rejects  $\mathbb{H}_0$  in favor of  $\mathbb{H}_1$  if  $S > c$  for some critical value  $c$ . An asymptotic test sets  $c$  as the  $1 - \alpha$  quantile of the  $\chi_{\ell-k}^2$  distribution. This is justified by the asymptotic null distribution of  $S$  which we now derive.

**Theorem 12.16** Under Assumption 12.2 and  $\mathbb{E}(e_i^2 | \mathbf{z}_i) = \sigma^2$ , then as  $n \rightarrow \infty$

$$S \xrightarrow{d} \chi_{\ell-k}^2.$$

For  $c$  satisfying  $\alpha = 1 - G_{\ell-k}(c)$ ,

$$\mathbb{P}(S > c | \mathbb{H}_0) \longrightarrow \alpha$$

so the test “Reject  $\mathbb{H}_0$  if  $S > c$ ” has asymptotic size  $\alpha$ .

We prove Theorem 12.16 below.

The Sargan statistic  $S$  is an asymptotic test of the overidentifying restrictions under the assumption of conditional homoskedasticity. It has some limitations. First, it is an asymptotic test, and does not have a finite sample (e.g.  $F$ ) counterpart. Simulation evidence suggests that the test can be oversized (reject too frequently) in small and moderate sample sizes. Consequently, p-values should be interpreted cautiously. Second, the assumption of conditional homoskedasticity is unrealistic in applications. The best

way to generalize the Sargan statistic to allow heteroskedasticity is to use the GMM overidentification statistic – which we will examine in Chapter 13. For 2SLS, Wooldridge (1995) suggested a robust score test, but Baum, Schaffer and Stillman (2003) point out that it is numerically equivalent to the GMM overidentification statistic. Hence the bottom line appears to be that to allow heteroskedasticity or clustering, it is best to use a GMM approach.

In overidentified applications, it is always prudent to report an overidentification test. If the test is insignificant it means that the overidentifying restrictions are not rejected, supporting the estimated model. If the overidentifying test statistic is highly significant (if the p-value is very small) this is evidence that the overidentifying restrictions are violated. In this case we should be concerned that the model is misspecified and interpreting the parameter estimates should be done cautiously.

When reporting the results of an overidentification test, it seems reasonable to focus on very small significance levels, such as 1%. This means that we should only treat a model as “rejected” if the Sargan p-value is very small, e.g. less than 0.01. The reason to focus on very small significance levels is because it is very difficult to interpret the result “The model is rejected”. Stepping back a bit, it does not seem credible that any overidentified model is literally true, rather what seems potentially credible is that an overidentified model is a reasonable approximation. A test is asking the question “Is there evidence that a model is not true” when we really want to know the answer to “Is there evidence that the model is a poor approximation”. Consequently it seems reasonable to require strong evidence to lead to the conclusion “Let’s reject this model”. The recommendation is that mild rejections (p-values between 1% and 5%) should be viewed as mildly worrisome, but not critical evidence against a model. The results of an overidentification test should be integrated with other information before making a strong decision.

We illustrate the methods with the Card college proximity example. We have estimated two overidentified models by 2SLS, in columns 4 & 5 of Table 12.1. In each case, the number of overidentifying restrictions is 1. We report the Sargan statistic and its asymptotic p-value (calculated using the  $\chi^2_1$  distribution) in the table. Both p-values (0.37 and 0.47) are far from significant, indicating that there is no evidence that the models are misspecified.

We now prove Theorem 12.16. The statistic  $S$  is invariant to rotations of  $\mathbf{Z}$  (replacing  $\mathbf{Z}$  with  $\mathbf{Z}\mathbf{C}$ ) so without loss of generality we assume  $\mathbb{E}(\mathbf{z}_i\mathbf{z}'_i) = \mathbf{I}_\ell$ . As  $n \rightarrow \infty$ ,  $n^{-1/2}\mathbf{Z}'\mathbf{e} \xrightarrow{d} \sigma\mathbf{Z}$  where  $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_\ell)$ . Also  $\frac{1}{n}\mathbf{Z}'\mathbf{Z} \xrightarrow{p} \mathbf{I}_\ell$  and  $\frac{1}{n}\mathbf{Z}'\mathbf{X} \xrightarrow{p} \mathbf{Q}$ , say. Then

$$\begin{aligned} n^{-1/2}\mathbf{Z}'\widehat{\mathbf{e}} &= \left( \mathbf{I}_\ell - \left( \frac{1}{n}\mathbf{Z}'\mathbf{X} \right) \left( \frac{1}{n}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X} \right)^{-1} \left( \frac{1}{n}\mathbf{X}'\mathbf{Z} \right) \left( \frac{1}{n}\mathbf{Z}'\mathbf{Z} \right)^{-1} \right) n^{-1/2}\mathbf{Z}'\mathbf{e} \\ &\xrightarrow{d} \sigma \left( \mathbf{I}_\ell - \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}' \right) \mathbf{Z}. \end{aligned}$$

Since  $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$  it follows that

$$S \xrightarrow{d} \mathbf{Z}' \left( \mathbf{I}_\ell - \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}' \right) \mathbf{Z} \sim \chi^2_{\ell-k}.$$

The distribution is  $\chi^2_{\ell-k}$  since  $\mathbf{I}_\ell - \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'$  is idempotent with rank  $\ell - k$ .

The Sargan statistic test can be implemented in Stata using the command `estat overid` after `ivregress 2sls` or `ivregres liml` if a standard (non-robust) covariance matrix has been specified (that is, without the ‘,r’ option), or by the command `estat overid, forcenonrobust` otherwise.

### Denis Sargan

The British econometrician John Denis Sargan (1924-1996) was a pioneer in the field of econometrics. He made a range of fundamental contributions, including the overidentification test, Edgeworth expansions, and unit root theory. He was also influential in British econometrics as the dissertation advisor for many influential econometricians.

## 12.32 Subset OverIdentification Tests

Tests of  $\mathbb{H}_0 : \mathbb{E}(\mathbf{z}_i e_i) = \mathbf{0}$  are typically interpreted as tests of model specification. The alternative  $\mathbb{H}_1 : \mathbb{E}(\mathbf{z}_i e_i) \neq \mathbf{0}$  means that at least one element of  $\mathbf{z}_i$  is correlated with the error  $e_i$  and is thus an invalid instrumental variable. In some cases it may be reasonable to test only a subset of the moment conditions.

As in the previous section we restrict attention to the homoskedasticity case  $\mathbb{E}(e_i^2 | \mathbf{z}_i) = \sigma^2$ .

Partition  $\mathbf{z}_i = (\mathbf{z}_{ai}, \mathbf{z}_{bi})$  with dimensions  $\ell_a$  and  $\ell_b$ , respectively, where  $\mathbf{z}_{ai}$  contains the instruments which are believed to be uncorrelated with  $e_i$ , and  $\mathbf{z}_{bi}$  contains the instruments which may be correlated with  $e_i$ . It is necessary to select this partition so that  $\ell_a > k$ , or equivalently  $\ell_b < \ell - k$ . This means that the model with just the instruments  $\mathbf{z}_{ai}$  is over-identified, or that  $\ell_b$  is smaller than the number of overidentifying restrictions. (If  $\ell_a = k$  then the tests described here exist but reduce to the Sargan test so are not interesting.) Hence the tests require that  $\ell - k > 1$ , that the number of overidentifying restrictions exceeds one.

Given this partition, the maintained hypothesis is that  $\mathbb{E}(\mathbf{z}_{ai} e_i) = \mathbf{0}$ . The null and alternative hypotheses are

$$\begin{aligned}\mathbb{H}_0 &: \mathbb{E}(\mathbf{z}_{bi} e_i) = \mathbf{0} \\ \mathbb{H}_1 &: \mathbb{E}(\mathbf{z}_{bi} e_i) \neq \mathbf{0}.\end{aligned}$$

That is, the null hypothesis is that the full set of moment conditions are valid, while the alternative hypothesis is that the instrument subset  $\mathbf{z}_{bi}$  is correlated with  $e_i$  and thus an invalid instrument. Rejection of  $\mathbb{H}_0$  in favor of  $\mathbb{H}_1$  is then interpreted as evidence that  $\mathbf{z}_{bi}$  is misspecified as an instrument.

Based on the same reasoning as described in the previous section, to test  $\mathbb{H}_0$  against  $\mathbb{H}_1$  we consider a partitioned version of the regression (12.70)

$$e_i = \mathbf{z}'_{ai} \boldsymbol{\alpha}_a + \mathbf{z}'_{bi} \boldsymbol{\alpha}_b + \varepsilon_i$$

but now focus on the coefficient  $\boldsymbol{\alpha}_b$ . Given  $\mathbb{E}(\mathbf{z}_{ai} e_i) = \mathbf{0}$ ,  $\mathbb{H}_0$  is equivalent to  $\boldsymbol{\alpha}_b = \mathbf{0}$ . The equation is estimated by least-squares, replacing the unobserved  $e_i$  with the 2SLS residual  $\hat{e}_i$ . The estimate of  $\boldsymbol{\alpha}_b$  is

$$\hat{\boldsymbol{\alpha}}_b = (\mathbf{Z}'_b \mathbf{M}_a \mathbf{Z}_b)^{-1} \mathbf{Z}'_b \mathbf{M}_a \hat{\mathbf{e}}$$

where  $\mathbf{M}_a = \mathbf{I}_n - \mathbf{Z}_a (\mathbf{Z}'_a \mathbf{Z}_a)^{-1} \mathbf{Z}'_a$ . Newey (1985) showed that an optimal (asymptotically most powerful) test of  $\mathbb{H}_0$  against  $\mathbb{H}_1$  is to reject for large values of the score statistic

$$\begin{aligned}N &= \hat{\boldsymbol{\alpha}}'_b \left( \widehat{\text{var}(\hat{\boldsymbol{\alpha}})} \right)^{-1} \hat{\boldsymbol{\alpha}}_b \\ &= \frac{\hat{\mathbf{e}}' \mathbf{R} \left( \mathbf{R}' \mathbf{R} - \mathbf{R}' \hat{\mathbf{X}} \left( \hat{\mathbf{X}}' \hat{\mathbf{X}} \right)^{-1} \hat{\mathbf{X}}' \mathbf{R} \right)^{-1} \mathbf{R}' \hat{\mathbf{e}}}{\hat{\sigma}^2}\end{aligned}$$

where  $\hat{\mathbf{X}} = \mathbf{P} \mathbf{X}$ ,  $\mathbf{P} = \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}'$ ,  $\mathbf{R} = \mathbf{M}_a \mathbf{Z}_b$ , and  $\hat{\sigma}^2 = \frac{1}{n} \hat{\mathbf{e}}' \hat{\mathbf{e}}$ .

Independently from Newey (1985), Eichenbaum, Hansen, and Singleton (1988) proposed a test based on the difference of Sargan statistics. Letting  $S$  be the Sargan test statistic (12.71) based on the full instrument set and  $S_a$  be the Sargan test based on the instrument set  $\mathbf{z}_{ai}$ , the Sargan difference statistic is

$$C = S - S_a.$$

Specifically, let  $\tilde{\boldsymbol{\beta}}_{2\text{sls}}$  be the 2SLS estimator using the instruments  $\mathbf{z}_{ai}$  only, set  $\tilde{e}_i = y_i - \mathbf{x}'_i \tilde{\boldsymbol{\beta}}_{2\text{sls}}$ , and set  $\tilde{\sigma}^2 = \frac{1}{n} \tilde{\mathbf{e}}' \tilde{\mathbf{e}}$ . Then

$$S_a = \frac{\tilde{\mathbf{e}}' \mathbf{Z}_a (\mathbf{Z}'_a \mathbf{Z}_a)^{-1} \mathbf{Z}'_a \tilde{\mathbf{e}}}{\tilde{\sigma}^2}.$$

An advantage of the  $C$  statistic is that it is quite simple to calculate from the standard regression output.

At this point it is useful to reflect on our stated requirement that  $\ell_a > k$ . Indeed, if  $\ell_a < k$  then  $\mathbf{z}_{ai}$  fails the order condition for identification and  $\tilde{\beta}_{2\text{SLS}}$  cannot be calculated. Thus  $\ell_a \geq k$  is necessary to compute  $S_a$  and hence  $S$ . Furthermore, if  $\ell_a = k$  then  $\mathbf{z}_{ai}$  is just identified so while  $\tilde{\beta}_{2\text{SLS}}$  can be calculated, the statistic  $S_a = 0$  so  $C = S$ . Thus when  $\ell_a = k$  the subset test equals the full overidentification test so there is no gain from considering subset tests.

The  $C$  statistic  $S_a$  is asymptotically equivalent to replacing  $\tilde{\sigma}^2$  in  $S_a$  with  $\hat{\sigma}^2$ , yielding the statistic

$$C^* = \frac{\tilde{\mathbf{e}}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \tilde{\mathbf{e}}}{\hat{\sigma}^2} - \frac{\tilde{\mathbf{e}}' \mathbf{Z}_a (\mathbf{Z}'_a \mathbf{Z}_a)^{-1} \mathbf{Z}'_a \tilde{\mathbf{e}}}{\hat{\sigma}^2}.$$

It turns out that this is Newey's statistic  $N$ . These tests have chi-square asymptotic distributions.

Let  $c$  satisfy  $\alpha = 1 - G_{\ell_b}(c)$ .

**Theorem 12.17** Algebraically,  $N = C^*$ . Under Assumption 12.2 and  $\mathbb{E}(e_i^2 | \mathbf{z}_i) = \sigma^2$ , as  $n \rightarrow \infty$ ,  $N \xrightarrow{d} \chi_{\ell_b}^2$  and  $C \xrightarrow{d} \chi_{\ell_b}^2$ . Thus the tests "Reject  $H_0$  if  $N > c$ " and "Reject  $H_0$  if  $C > c$ " are asymptotically equivalent and have asymptotic size  $\alpha$ .

Theorem 12.17 shows that  $N$  and  $C^*$  are identical, and are near equivalents to the convenient statistic  $C^*$ , and the appropriate asymptotic distribution is  $\chi_{\ell_b}^2$ . Computationally, the easiest method to implement a subset overidentification test is to estimate the model twice by 2SLS, first using the full instrument set  $\mathbf{z}_i$  and the second using the partial instrument set  $\mathbf{z}_{ai}$ . Compute the Sargan statistics for both 2SLS regressions, and compute  $C$  as the difference in the Sargan statistics. In Stata, for example, this is simple to implement with a few lines of code.

We illustrate using the Card college proximity example. Our reported 2SLS estimates have  $\ell - k = 1$  so there is no role for a subset overidentification test. (Recall, the number of overidentifying restrictions must exceed one.) To illustrate we consider adding extra instruments to the estimates in column 5 of Table 12.1 (the 2SLS estimates using *public*, *private*, *age*, and *age*<sup>2</sup> as instruments for *education*, *experience*, and *experience*<sup>2</sup>/100). We add two instruments: the years of education of the *father* and the *mother* of the worker. These variables had been used in the earlier labor economics literature as instruments, but Card did not. (He used them as regression controls in some specifications.) The motivation for using parent's education as instruments is the hypothesis that parental education influences children's educational attainment, but does not directly influence their ability. The more modern labor economics literature has disputed this idea, arguing that children are educated in part at home, and thus parent's education has a direct impact on the skill attainment of children (and not just an indirect impact via educational attainment). The older view was that parent's education is a valid instrument, the modern view is that it is not valid. We can test this dispute using a overidentification subset test.

We do this by estimating the wage equation by 2SLS using *public*, *private*, *age*, *age*<sup>2</sup>, *father*, and *mother*, as instruments for *education*, *experience*, and *experience*<sup>2</sup>/100). We do not report the parameter estimates here, but observe that this model is overidentified with 3 overidentifying restrictions. We calculate the Sargan overidentification statistic. It is 7.9 with an asymptotic p-value (calculated using  $\chi_3^2$ ) of 0.048. This is a mild rejection of the null hypothesis of correct specification. As we argued in the previous section, this by itself is not reason to reject the model. Now we consider a subset overidentification test. We are interested in testing the validity of the two instruments *father* and *mother*, not the instruments *public*, *private*, *age*, *age*<sup>2</sup>. To test the hypothesis that these two instruments are uncorrelated with the structural error, we compute the difference in Sargan statistic,  $C = 7.9 - 0.5 = 7.4$ , which has a p-value (calculated using  $\chi_2^2$ ) of 0.025. This is marginally statistically significant, meaning that there is evidence that *father* and *mother* are not valid instruments for the wage equation. Since the p-value is not smaller than 1%, it is not overwhelming evidence, but it still supports Card's decision to not use parental education as instruments for the wage equation.

We now prove the results in Theorem 12.17.

We first show that  $N = C^*$ . Define  $\mathbf{P}_a = \mathbf{Z}_a(\mathbf{Z}'_a \mathbf{Z}_a)^{-1} \mathbf{Z}'_a$  and  $\mathbf{P}_R = \mathbf{R}(\mathbf{R}' \mathbf{R})^{-1} \mathbf{R}'$ . Since  $[\mathbf{Z}_a, \mathbf{R}]$  span  $\mathbf{Z}$  we find  $\mathbf{P} = \mathbf{P}_R + \mathbf{P}_a$  and  $\mathbf{P}_R \mathbf{P}_a = \mathbf{0}$ . It will be useful to note that

$$\begin{aligned}\mathbf{P}_R \hat{\mathbf{X}} &= \mathbf{P}_R \mathbf{P} \mathbf{X} = \mathbf{P}_R \mathbf{X} \\ \hat{\mathbf{X}}' \hat{\mathbf{X}} - \hat{\mathbf{X}}' \mathbf{P}_R \hat{\mathbf{X}} &= \mathbf{X}' (\mathbf{P} - \mathbf{P}_R) \mathbf{X} = \mathbf{X}' \mathbf{P}_a \mathbf{X}.\end{aligned}$$

The fact that  $\mathbf{X}' \mathbf{P} \hat{\mathbf{e}} = \hat{\mathbf{X}}' \hat{\mathbf{e}} = \mathbf{0}$  implies  $\mathbf{X}' \mathbf{P}_R \hat{\mathbf{e}} = -\mathbf{X}' \mathbf{P}_a \hat{\mathbf{e}}$ . Finally, since  $\mathbf{y} = \mathbf{X} \hat{\boldsymbol{\beta}} + \hat{\mathbf{e}}$ ,

$$\tilde{\mathbf{e}} = \left( \mathbf{I}_n - \mathbf{X} (\mathbf{X}' \mathbf{P}_a \mathbf{X})^{-1} \mathbf{X}' \mathbf{P}_a \right) \hat{\mathbf{e}}$$

so

$$\tilde{\mathbf{e}}' \mathbf{P}_a \tilde{\mathbf{e}} = \hat{\mathbf{e}}' \left( \mathbf{P}_a - \mathbf{P}_a \mathbf{X} (\mathbf{X}' \mathbf{P}_a \mathbf{X})^{-1} \mathbf{X}' \mathbf{P}_a \right) \hat{\mathbf{e}}.$$

Applying the Woodbury matrix equality to the definition of  $N$ , and the above algebraic relationships,

$$\begin{aligned}N &= \frac{\hat{\mathbf{e}}' \mathbf{P}_R \hat{\mathbf{e}} + \hat{\mathbf{e}}' \mathbf{P}_R \hat{\mathbf{X}} \left( \hat{\mathbf{X}}' \hat{\mathbf{X}} - \hat{\mathbf{X}}' \mathbf{P}_R \hat{\mathbf{X}} \right)^{-1} \hat{\mathbf{X}}' \mathbf{P}_R \hat{\mathbf{e}}}{\hat{\sigma}^2} \\ &= \frac{\hat{\mathbf{e}}' \mathbf{P} \hat{\mathbf{e}} - \hat{\mathbf{e}}' \mathbf{P}_a \hat{\mathbf{e}} + \hat{\mathbf{e}}' \mathbf{P}_a \mathbf{X} (\mathbf{X}' \mathbf{P}_a \mathbf{X})^{-1} \mathbf{X}' \mathbf{P}_a \hat{\mathbf{e}}}{\hat{\sigma}^2} \\ &= \frac{\hat{\mathbf{e}}' \mathbf{P} \hat{\mathbf{e}} - \hat{\mathbf{e}}' \mathbf{P}_a \tilde{\mathbf{e}}}{\hat{\sigma}^2} \\ &= C^*\end{aligned}$$

as claimed.

We next establish the asymptotic distribution. Since  $\mathbf{Z}_a$  is a subset of  $\mathbf{Z}$ ,  $\mathbf{P} \mathbf{M}_a = \mathbf{M}_a \mathbf{P}$ , thus  $\mathbf{P} \mathbf{R} = \mathbf{R}$  and  $\mathbf{R}' \mathbf{X} = \mathbf{R}' \hat{\mathbf{X}}$ . Consequently

$$\begin{aligned}\frac{1}{\sqrt{n}} \mathbf{R}' \hat{\mathbf{e}} &= \frac{1}{\sqrt{n}} \mathbf{R}' (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}) \\ &= \frac{1}{\sqrt{n}} \mathbf{R}' \left( \mathbf{I}_n - \mathbf{X} (\hat{\mathbf{X}}' \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}' \right) \mathbf{e} \\ &= \frac{1}{\sqrt{n}} \mathbf{R}' \left( \mathbf{I}_n - \hat{\mathbf{X}} (\hat{\mathbf{X}}' \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}' \right) \mathbf{e} \\ &\xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}_2)\end{aligned}$$

where

$$\mathbf{V}_2 = \operatorname{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{R}' \mathbf{R} - \frac{1}{n} \mathbf{R}' \hat{\mathbf{X}} \left( \frac{1}{n} \hat{\mathbf{X}}' \hat{\mathbf{X}} \right)^{-1} \frac{1}{n} \hat{\mathbf{X}}' \mathbf{R} \right).$$

It follows that  $N = C^* \xrightarrow{d} \chi^2_{\ell_b}$  as claimed. Since  $C = C^* + o_p(1)$  it has the same limiting distribution.

### 12.33 Bootstrap Overidentification Tests

The bootstrap for 2SLS (Section 12.23) can be used for overidentification tests, but the bootstrap version of the overidentification statistic must be adjusted. This is because in the bootstrap universe the overidentified moment conditions are not satisfied. One solution is to center the moment conditions.

For the 2SLS estimator the standard overidentification test is based on the Sargan statistic

$$\begin{aligned}S &= n \frac{\hat{\mathbf{e}}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \hat{\mathbf{e}}}{\hat{\mathbf{e}}' \hat{\mathbf{e}}} \\ \hat{\mathbf{e}} &= \mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}_{2\text{sls}}.\end{aligned}$$

The recentered bootstrap analog is

$$S^{**} = n \frac{(\hat{\mathbf{e}}^{*'} \mathbf{Z}^* - \mathbf{Z}' \hat{\mathbf{e}}) (\mathbf{Z}^{*'} \mathbf{Z}^*)^{-1} (\mathbf{Z}^{*'} \hat{\mathbf{e}}^* - \mathbf{Z}' \hat{\mathbf{e}})}{\hat{\mathbf{e}}^{*'} \hat{\mathbf{e}}^*}$$

$$\hat{\mathbf{e}}^* = \mathbf{y}^* - \mathbf{X}^* \hat{\boldsymbol{\beta}}_{2\text{sls}}^*.$$

On each bootstrap sample  $S^{**}(b)$  is calculated and stored. The bootstrap p-value is

$$p^* = \frac{1}{B} \sum_{b=1}^B \mathbb{1}(S^{**}(b) > S).$$

This bootstrap p-value is asymptotically valid because the statistic  $S^{**}$  satisfies the overidentified moment conditions.

### 12.34 Local Average Treatment Effects

In a pair of influential papers, Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996) proposed a new interpretation of the instrumental variables estimator using the potential outcomes model introduced in Section 2.30.

We will restrict attention to the case that the endogenous regressor  $x$  and excluded instrument  $z$  are binary variables. We write the model as a pair of potential outcome functions. The dependent variable  $y$  is a function of the regressor and an unobservable vector  $\mathbf{u}$

$$y = h(x, \mathbf{u})$$

and the endogenous regressor  $x$  is a function of the instrument  $z$  and  $\mathbf{u}$

$$x = g(z, \mathbf{u}).$$

By specifying  $\mathbf{u}$  as a vector there is no loss of generality in letting both equations depend on  $\mathbf{u}$ .

In this framework, the outcomes are determined by the random vector  $\mathbf{u}$  and the exogenous instrument  $z$ . This determines  $x$ , which determines  $y$ . To put this in the context of the college proximity example, the variable  $\mathbf{u}$  is everything specific about an individual. Given college proximity  $z$ , the person decides to attend college or not. The person's wage is determined by the individual attributes  $\mathbf{u}$  as well as college attendance  $x$ , but is not directly affected by college proximity  $z$ .

We can omit the random variable  $\mathbf{u}$  from the notation as follows. An individual  $i$  has a realization  $\mathbf{u}_i$ . We then set  $y_i(x) = h(x, \mathbf{u}_i)$  and  $x_i(z) = g(z, \mathbf{u}_i)$ . Also, given a realization  $z_i$  the observables are  $x_i = x_i(z_i)$  and  $y_i = y_i(x_i)$ .

In this model the causal effect of college for individual  $i$  is

$$C_i = y_i(1) - y_i(0).$$

As discussed in Section 2.30, in general this is individual-specific.

We would like to learn about the distribution of the causal effects, or at least features of the distribution. A common feature of interest is the average treatment effect (ATE)

$$ATE = \mathbb{E}(C_i) = \mathbb{E}(y_i(1) - y_i(0)).$$

This, however, is typically not feasible to estimate allowing for endogenous  $x$  without strong assumptions (such as that the causal effect  $C_i$  is constant across individuals). The treatment effect literature has explored what features of the distribution of  $C_i$  can be estimated.

One particular feature of interest, and emphasized by Imbens and Angrist (1994), is known as the local average treatment effect (LATE), and is roughly the average effect upon those effected by the instrumental variable. To understand LATE, it is helpful to consider the college proximity example using

the potential outcomes framework. In this framework, each person is fully characterized by their individual unobservable  $\mathbf{u}_i$ . Given  $\mathbf{u}_i$ , their decision to attend college is a function of the proximity indicator  $z_i$ . For some students, proximity has no effect on their decision. For other students, it has an effect in the specific sense that given  $z_i = 1$  they choose to attend college while if  $z_i = 0$  they choose to not attend. We can summarize the possibilities with the following chart, which is based on labels developed by Angrist, Imbens and Rubin (1996).

	$x(0) = 0$	$x(0) = 1$
$x(1) = 0$	Never Takers	Deniers
$x(1) = 1$	Compliers	Always Takers

The columns indicate the college attendance decision given  $z = 0$ . The rows indicate the college attendance decision given  $z = 1$ . The four entries are labels for the four types of individuals based on these decisions. The upper-left entry are the individuals who do not attend college regardless of  $z$ . They are called “Never Takers”. The lower-right entry are the individuals who conversely attend college regardless of  $z$ . They are called “Always Takers”. The bottom left are the individuals who only attend college if they live close to one. They are called “Compliers”. The upper right entry is a bit of a challenge. These are individuals who attend college only if they do not live close to one. They are called “Deniers”. Imbens and Angrist discovered that to identify the parameters of interest we need to assume that there are no Deniers, or equivalently that  $x(1) \geq x(0)$ , which they label as a “monotonicity” condition – that increasing the instrument cannot decrease  $x$  for any individual.

We can distinguish the types in the table by the relative values of  $x(1) - x(0)$ . For Never-Takers and Always-Takers,  $x(1) - x(0) = 0$ , while for Compliers,  $x(1) - x(0) = 1$ .

We are interested in the causal effect  $C_i = h(1, \mathbf{u}) - h(0, \mathbf{u})$  of college attendance on wages. Consider the average causal effect among the different types. Among Never-Takers and Always-Takers,  $x(1) = x(0)$  so

$$\mathbb{E}(y_i(1) - y_i(0)|x_i(1) = x_i(0)).$$

Suppose we try and estimate its average value, conditional for each the three types of individuals: Never-Takers, Always-Takers, and Compliers. It would impossible for the Never-Takers and Always-Takers. For the former, none attend college so it would be impossible to ascertain the effect of college attendance, and similarly for the latter since they all attend college. Thus the only group for which we can estimate a causal effect are the Compliers. This is

$$\text{LATE} = \mathbb{E}(y_i(1) - y_i(0)|x_i(1) > x_i(0)).$$

Imbens and Angrist called this the **local average treatment effect (LATE)** as it is the average treatment effect for the sub-population whose endogenous regressor is affected by changes in the instrumental variable.

Interestingly, we show below that

$$\text{LATE} = \frac{\mathbb{E}(y_i | z_i = 1) - \mathbb{E}(y_i | z_i = 0)}{\mathbb{E}(x_i | z_i = 1) - \mathbb{E}(x_i | z_i = 0)}. \quad (12.72)$$

That is, LATE equals the Wald expression (12.29) for the slope coefficient in the IV regression model. This means that the standard IV estimator is an estimator of LATE. Thus when treatment effects are potentially heterogeneous, we can interpret IV as an estimator of LATE. The equality (12.72) occurs under the following conditions.

**Assumption 12.3**  $\mathbf{u}_i$  and  $z_i$  are independent; and  $\mathbb{P}(x_i(1) - x_i(0) < 0) = 0$ .

One interesting feature about LATE is that its value can depend on the instrument  $z_i$  and the distribution of causal effects  $C_i$  in the population. To make this concrete, suppose that instead of the Card proximity instrument, we consider an instrument based on the financial cost of local college attendance. It is reasonable to expect that while the set of students affected by these two instruments are similar, the two sets of students will not be the same. That is, some students may be responsive to proximity but not finances, and conversely. If the causal effect  $C_i$  has a different average in these two groups of students, then LATE will be different when calculated with these two instruments. Thus LATE can vary by the choice of instrument.

How can that be? How can a well-defined parameter depend on the choice of instrument? Doesn't this contradict the basic IV regression model? The answer is that the basic IV regression model is more restrictive – it specifies that the causal effect  $\beta$  is common across all individuals. Thus its value is the same regardless of the choice of specific instrument (so long as it satisfies the instrumental variables assumptions). In contrast, the potential outcomes framework is more general, allowing for the causal effect to vary across individuals. What this analysis shows us is that in this context is quite possible for the LATE coefficient to vary by instrument. This occurs when causal effects are heterogeneous.

One implication of the LATE framework is that IV estimates should be interpreted as causal effects only for the population of compliers. Interpretation should focus on the population of potential compliers and extension to other populations should be done with caution. For example, in the Card proximity model, the IV estimates of the causal return to schooling presented in Table 12.1 should be interpreted as applying to the population of students who are incentivized to attend college by the presence of a college within their home county. The estimates should not be applied to other students.

Formally, the analysis of this section examined the case of a binary instrument and endogenous regressor. How does this generalize? Suppose that the regressor  $x$  is discrete, taking  $J + 1$  discrete values. We can then rewrite the model as one with  $J$  binary endogenous regressors. If we then have  $J$  binary instruments, we are back in the Imbens-Angrist framework (assuming the instruments have a monotonic impact on the endogenous regressors). A benefit is that with a larger set of instruments it is plausible that the set of compliers in the population is expanded.

We close this section by showing (12.72) under Assumption 12.3. The realized value of  $x_i$  can be written as

$$x_i = (1 - z_i)x_i(0) + z_i x_i(1) = x_i(0) + z_i(x_i(1) - x_i(0)).$$

Similarly

$$y_i = y_i(0) + x_i(y_i(1) - y_i(0)) = y_i(0) + x_i C_i.$$

Combining,

$$y_i = y_i(0) + x_i(0)C_i + z_i(x_i(1) - x_i(0))C_i.$$

The independence of  $u_i$  and  $z_i$  implies independence of  $(y_i(0), y_i(1), x_i(0), x_i(1), C_i)$  and  $z_i$ . Thus

$$\mathbb{E}(y_i|z_i = 1) = \mathbb{E}(y_i(0)) + \mathbb{E}(x_i(0)C_i) + \mathbb{E}((x_i(1) - x_i(0))C_i)$$

and

$$\mathbb{E}(y_i|z_i = 0) = \mathbb{E}(y_i(0)) + \mathbb{E}(x_i(0)C_i).$$

Subtracting we obtain

$$\begin{aligned} \mathbb{E}(y_i|z_i = 1) - \mathbb{E}(y_i|z_i = 0) &= \mathbb{E}((x_i(1) - x_i(0))C_i) \\ &= 1 \cdot \mathbb{E}(C_i|x_i(1) - x_i(0) = 1)\mathbb{P}(x_i(1) - x_i(0) = 1) \\ &\quad + 0 \cdot \mathbb{E}(C_i|x_i(1) - x_i(0) = 0)\mathbb{P}(x_i(1) - x_i(0) = 0) \\ &\quad + (-1) \cdot \mathbb{E}(C_i|x_i(1) - x_i(0) = -1)\mathbb{P}(x_i(1) - x_i(0) = -1) \\ &= \mathbb{E}(C_i|x_i(1) - x_i(0) = 1)(\mathbb{E}(x_i|z_i = 1) - \mathbb{E}(x_i|z_i = 0)) \end{aligned}$$

where the final equality uses  $\mathbb{P}(x_i(1) - x_i(0) < 0) = 0$  and

$$\mathbb{P}(x_i(1) - x_i(0) = 1) = \mathbb{E}(x_i(1) - x_i(0)) = \mathbb{E}(x_i|z_i = 1) - \mathbb{E}(x_i|z_i = 0).$$

Rearranging

$$\text{LATE} = \mathbb{E}(C_i | x_i(1) - x_i(0) = 1) = \frac{\mathbb{E}(y_i | z_i = 1) - \mathbb{E}(y_i | z_i = 0)}{\mathbb{E}(x_i | z_i = 1) - \mathbb{E}(x_i | z_i = 0)}$$

as claimed.

### 12.35 Identification Failure

Recall the reduced form equation

$$\mathbf{x}_{2i} = \boldsymbol{\Gamma}'_{12}\mathbf{z}_{1i} + \boldsymbol{\Gamma}'_{22}\mathbf{z}_{2i} + \mathbf{u}_{2i}.$$

The parameter  $\beta$  fails to be identified if  $\boldsymbol{\Gamma}_{22}$  has deficient rank. The consequences of identification failure for inference are quite severe.

Take the simplest case where  $k_1 = 0$  and  $k_2 = \ell_2 = 1$ . Then the model may be written as

$$\begin{aligned} y_i &= x_i\beta + e_i \\ x_i &= z_i\gamma + u_i \end{aligned} \tag{12.73}$$

and  $\boldsymbol{\Gamma}_{22} = \gamma = \mathbb{E}(z_i x_i) / \mathbb{E}(z_i^2)$ . We see that  $\beta$  is identified if and only if  $\gamma \neq 0$ , which occurs when  $\mathbb{E}(x_i z_i) \neq 0$ . Thus identification hinges on the existence of correlation between the excluded exogenous variable and the included endogenous variable.

Suppose this condition fails. In this case  $\gamma = 0$  and  $\mathbb{E}(x_i z_i) = 0$ . We now analyze the distribution of the least-squares and IV estimators of  $\beta$ . For simplicity we assume conditional homoskedasticity and normalize the variances to unity. Thus

$$\text{var}\left(\begin{pmatrix} e_i \\ u_i \end{pmatrix} | z_i\right) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \tag{12.74}$$

The errors have non-zero correlation  $\rho \neq 0$  which occurs when the variables are endogenous.

By the CLT we have the joint convergence

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \begin{pmatrix} z_i e_i \\ z_i u_i \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

It is convenient to define  $\xi_0 = \xi_1 - \rho \xi_2$  which is normal and independent of  $\xi_2$ .

As a benchmark, it is useful to observe that the least-squares estimator of  $\beta$  satisfies

$$\hat{\beta}_{\text{ols}} - \beta = \frac{n^{-1} \sum_{i=1}^n u_i e_i}{n^{-1} \sum_{i=1}^n u_i^2} \xrightarrow{p} \rho \neq 0$$

so endogeneity causes  $\hat{\beta}_{\text{ols}}$  to be inconsistent for  $\beta$ .

Under identification failure  $\gamma = 0$  the asymptotic distribution of the IV estimator is

$$\hat{\beta}_{\text{iv}} - \beta = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n z_i e_i}{\frac{1}{\sqrt{n}} \sum_{i=1}^n z_i x_i} \xrightarrow{d} \frac{\xi_1}{\xi_2} = \rho + \frac{\xi_0}{\xi_2}.$$

This asymptotic convergence result uses the continuous mapping theorem, which applies since the function  $\xi_1 / \xi_2$  is continuous everywhere except at  $\xi_2 = 0$ , which occurs with probability equal to zero.

This limiting distribution has several notable features.

First,  $\hat{\beta}_{\text{iv}}$  does not converge in probability to a limit, rather it converges in distribution to a random variable. Thus the IV estimator is inconsistent. Indeed, it is not possible to consistently estimate an unidentified parameter and  $\beta$  is not identified when  $\gamma = 0$ .

Second, the ratio  $\xi_0/\xi_2$  is symmetrically distributed about zero, so the median of the limiting distribution of  $\hat{\beta}_{\text{iv}}$  is  $\beta + \rho$ . This means that the IV estimator is median biased under endogeneity. Thus under identification failure the IV estimator does not correct the centering (median bias) of least-squares.

Third, the ratio  $\xi_0/\xi_2$  of two independent normal random variables is Cauchy distributed. This is particularly nasty, as the Cauchy distribution does not have a finite mean. The distribution has thick tails meaning that extreme values occur with higher frequency than the normal, and inferences based on the normal distribution can be quite incorrect.

Together, these results show that  $\gamma = 0$  renders the IV estimator particularly poorly behaved – it is inconsistent, median biased, and non-normally distributed.

We can also examine the behavior of the t-statistic. For simplicity consider the classical (homoskedastic) t-statistic. The error variance estimate has the asymptotic distribution

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - x_i \hat{\beta}_{\text{iv}})^2 \\ &= \frac{1}{n} \sum_{i=1}^n e_i^2 - \frac{2}{n} \sum_{i=1}^n e_i x_i (\hat{\beta}_{\text{iv}} - \beta) + \frac{1}{n} \sum_{i=1}^n x_i^2 (\hat{\beta}_{\text{iv}} - \beta)^2 \\ &\xrightarrow{d} 1 - 2\rho \frac{\xi_1}{\xi_2} + \left( \frac{\xi_1}{\xi_2} \right)^2.\end{aligned}$$

Thus the t-statistic has the asymptotic distribution

$$T = \frac{\hat{\beta}_{\text{iv}} - \beta}{\sqrt{\hat{\sigma}^2 \sum_{i=1}^n z_i^2 / |\sum_{i=1}^n z_i x_i|}} \xrightarrow{d} \frac{\xi_1 / \xi_2}{\sqrt{1 - 2\rho \frac{\xi_1}{\xi_2} + \left( \frac{\xi_1}{\xi_2} \right)^2}}.$$

The limiting distribution is non-normal, meaning that inference using the normal distribution will be (considerably) incorrect. This distribution depends on the correlation  $\rho$ . The distortion is increasing in  $\rho$ . Indeed as  $\rho \rightarrow 1$  we have  $\xi_1/\xi_2 \rightarrow_p 1$  and the unexpected finding  $\hat{\sigma}^2 \rightarrow_p 0$ . The latter means that the conventional standard error  $s(\hat{\beta}_{\text{iv}})$  for  $\hat{\beta}_{\text{iv}}$  also converges in probability to zero. This implies that the t-statistic diverges in the sense  $|T| \rightarrow_p \infty$ . In this situations users may incorrectly interpret estimates as precise, despite the fact that they are useless.

## 12.36 Weak Instruments

In the previous section we examined the extreme consequences of full identification failure. Similar problems occur when identification is weak in the sense that the reduced form coefficients are of small magnitude. In this section we derive an asymptotic distribution of the OLS, 2SLS, and LIML estimators when the reduced form coefficients are treated as weak. We show that the estimators are inconsistent, and the 2SLS and LIML estimators remain random in large samples.

To simplify the exposition we assume that there are no included exogenous variables (no  $x_1$ ) so we write  $x_2, z_2$  and  $\beta_2$  simply as  $x, z$  and  $\beta$ . Thus the model is

$$\begin{aligned}y_i &= x'_i \beta + e_i \\ x_i &= \Gamma' z_i + u_{2i}.\end{aligned}$$

Define the reduced form error vector  $a_i = (v_i, u_{2i})$  and its variance matrix

$$\mathbb{E}(a_i a_i') = \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Recall that the structural error is  $e_i = v_i - \beta' u_{2i} = \gamma' u_i$  where  $\gamma = (1, -\beta)$ , which has variance  $\mathbb{E}(e_i^2 | z_i) = \gamma' \Sigma \gamma$ . Also define the covariance  $\Sigma_{2e} = \mathbb{E}(u_{2i} e_i | z_i) = \Sigma_{21} - \Sigma_{22} \beta$ .

In Section 12.35 we assumed complete identification failure in the sense that  $\Gamma = \mathbf{0}$ . We now want to assume that identification does not completely fail, but is weak in the sense that  $\Gamma$  is small. A rich asymptotic distribution theory has been developed to understand this setting by modeling  $\Gamma$  as “local-to-zero”. The seminal contribution is Staiger and Stock (1997). The theory was extended to nonlinear GMM estimation by Stock and Wright (2000).

The technical device introduced by Staiger and Stock (1997) is to assume that the reduced form parameter is **local-to-zero**, specifically

$$\Gamma = n^{-1/2} \mathbf{C} \quad (12.75)$$

where  $\mathbf{C}$  is a free matrix. The  $n^{-1/2}$  scaling is picked because it provides just the right balance to allow a useful distribution theory. The local-to-zero assumption (12.75) is not meant to be taken literally but rather is meant to be a useful distributional approximation. The parameter  $\mathbf{C}$  indexes the degree of identification. Larger  $\|\mathbf{C}\|$  implies stronger identification; smaller  $\|\mathbf{C}\|$  implies weaker identification.

We now derive the asymptotic distribution of the least-squares, 2SLS and LIML estimators under the local-to-unity assumption (12.75).

The least-squares estimator satisfies

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\beta} &= (n^{-1} \mathbf{X}' \mathbf{X})^{-1} (n^{-1} \mathbf{X}' \mathbf{e}) \\ &= (n^{-1} \mathbf{U}_2' \mathbf{U}_2)^{-1} (n^{-1} \mathbf{U}_2' \mathbf{e}) + o_p(1) \\ &\xrightarrow{p} \Sigma_{22}^{-1} \Sigma_{2e}. \end{aligned}$$

Thus the least-squares estimator is inconsistent for  $\boldsymbol{\beta}$ .

To examine the 2SLS estimator, by the central limit theorem

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{z}_i \mathbf{a}'_i \xrightarrow{d} \boldsymbol{\xi} = [\boldsymbol{\xi}_1, \boldsymbol{\xi}_2]$$

where

$$\text{vec}(\boldsymbol{\xi}) \sim N(0, \mathbb{E}(\mathbf{a}_i \mathbf{a}'_i \otimes \mathbf{z}_i \mathbf{z}'_i)).$$

This implies

$$\frac{1}{\sqrt{n}} \mathbf{Z}' \mathbf{e} \xrightarrow{d} \boldsymbol{\xi}_e = \boldsymbol{\xi} \boldsymbol{\gamma}.$$

We also find that

$$\frac{1}{\sqrt{n}} \mathbf{Z}' \mathbf{X} = \frac{1}{n} \mathbf{Z}' \mathbf{Z} \mathbf{C} + \frac{1}{\sqrt{n}} \mathbf{Z}' \mathbf{U}_2 \xrightarrow{d} \mathbf{Q}_z \mathbf{C} + \boldsymbol{\xi}_2.$$

Thus

$$\begin{aligned} \mathbf{X}' \mathbf{P}_Z \mathbf{X} &= \left( \frac{1}{\sqrt{n}} \mathbf{X}' \mathbf{Z} \right) \left( \frac{1}{n} \mathbf{Z}' \mathbf{Z} \right)^{-1} \left( \frac{1}{\sqrt{n}} \mathbf{Z}' \mathbf{X} \right) \\ &\xrightarrow{d} (\mathbf{Q}_z \mathbf{C} + \boldsymbol{\xi}_2)' \mathbf{Q}_z^{-1} (\mathbf{Q}_z \mathbf{C} + \boldsymbol{\xi}_2) \end{aligned}$$

and

$$\begin{aligned} \mathbf{X}' \mathbf{P}_Z \mathbf{e} &= \left( \frac{1}{\sqrt{n}} \mathbf{X}' \mathbf{Z} \right) \left( \frac{1}{n} \mathbf{Z}' \mathbf{Z} \right)^{-1} \left( \frac{1}{\sqrt{n}} \mathbf{Z}' \mathbf{e} \right) \\ &\xrightarrow{d} (\mathbf{Q}_z \mathbf{C} + \boldsymbol{\xi}_2)' \mathbf{Q}_z^{-1} \boldsymbol{\xi}_e. \end{aligned}$$

We find that the 2SLS estimator has the asymptotic distribution

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{2sls}} - \boldsymbol{\beta} &= (\mathbf{X}' \mathbf{P}_Z \mathbf{X})^{-1} (\mathbf{X}' \mathbf{P}_Z \mathbf{e}) \\ &\xrightarrow{d} \left( (\mathbf{Q}_z \mathbf{C} + \boldsymbol{\xi}_2)' \mathbf{Q}_z^{-1} (\mathbf{Q}_z \mathbf{C} + \boldsymbol{\xi}_2) \right)^{-1} (\mathbf{Q}_z \mathbf{C} + \boldsymbol{\xi}_2)' \mathbf{Q}_z^{-1} \boldsymbol{\xi}_e. \end{aligned} \quad (12.76)$$

As in the case of complete identification failure, we find that  $\hat{\beta}_{2\text{sls}}$  is inconsistent for  $\beta$ , it is asymptotically random, and its asymptotic distribution is non-normal. The distortion is affected by the coefficient  $C$ . As  $\|C\| \rightarrow \infty$  the distribution in (12.76) converges in probability to zero, suggesting that  $\hat{\beta}_{2\text{sls}}$  is consistent for  $\beta$ . This corresponds to the classic “strong identification” context.

Now consider the LIML estimator. The reduced form is  $Y = Z\Pi + a$ . This implies  $M_Z Y = M_Z a$  and by standard asymptotic theory

$$\frac{1}{n} Y' M_Z Y = \frac{1}{n} a' M_Z a \xrightarrow{p} \Sigma = \mathbb{E}(a_i a_i').$$

Define  $\bar{\beta} = [\beta, I_k]$  so that the reduced form coefficients equal  $\Pi = [\Gamma\beta, \Gamma] = n^{-1/2} C \bar{\beta}$ . Then

$$\frac{1}{\sqrt{n}} Z' Y = \frac{1}{n} Z' Z C \bar{\beta} + \frac{1}{\sqrt{n}} Z' U \xrightarrow{d} Q_z C \bar{\beta} + \xi$$

and

$$Y' Z (Z' Z)^{-1} Z' Y \xrightarrow{d} (Q_z C \bar{\beta} + \xi)' Q_z^{-1} (Q_z C \bar{\beta} + \xi).$$

This allows us to calculate that by the continuous mapping theorem

$$\begin{aligned} n\hat{\mu} &= \min_{\gamma} \frac{\gamma' Y' Z (Z' Z)^{-1} Z' Y \gamma}{\gamma' \frac{1}{n} Y' M_Z Y \gamma} \\ &\xrightarrow{d} \min_{\gamma} \frac{\gamma' (Q_z C \bar{\beta} + \xi)' Q_z^{-1} (Q_z C \bar{\beta} + \xi) \gamma}{\gamma' \Sigma \gamma} \\ &= \mu^* \end{aligned}$$

say, which is a function of  $\xi$  and thus random. We deduce that the asymptotic distribution of the LIML estimator is

$$\begin{aligned} \hat{\beta}_{\text{lml}} - \beta &= \left( X' P_Z X - n\hat{\mu} \frac{1}{n} X' M_Z X \right)^{-1} \left( X' P_Z e - n\hat{\mu} \frac{1}{n} X' M_Z e \right) \\ &\xrightarrow{d} \left( (Q_z C + \xi_2)' Q_z^{-1} (Q_z C + \xi_2) - \mu^* \Sigma_{22} \right)^{-1} \left( (Q_z C + \xi_2)' Q_z^{-1} \xi_e - \mu^* \Sigma_{2e} \right). \end{aligned}$$

Similarly to 2SLS, the LIML estimator is inconsistent for  $\beta$ , is asymptotically random, and non-normally distributed.

We summarize.

**Theorem 12.18** Under (12.75),

$$\hat{\beta}_{\text{ols}} - \beta \xrightarrow{p} \Sigma_{22}^{-1} \Sigma_{2e}$$

$$\hat{\beta}_{2\text{sls}} - \beta \xrightarrow{d} \left( (Q_z C + \xi_2)' Q_z^{-1} (Q_z C + \xi_2) \right)^{-1} (Q_z C + \xi_2)' Q_z^{-1} \xi_e$$

and

$$\begin{aligned} \hat{\beta}_{\text{lml}} - \beta &\xrightarrow{d} \left( (Q_z C + \xi_2)' Q_z^{-1} (Q_z C + \xi_2) - \mu^* \Sigma_{22} \right)^{-1} \\ &\quad \cdot \left( (Q_z C + \xi_2)' Q_z^{-1} \xi_e - \mu^* \Sigma_{2e} \right) \end{aligned}$$

where

$$\mu^* = \min_{\gamma} \frac{\gamma' (Q_z C \bar{\beta} + \xi)' Q_z^{-1} (Q_z C \bar{\beta} + \xi) \gamma}{\gamma' \Sigma \gamma}.$$

All three estimators are inconsistent. The 2SLS and LIML estimators are asymptotically random with non-standard distributions, similar to the asymptotic distribution of the IV estimator under complete identification failure explored in the previous section. The difference under weak identification is the presence of the coefficient matrix  $\mathbf{C}$ .

### 12.37 Many Instruments

Some applications have available a large number  $\ell$  of instruments. If they are all valid, using a large number should reduce the asymptotic variance relative to estimation with a smaller number of instruments. Is it then good practice to use many instruments? Or is there a cost to this practice? Bekker (1994) initiated a large literature investigating this question by formalizing the idea of “many instruments”. Bekker proposed an asymptotic approximation which treats the number of instruments  $\ell$  as proportional to the sample size, that is  $\ell = \alpha n$ , or equivalently that  $\ell/N \rightarrow \alpha \in [0, 1)$ . The distributional theory obtained is similar in many respects to the weak instrument theory outlined in the previous section. Consequently the impact of “weak” and “many” instruments is similar.

Again for simplicity we assume that there are no included exogenous regressors so that the model is

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbf{x}_i &= \mathbf{\Gamma}' \mathbf{z}_i + \mathbf{u}_{2i} \end{aligned} \tag{12.77}$$

with  $\mathbf{z}_i \ell \times 1$ . We also make the simplifying assumption that the errors are conditionally homoskedastic. Specifically, for  $\mathbf{a}_i = (v_i, \mathbf{u}_{2i})$

$$\mathbb{E}(\mathbf{a}_i \mathbf{a}'_i | \mathbf{z}_i) = \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}. \tag{12.78}$$

In addition we assume that the conditional fourth moments are bounded

$$\mathbb{E}(\|\mathbf{a}_i\|^4 | \mathbf{z}_i) \leq B < \infty. \tag{12.79}$$

The idea that there are “many instruments” is formalized by the assumption that the number of instruments is increasing proportionately with the sample size

$$\frac{\ell}{n} \longrightarrow \alpha. \tag{12.80}$$

The best way to think about this is to view  $\alpha$  as the ratio of  $\ell$  to  $n$  in a given sample. Thus if an application has  $n = 100$  observations and  $\ell = 10$  instruments, then we should treat  $\alpha = 0.10$ .

Suppose that there is a single endogenous regressor  $x_i$ . Calculate its variance using the reduced form:  $\text{var}(x_i) = \text{var}(\mathbf{z}'_i \mathbf{\Gamma}) + \text{var}(u_i)$ . Suppose as well that  $\text{var}(x_i)$  and  $\text{var}(u_i)$  are unchanging as  $\ell$  increases. This implies that  $\text{var}(\mathbf{z}'_i \mathbf{\Gamma})$  is unchanging, even though the dimension  $\ell$  is increasing. This is a useful assumption, as it implies that the population  $R^2$  of the reduced form is not changing with  $\ell$ . We don't need this exact condition, rather we simply assume that the sample version converges in probability to a fixed constant. Specifically, we assume that

$$\frac{1}{n} \sum_{i=1}^n \mathbf{\Gamma}' \mathbf{z}_i \mathbf{z}'_i \mathbf{\Gamma} \xrightarrow{p} \mathbf{H} \tag{12.81}$$

for some matrix  $\mathbf{H} > 0$ . Again, this essentially implies that the  $R^2$  of the reduced form regressions for each regressor in  $\mathbf{x}_i$  converge to constants.

As a baseline it is useful to examine the behavior of the least-squares estimator of  $\boldsymbol{\beta}$ . First, observe that the variance of  $\text{vec}(n^{-1} \sum_{i=1}^n \mathbf{\Gamma}' \mathbf{z}_i \mathbf{u}'_i)$ , conditional on  $\mathbf{Z}$ , is

$$\boldsymbol{\Sigma} \otimes n^{-2} \sum_{i=1}^n \mathbf{\Gamma}' \mathbf{z}_i \mathbf{z}'_i \mathbf{\Gamma} \xrightarrow{p} \mathbf{0}$$

by (12.81). Thus it converges in probability to zero:

$$n^{-1} \sum_{i=1}^n \boldsymbol{\Gamma}' \mathbf{z}_i \mathbf{a}'_i \xrightarrow{p} \mathbf{0}. \quad (12.82)$$

Combined with (12.81) and the WLLN we find

$$\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i e_i = \frac{1}{n} \sum_{i=1}^n \boldsymbol{\Gamma}' \mathbf{z}_i e_i + \frac{1}{n} \sum_{i=1}^n \mathbf{u}_{2i} e_i \xrightarrow{p} \boldsymbol{\Sigma}_{2e}$$

and

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i &= \frac{1}{n} \sum_{i=1}^n \boldsymbol{\Gamma}' \mathbf{z}_i \mathbf{z}'_i \boldsymbol{\Gamma} + \frac{1}{n} \sum_{i=1}^n \boldsymbol{\Gamma}' \mathbf{z}_i \mathbf{u}'_{2i} + \frac{1}{n} \sum_{i=1}^n \mathbf{u}_{2i} \mathbf{z}'_i \boldsymbol{\Gamma} + \frac{1}{n} \sum_{i=1}^n \mathbf{u}_{2i} \mathbf{u}'_{2i} \\ &\xrightarrow{p} \mathbf{H} + \boldsymbol{\Sigma}_{22}. \end{aligned}$$

Hence

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{ols}} &= \boldsymbol{\beta} + \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i e_i \right) \\ &\xrightarrow{p} \boldsymbol{\beta} + (\mathbf{H} + \boldsymbol{\Sigma}_{22})^{-1} \boldsymbol{\Sigma}_{2e}. \end{aligned}$$

Thus least-squares is inconsistent for  $\boldsymbol{\beta}$ .

Now consider the 2SLS estimator. In matrix notation, setting  $\mathbf{P}_Z = Z(Z'Z)^{-1}Z'$ ,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{2sls}} - \boldsymbol{\beta} &= \left( \frac{1}{n} X' \mathbf{P}_Z X \right)^{-1} \left( \frac{1}{n} X' \mathbf{P}_Z \mathbf{e} \right) \\ &= \left( \frac{1}{n} \boldsymbol{\Gamma}' Z' Z \boldsymbol{\Gamma} + \frac{1}{n} \boldsymbol{\Gamma}' Z' \mathbf{u}_2 + \frac{1}{n} \mathbf{u}'_2 Z \boldsymbol{\Gamma} + \frac{1}{n} \mathbf{u}'_2 \mathbf{P}_Z \mathbf{u}_2 \right)^{-1} \left( \frac{1}{n} \boldsymbol{\Gamma}' Z' \mathbf{e} + \frac{1}{n} \mathbf{u}'_2 \mathbf{P}_Z \mathbf{e} \right). \end{aligned} \quad (12.83)$$

In the expression on the right-side of (12.83), several of the components have been examined in (12.81) and (12.82). We now examine the remaining components  $\frac{1}{n} \mathbf{u}'_2 \mathbf{P}_Z \mathbf{e}$  and  $\frac{1}{n} \mathbf{u}'_2 \mathbf{P}_Z \mathbf{u}_2$  which are sub-components of the matrix  $\frac{1}{n} \mathbf{a}' \mathbf{P}_Z \mathbf{a}$ . Take the  $jk^{th}$  element  $\frac{1}{n} \mathbf{a}'_j \mathbf{P}_Z \mathbf{a}_k$ .

First, take its expectation. We have (given under the conditional homoskedasticity assumption (12.78))

$$\mathbb{E} \left( \frac{1}{n} \mathbf{a}'_j \mathbf{P}_Z \mathbf{a}_k \mid Z \right) = \frac{1}{n} \text{tr} \mathbb{E} \left( \mathbf{P}_Z \mathbf{a}_k \mathbf{a}'_j \mid Z \right) = \frac{1}{n} \text{tr}(\mathbf{P}) \boldsymbol{\Sigma}_{jk} = \frac{\ell}{n} \boldsymbol{\Sigma}_{jk} \quad (12.84)$$

the final equality since  $\text{tr}(\mathbf{P}_Z) = \ell$ .

Second, we calculate its variance, which is a more cumbersome exercise. Let  $P_{im} = \mathbf{z}'_i (Z'Z)^{-1} \mathbf{z}_m$  be the  $im^{th}$  element of  $\mathbf{P}_Z$ . Then  $\mathbf{a}'_j \mathbf{P} \mathbf{a}_k = \sum_{i=1}^n \sum_{m=1}^n a_{ji} a_{km} P_{im}$ . The matrix  $\mathbf{P}_Z$  is idempotent. It therefore has the properties  $\sum_{i=1}^n P_{ii} = \text{tr}(\mathbf{P}_Z) = \ell$  and  $0 \leq P_{ii} \leq 1$ . The property  $\mathbf{P}_Z \mathbf{P}_Z = \mathbf{P}_Z$  also implies

$\sum_{m=1}^n P_{im}^2 = P_{ii}$ . Then

$$\begin{aligned} \text{var}\left(\frac{1}{n} \mathbf{a}'_j \mathbf{P}_Z \mathbf{a}_k \mid \mathbf{Z}\right) &= \frac{1}{n^2} \mathbb{E} \left( \sum_{i=1}^n \sum_{m=1}^n (a_{ji} a_{km} - \mathbb{E}(a_{ji} a_{km}) \mathbf{1}(i=m)) P_{im} \mid \mathbf{Z} \right)^2 \\ &= \frac{1}{n^2} \mathbb{E} \left( \sum_{i=1}^n \sum_{m=1}^n \sum_{q=1}^n \sum_{r=1}^n (a_{ji} a_{km} - \Sigma_{jk} \mathbf{1}(i=m)) P_{im} (a_{jq} a_{kr} - \Sigma_{jk} \mathbf{1}(q=r)) P_{qr} \right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \mathbb{E} \left( (a_{ji} a_{ki} - \Sigma_{jk})^2 \right) P_{ii}^2 \\ &\quad + \frac{1}{n^2} \sum_{i=1}^n \sum_{m \neq i} \mathbb{E} \left( a_{ji}^2 a_{km}^2 \right) P_{im}^2 + \frac{1}{n^2} \sum_{i=1}^n \sum_{m \neq i} \mathbb{E} \left( a_{ji} a_{km} a_{jm} a_{ki} \right) P_{im}^2 \\ &\leq \frac{B}{n^2} \left( \sum_{i=1}^n P_{ii}^2 + 2 \sum_{i=1}^n \sum_{m=1}^n P_{im}^2 \right) \\ &\leq \frac{3B}{n^2} \sum_{i=1}^n P_{ii} \\ &= 3B \frac{\ell}{n^2} \rightarrow 0. \end{aligned}$$

The third equality holds because the remaining cross-products have zero expectation since the observations are independent and the errors have zero mean. The first inequality is (12.79). The second uses  $P_{ii}^2 \leq P_{ii}$  and  $\sum_{m=1}^n P_{im}^2 = P_{ii}$ . The final equality is  $\sum_{i=1}^n P_{ii} = \ell$ . Together, we have shown that

$$\text{var}\left(\frac{1}{n} \mathbf{a}'_j \mathbf{P}_Z \mathbf{a}_k\right) \rightarrow 0.$$

Using (12.80), (12.84), Markov's inequality (B.35), and combining across all  $j$  and  $k$  we deduce that

$$\frac{1}{n} \mathbf{a}' \mathbf{P}_Z \mathbf{a} \xrightarrow{p} \alpha \Sigma. \quad (12.85)$$

Returning to the 2SLS estimator (12.83) and combining (12.81), (12.82), and (12.85), we find

$$\hat{\beta}_{\text{2sls}} - \beta \xrightarrow{p} (\mathbf{H} + \alpha \Sigma_{22})^{-1} \alpha \Sigma_{2e}.$$

Thus 2SLS is also inconsistent for  $\beta$ . The limit, however, depends on the magnitude of  $\alpha$ .

We finally examine the LIML estimator. (12.85) implies

$$\frac{1}{n} \mathbf{Y}' \mathbf{M}_Z \mathbf{Y} = \frac{1}{n} \mathbf{a}' \mathbf{a} - \frac{1}{n} \mathbf{a}' \mathbf{P}_Z \mathbf{a} \xrightarrow{p} (1 - \alpha) \Sigma.$$

Similarly

$$\begin{aligned} \frac{1}{n} \mathbf{Y}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{Y} &= \bar{\beta}' \Gamma' \left( \frac{1}{n} \mathbf{Z}' \mathbf{Z} \right) \Gamma \bar{\beta} + \bar{\beta}' \Gamma' \left( \frac{1}{n} \mathbf{Z}' \mathbf{a} \right) + \left( \frac{1}{n} \mathbf{a}' \mathbf{Z} \right) \Gamma \bar{\beta} + \frac{1}{n} \mathbf{a}' \mathbf{P}_Z \mathbf{a} \\ &\xrightarrow{d} \bar{\beta}' \mathbf{H} \bar{\beta} + \alpha \Sigma. \end{aligned}$$

Hence

$$\begin{aligned} \hat{\mu} &= \min_{\gamma} \frac{\gamma' \mathbf{Y}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{Y} \gamma}{\gamma' \mathbf{Y}' \mathbf{M}_Z \mathbf{Y} \gamma} \\ &\xrightarrow{d} \min_{\gamma} \frac{\gamma' (\bar{\beta}' \mathbf{H} \bar{\beta} + \alpha \Sigma) \gamma}{\gamma' (1 - \alpha) \Sigma \gamma} \\ &= \frac{\alpha}{1 - \alpha} \end{aligned}$$

and

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{lml}} - \boldsymbol{\beta} &= \left( \frac{1}{n} \mathbf{X}' \mathbf{P}_Z \mathbf{X} - \widehat{\mu} \frac{1}{n} \mathbf{X}' \mathbf{M}_Z \mathbf{X} \right)^{-1} \left( \frac{1}{n} \mathbf{X}' \mathbf{P}_Z \mathbf{e} - \widehat{\mu} \frac{1}{n} \mathbf{X}' \mathbf{M}_Z \mathbf{e} \right) \\ &\xrightarrow{d} \left( \mathbf{H} + \alpha \boldsymbol{\Sigma}_{22} - \frac{\alpha}{1-\alpha} (1-\alpha) \boldsymbol{\Sigma}_{22} \right)^{-1} \left( \alpha \boldsymbol{\Sigma}_{2e} - \frac{\alpha}{1-\alpha} (1-\alpha) \boldsymbol{\Sigma}_{2e} \right) \\ &= \mathbf{H}^{-1} \mathbf{0} \\ &= \mathbf{0}.\end{aligned}$$

Thus LIML is consistent for  $\boldsymbol{\beta}$ , unlike 2SLS.

We state these results formally.

**Theorem 12.19** In model (12.77), under assumptions (12.78), (12.79) and (12.80), then as  $n \rightarrow \infty$ .

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{ols}} &\xrightarrow{p} \boldsymbol{\beta} + (\mathbf{H} + \boldsymbol{\Sigma}_{22})^{-1} \boldsymbol{\Sigma}_{2e} \\ \widehat{\boldsymbol{\beta}}_{\text{2sls}} &\xrightarrow{p} \boldsymbol{\beta} + (\mathbf{H} + \alpha \boldsymbol{\Sigma}_{22})^{-1} \alpha \boldsymbol{\Sigma}_{2e} \\ \widehat{\boldsymbol{\beta}}_{\text{lml}} &\xrightarrow{p} \boldsymbol{\beta}.\end{aligned}$$

This result is quite insightful. It shows that while endogeneity ( $\boldsymbol{\Sigma}_{2e} \neq 0$ ) renders the least-squares estimator inconsistent, the 2SLS estimator is also inconsistent if the number of instruments diverges proportionately with  $n$ . The limit in Theorem 12.19 shows a continuity between least-squares and 2SLS. The probability limit of the 2SLS estimator is continuous in  $\alpha$ , with the extreme case ( $\alpha = 1$ ) implying that 2SLS and least-squares have the same probability limit. The general implication is that the inconsistency of 2SLS is increasing in  $\alpha$ .

The theorem also shows that unlike 2SLS, the LIML estimator is consistent under the many instruments assumption. Effectively, LIML makes a bias-correction.

Theorems 12.18 (weak instruments) and 12.19 (many instruments) tell a cautionary tale. They show that when instruments are weak and/or many, that the 2SLS estimator is inconsistent. The degree of inconsistency depends on the weakness of the instruments (the magnitude of the matrix  $\mathbf{C}$  in Theorem 12.18) and the degree of overidentification (the ratio  $\alpha$  in Theorem 12.19). The Theorems also show that the LIML estimator is inconsistent under the weak instrument assumption but with a bias-correction, and is consistent under the many instrument assumption. This suggests that LIML is more robust than 2SLS to weak and many instruments.

An important limitation of the results in Theorem 12.19 is the assumption of conditional homoskedasticity. It appears likely that the consistency of LIML may fail in the many instrument setting if the errors are heteroskedastic.

In an application, users should be aware of the potential consequences of the many instrument framework. It may be useful to calculate the “many instrument ratio”  $\alpha = \ell/n$ . Unfortunately there is no known rule-of-thumb for  $\alpha$  which should lead to acceptable inference, but a minimum criterion is that if  $\alpha \geq 0.05$  you should be seriously concerned about the many-instrument problem. In general, when  $\alpha$  is large it seems preferable to use LIML instead of 2SLS.

## 12.38 Testing for Weak Instruments

In the previous sections we have found that weak instruments results in non-standard asymptotic distributions for the 2SLS and LIML estimators. In practice how do we know if this is a problem? Is there a way to test if the instruments are weak?

This question was addressed in an influential paper by Stock and Yogo (2005) as an extension of Staiger and Stock (1997). Stock-Yogo focus on two implications of weak instruments: (1) estimation bias and (2) inference distortion. They show how to test the hypothesis that these distortions are not “too big”. These tests are simply  $F$  tests for the excluded instruments in the reduced form regressions, but with non-standard critical values. In particular, when there is one endogenous regressor and a single instrument, the Stock-Yogo test rejects the null of weak instruments when this  $F$  statistic exceeds 10. While Stock and Yogo explore two types of distortions, we focus exclusively on inference as that is the more challenging problem. In this section we describe the Stock-Yogo theory and tests for the case of a single endogenous regressor ( $k_2 = 1$ ), and in the following section describe their methods for the case of multiple endogenous regressors.

While the theory in Stock and Yogo allows for an arbitrary number of exogenous regressors and instruments, for the sake of clear exposition we will focus on the very simple case of no included exogenous variables ( $k_1 = 0$ ) and just one exogenous instrument ( $\ell_2 = 1$ ), which is model (12.73) from Section 12.35

$$\begin{aligned} y_i &= x_i \beta + e_i \\ x_i &= z_i \Gamma + u_i. \end{aligned}$$

Furthermore, as in Section 12.35 we assume conditional homoskedasticity and normalize the variances as in (12.74). Since the model is just-identified the 2SLS, LIML and IV estimators are all equivalent.

The question of primary interest is to determine conditions on the reduced form under which the IV estimator of the structural equation is well behaved, and secondly, what statistical tests can be used to learn if these conditions are satisfied. As in Section 12.36 we assume that the reduced form coefficient  $\Gamma$  is **local-to-zero**, specifically

$$\Gamma = n^{-1/2} \mu.$$

The asymptotic distribution of the IV estimator is presented in Theorem 12.18. Given the simplifying assumptions the result is

$$\hat{\beta}_{\text{iv}} - \beta \xrightarrow{d} \frac{\xi_e}{\mu + \xi_2}$$

where  $(\xi_e, \xi_2)$  are bivariate normal. For inference we also examine the behavior of the classical (homoskedastic) t-statistic for the IV estimator. Note

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - x_i \hat{\beta}_{\text{iv}})^2 \\ &= \frac{1}{n} \sum_{i=1}^n e_i^2 - \frac{2}{n} \sum_{i=1}^n e_i x_i (\hat{\beta}_{\text{iv}} - \beta) + \frac{1}{n} \sum_{i=1}^n x_i^2 (\hat{\beta}_{\text{iv}} - \beta)^2 \\ &\xrightarrow{d} 1 - 2\rho \frac{\xi_e}{\mu + \xi_2} + \left( \frac{\xi_e}{\mu + \xi_2} \right)^2. \end{aligned}$$

Thus

$$T = \frac{\hat{\beta}_{\text{iv}} - \beta}{\sqrt{\hat{\sigma}^2 \sum_{i=1}^n z_i^2 / |\sum_{i=1}^n z_i x_i|}} \xrightarrow{d} \frac{\xi_1}{\sqrt{1 - 2\rho \frac{\xi_1}{\mu + \xi_2} + \left( \frac{\xi_1}{\mu + \xi_2} \right)^2}} \stackrel{\text{def}}{=} S. \quad (12.86)$$

In general,  $S$  is non-normal, and its distribution depends on the parameters  $\rho$  and  $\mu$ .

Can we use the distribution  $S$  for inference on  $\beta$ ? The distribution depends on two unknown parameters, and neither is consistently estimable. (Thus we cannot simply use the distribution in (12.86) with  $\rho$  and  $\mu$  replaced with estimates.) To eliminate the dependence on  $\rho$  one possibility is to use the “worst case” value, which turns out to be  $\rho = 1$ . By worst-case we mean that value which causes the greatest distortion away from normal critical values. Setting  $\rho = 1$  we have the considerable simplification

$$S = S_1 = \xi \left| 1 + \frac{\xi}{\mu} \right| \quad (12.87)$$

where  $\xi \sim N(0, 1)$ . When the model is strongly identified (so  $|\mu|$  is very large) then  $S_1 \approx \xi$  is standard normal, consistent with classical theory. However when  $|\mu|$  is very small (but non-zero)  $|S_1| \approx \xi^2/\mu$  (in the sense that this term dominates), which is a scaled  $\chi_1^2$  and quite far from normal. As  $|\mu| \rightarrow 0$  we find the extreme case  $|S_1| \rightarrow_p \infty$ .

While (12.87) is a convenient simplification it does not yield a useful approximation for inference since the distribution in (12.87) is highly dependent on the unknown  $\mu$ . If we try to take the worst-case value of  $\mu$ , which is  $\mu = 0$ , we find that  $|S_1|$  diverges and all distributional approximations fail.

To break this impasse, Stock and Yogo (2005) recommended a constructive alternative. Rather than using the worst-case  $\mu$ , they suggested finding a threshold such that if  $\mu$  exceeds this threshold then the distribution (12.87) is not “too badly” distorted from the normal distribution.

Specifically, the Stock-Yogo recommendation can be summarized by two steps. First, the distribution result (12.87) can be used to find a threshold value  $\tau^2$  such that if  $\mu^2 \geq \tau^2$  then the size of the nominal<sup>1</sup> 5% test “Reject if  $|T| \geq 1.96$ ” has asymptotic size  $\mathbb{P}(|S_1| \geq 1.96) \leq 0.15$ . This means that while the goal is to obtain a test with size 5%, we recognize that there may be size distortion due to weak instruments and are willing to tolerate a specific size distortion, for example 10% distortion (allow for actual size up to 15%, or more generally  $r$ ). Second, they use the asymptotic distribution of the reduced-form (first stage)  $F$  statistic to test if the actual unknown value of  $\mu^2$  exceeds the threshold  $\tau^2$ . These two steps together give rise to the rule-of-thumb that the first-stage  $F$  statistic should exceed 10 in order to achieve reliable IV inference. (This is for the case of one instrumental variable. If there is more than one instrument then the rule-of-thumb changes.) We now describe the steps behind this reasoning in more detail.

The first step is to use the distribution (12.86) to determine the threshold  $\tau^2$ . Formally, the goal is to find the value of  $\tau^2 = \mu^2$  at which the asymptotic size of a nominal 5% test is actually  $r$  (e.g.  $r = 0.15$ )

$$\mathbb{P}(|S_1| \geq 1.96) \leq r.$$

By some algebra and using the quadratic formula the event  $|\xi(1 + \xi/\mu)| < x$  is the same as

$$\frac{\mu^2}{4} - x\mu < \left(\xi + \frac{\mu}{2}\right)^2 < \frac{\mu^2}{4} + x\mu.$$

The random variable between the inequalities is distributed  $\chi_1^2(\mu^2/4)$ , a noncentral chi-square with one degree of freedom and noncentrality parameter  $\mu^2/4$ . Thus

$$\begin{aligned} \mathbb{P}(|S_1| \geq x) &= \mathbb{P}\left(\chi_1^2\left(\frac{\mu^2}{4}\right) \geq \frac{\mu^2}{4} + x\mu\right) + \mathbb{P}\left(\chi_1^2\left(\frac{\mu^2}{4}\right) \leq \frac{\mu^2}{4} - x\mu\right) \\ &= 1 - G\left(\frac{\mu^2}{4} + x\mu, \frac{\mu^2}{4}\right) + G\left(\frac{\mu^2}{4} - x\mu, \frac{\mu^2}{4}\right) \end{aligned} \quad (12.88)$$

where  $G(u, \lambda)$  is the distribution function of  $\chi_1^2(\lambda)$ . Hence the desired threshold  $\tau^2$  solves

$$1 - G\left(\frac{\tau^2}{4} + 1.96\tau, \frac{\tau^2}{4}\right) + G\left(\frac{\tau^2}{4} - 1.96\tau, \frac{\tau^2}{4}\right) = r$$

or effectively

$$G\left(\frac{\tau^2}{4} + 1.96\tau, \frac{\tau^2}{4}\right) = 1 - r$$

since  $\tau^2/4 - 1.96\tau < 0$  for relevant values of  $\tau$ . The numerical solution (computed with the non-central chi-square distribution function, e.g. `ncx2cdf` in MATLAB) is  $\tau^2 = 1.70$  when  $r = 0.15$ . (That is, the command

```
ncx2cdf(1.7/4 + 1.96 * sqrt(1.7), 1, 1.7/4)
```

---

<sup>1</sup>The term “nominal size” of a test is the official intended size – the size which would obtain under ideal circumstances. In this context the test “Reject if  $|T| \geq 1.96$ ” has nominal size 0.05 as this would be the asymptotic rejection probability in the ideal context of strong instruments.

yields the answer 0.8500. Stock and Yogo (2005) approximate the same calculation using simulation methods and report  $\tau^2 = 1.82$ .)

This calculation means that if the true reduced form coefficient satisfies  $\mu^2 \geq 1.7$ , or equivalently if  $\Gamma^2 \geq 1.7/n$ , then the (asymptotic) size of a nominal 5% test on the structural parameter is no larger than 15%.

To summarize the Stock-Yogo first step, we calculate the minimum value  $\tau^2$  for  $\mu^2$  sufficient to ensure that the asymptotic size of a nominal 5% t-test does not exceed  $r$ , and find that  $\tau^2 = 1.70$  for  $r = 0.15$ .

The Stock-Yogo second step is to find a critical value for the first-stage  $F$  statistic sufficient to reject the hypothesis that  $H_0 : \mu^2 = \tau^2$  against  $H_1 : \mu^2 > \tau^2$ . We now describe this procedure.

They suggest testing  $H_0 : \mu^2 = \tau^2$  at the 5% size using the first stage  $F$  statistic. If the  $F$  statistic is small so that the test does not reject then we should be worried that the true value of  $\mu^2$  is small and there is a weak instrument problem. On the other hand if the  $F$  statistic is large so that the test rejects then we can have some confidence that the true value of  $\mu^2$  is sufficiently large that the weak instrument problem is not too severe.

To implement the test we need to calculate an appropriate critical value. It should be calculated under the null hypothesis  $H_0 : \mu^2 = \tau^2$ . This is different from a conventional  $F$  test (which has the null hypothesis  $H_0 : \mu^2 = 0$ ).

We start by calculating the asymptotic distribution of  $F$ . Since there is just one regressor and one instrument in our simplified setting, the first-stage  $F$  statistic is the squared t-statistic from the reduced form, and given our previous calculations has the asymptotic distribution

$$F = \frac{\hat{\gamma}^2}{s(\hat{\gamma})^2} = \frac{(\sum_{i=1}^n z_i x_i)^2}{(\sum_{i=1}^n x_i^2) \hat{\sigma}_u^2} \xrightarrow{d} (\mu + \xi_2)^2 \sim \chi_1^2(\mu^2).$$

This is a non-central chi-square distribution with one degree of freedom and non-centrality parameter  $\mu^2$ . The distribution function of the latter is  $G(u, \mu^2)$ .

To test  $H_0 : \mu^2 = \tau^2$  against  $H_1 : \mu^2 > \tau^2$  we reject for  $F \geq c$  where  $c$  is selected so that the asymptotic rejection probability

$$\mathbb{P}(F \geq c) \rightarrow \mathbb{P}(\chi_1^2(\mu^2) \geq c) = 1 - G(c, \mu^2)$$

equals 0.05 under  $H_0 : \mu^2 = \tau^2$ , or equivalently

$$G(c, \tau^2) = G(c, 1.7) = 0.95.$$

This can be found using the non-central chi-square quantile function, e.g. the function  $Q(p, d)$  which solves  $G(Q(p, d), d) = p$ . We find that

$$c = Q(0.95, 1.7) = 8.7.$$

In MATLAB, this can be computed by `ncx2inv(.95, 1.7)`. (Stock and Yogo (2005) report  $c = 9.0$  since they used  $\tau^2 = 1.82$ .)

This means that if  $F > 8.7$  we can reject  $H_0 : \mu^2 = 1.7$  against  $H_1 : \mu^2 > 1.7$  with an asymptotic 5% test. In this context we should expect the IV estimate and tests to be reasonably well behaved. However, if  $F < 8.7$  then we should be cautious about the IV estimator, confidence intervals, and tests. This finding led Staiger and Stock (1997) to propose the informal “rule of thumb” that the first stage  $F$  statistic should exceed 10. Notice that  $F$  exceeding 8.7 (or 10) is equivalent to the reduced form t-statistic exceeding 2.94 (or 3.16), which is considerably larger than a conventional check if the t-statistic is “significant”. Equivalently, the recommended rule-of-thumb for the case of a single instrument is to estimate the reduced form and verify that the t-statistic for exclusion of the instrumental variable exceeds 3 in absolute value.

Does the proposed procedure control the asymptotic size of a 2SLS test? The first step has asymptotic size bounded below  $r$  (e.g. 15%). The second step has asymptotic size 5%. By the Bonferroni bound (see Section 9.20) the two steps together have asymptotic size bounded below  $r + 0.05$  (e.g. 20%). We can thus call the Stock-Yogo procedure a rigorous test with asymptotic size  $r + 0.05$  (or 20%).

Our analysis has been confined to the case  $k_2 = \ell_2 = 1$ . Stock and Yogo (2005) also examine the case of  $\ell_2 > 1$  (which requires numerical simulation to solve), and both the 2SLS and LIML estimators. They

show that the  $F$  statistic critical values depend on the number of instruments  $\ell_2$  as well as the estimator. We report their calculations in Table 12.4.

Table 12.4: 5% Critical Value for Weak Instruments,  $k_2 = 1$

$\ell_2$	Maximal Size $r$							
	2SLS				LIML			
	0.10	0.15	0.20	0.25	0.10	0.15	0.20	0.25
1	16.4	9.0	6.7	5.5	16.4	9.0	6.7	5.5
2	19.9	11.6	8.7	7.2	8.7	5.3	4.4	3.9
3	22.3	12.8	9.5	7.8	6.5	4.4	3.7	3.3
4	24.6	14.0	10.3	8.3	5.4	3.9	3.3	3.0
5	26.9	15.1	11.0	8.8	4.8	3.6	3.0	2.8
6	29.2	16.2	11.7	9.4	4.4	3.3	2.9	2.6
7	31.5	17.4	12.5	9.9	4.2	3.2	2.7	2.5
8	33.8	18.5	13.2	10.5	4.0	3.0	2.6	2.4
9	36.2	19.7	14.0	11.1	3.8	2.9	2.5	2.3
10	38.5	20.9	14.8	11.6	3.7	2.8	2.5	2.2
15	50.4	26.8	18.7	12.2	3.3	2.5	2.2	2.0
20	62.3	32.8	22.7	17.6	3.2	2.3	2.1	1.9
25	74.2	38.8	26.7	20.6	3.8	2.2	2.0	1.8
30	86.2	44.8	30.7	23.6	3.9	2.2	1.9	1.7

One striking feature about these critical values is that those for the 2SLS estimator are strongly increasing in  $\ell_2$  while those for the LIML estimator are decreasing in  $\ell_2$ . This means that when the number of instruments  $\ell_2$  is large, 2SLS requires a much stronger reduced form (larger  $\mu^2$ ) in order for inference to be reliable, but this is not the case for LIML. This is direct evidence that inference is less sensitive to weak instruments when estimation is by LIML rather than 2SLS. This makes a strong case for using LIML rather than 2SLS, especially when  $\ell_2$  is large or the instruments are potentially weak.

We now summarize the recommended Staiger-Stock/Stock-Yogo procedure for  $k_1 \geq 1$ ,  $k_2 = 1$ , and  $\ell_2 \geq 1$ . The structural equation and reduced form equations are

$$\begin{aligned} y_i &= \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + x_{2i} \beta_2 + e_i \\ x_{2i} &= \mathbf{x}'_{1i} \boldsymbol{\gamma}_1 + \mathbf{z}'_{2i} \boldsymbol{\gamma}_2 + u_i. \end{aligned}$$

The reduced form is estimated by least-squares

$$x_{2i} = \mathbf{x}'_{1i} \hat{\boldsymbol{\gamma}}_1 + \mathbf{z}'_{2i} \hat{\boldsymbol{\gamma}}_2 + \hat{u}_i$$

and the structural equation by either 2SLS or LIML:

$$y_i = \mathbf{x}'_{1i} \hat{\boldsymbol{\beta}}_1 + x_{2i} \hat{\beta}_2 + \hat{e}_i.$$

Let  $F$  be the  $F$  statistic for  $H_0 : \boldsymbol{\gamma}_2 = 0$  in the reduced form equation. Let  $s(\hat{\beta}_2)$  be a standard error for  $\beta_2$  in the structural equation. The procedure is:

1. Compare  $F$  with the critical values  $c$  in the above table, with the row selected to match the number of excluded instruments  $\ell_2$ , and the columns to match the estimation method (2SLS or LIML) and the desired size  $r$ .
2. If  $F > c$  then report the 2SLS or LIML estimates with conventional inference.

The Stock-Yogo test can be implemented in Stata using the command `estat firststage after ivregress 2sls` or `ivregres liml` if a standard (non-robust) covariance matrix has been specified (that is, without the '`,r`' option).

There are possible extensions to the Stock-Yogo procedure.

One modest extension is to use the information to convey the degree of confidence in the accuracy of a confidence interval. Suppose in an application you have  $\ell_2 = 5$  excluded instruments and have estimated your equation by 2SLS. Now suppose that your reduced form  $F$  statistic equals 12. You check the Stock-Yogo table, and find that  $F = 12$  is significant with  $r = 0.20$ . Thus we can interpret the conventional 2SLS confidence interval as having coverage of 80% (or 75% if we make the Bonferroni correction). On the other hand if  $F = 27$  we would conclude that the test for weak instruments is significant with  $r = 0.10$ , meaning that the conventional 2SLS confidence interval can be interpreted as having coverage of 90% (or 85% after Bonferroni correction).

A more substantive extension, which we now discuss, reverses the steps. Unfortunately this discussion will be limited to the case  $\ell_2 = 1$ , where 2SLS and LIML are equivalent. First, use the reduced form  $F$  statistic to find a one-sided confidence interval for  $\mu^2$  of the form  $[\mu_L^2, \infty)$ . Second, use the lower bound  $\mu_L^2$  to calculate a critical value  $c$  for  $S_1$  such that the 2SLS test has asymptotic size bounded below 0.05. This produces better size control than the Stock-Yogo procedure and produces more informative confidence intervals for  $\beta_2$ . We now describe the steps in detail.

The first goal is to find a one-sided confidence interval for  $\mu^2$ . This is found by test inversion. As we described earlier, for any  $\tau^2$  we reject  $H_0 : \mu^2 = \tau^2$  in favor of  $H_1 : \mu^2 > \tau^2$  if  $F > c$  where  $G(c, \tau^2) = 0.95$ . Equivalently, we reject if  $G(F, \tau^2) > 0.95$ . By the test inversion principle, an asymptotic 95% confidence interval  $[\mu_L^2, \infty)$  can be formed as the set of all values of  $\tau^2$  which are not rejected by this test. Since  $G(F, \tau^2) \geq 0.95$  for all  $\tau^2$  in this set, the lower bound  $\mu_L^2$  satisfies  $G(F, \mu_L^2) = 0.95$ . The lower bound is found from this equation. Since this solution is not generally programmed, it needs to be found numerically. In MATLAB, the solution is  $\text{mu2}$  when  $\text{ncx2cdf}(F, 1, \text{mu2})$  returns 0.95.

The second goal is to find the critical value  $c$  such that  $\mathbb{P}(|S_1| \geq c) = 0.05$  when  $\mu^2 = \mu_L^2$ . From (12.88), this is achieved when

$$1 - G\left(\frac{\mu_L^2}{4} + c\mu_L, \frac{\mu_L^2}{4}\right) + G\left(\frac{\mu_L^2}{4} - c\mu_L, \frac{\mu_L^2}{4}\right) = 0.05. \quad (12.89)$$

This can be solved as

$$G\left(\frac{\mu_L^2}{4} + c\mu_L, \frac{\mu_L^2}{4}\right) = 0.95.$$

(The third term on the left-hand-side of (12.89) is zero for all solutions so can be ignored.) Using the non-central chi-square quantile function  $Q(p, d)$ , this  $C$  equals

$$c = \frac{Q\left(0.95, \frac{\mu_L^2}{4}\right) - \frac{\mu_L^2}{4}}{\mu_L}.$$

For example, in MATLAB this is found as  $c = (\text{ncx2inv}(0.95, 1, \text{mu2}/4) - \text{mu2}/4) / \sqrt{\text{mu2}}$ . 95% confidence intervals for  $\beta_2$  are then calculated as

$$\hat{\beta}_{IV} \pm cs(\hat{\beta}_{IV}).$$

We can also calculate a p-value for the t-statistic  $T$  for  $\beta_2$ . These are

$$p = 1 - G\left(\frac{\mu_L^2}{4} + |T|\mu_L, \frac{\mu_L^2}{4}\right) + G\left(\frac{\mu_L^2}{4} - |T|\mu_L, \frac{\mu_L^2}{4}\right)$$

where the third term equals zero if  $|T| \geq \mu_L/4$ . In MATLAB, for example, this can be calculated by the commands

```
T1 = mu2/4 + abs(T) * sqrt(mu2);
T2 = mu2/4 - abs(T) * sqrt(mu2);
p = -ncx2cdf(T1, 1, mu2/4) + ncx2cdf(T2, 1, mu2/4);
```

These confidence intervals and p-values will be larger than the conventional intervals and p-values, reflecting the incorporation of information about the strength of the instruments through the first-stage

$F$  statistic. Also, by the Bonferroni bound these tests have asymptotic size bounded below 10% and the confidence intervals have asymptotic coverage exceeding 90%, unlike the Stock-Yogo method which has size of 20% and coverage of 80%.

The augmented procedure suggested here, only for the  $\ell_2 = 1$  case, is

1. Find  $\mu_L^2$  which solves  $G(F, \mu_L^2) = 0.95$ . In MATLAB, the solution is `mu2` when `ncx2cdf(F, 1, mu2)` returns 0.95.
2. Find  $c$  which solves  $G(\mu_L^2/4 + c\mu_L, \mu_L^2/4) = 0.95$ . In MATLAB, the command is  
`c=(ncx2inv(.95, 1, mu2/4)-mu2/4)/sqrt(mu2)`
3. Report the confidence interval  $\hat{\beta}_2 \pm cs(\hat{\beta}_2)$  for  $\beta_2$ .
4. For the t statistic  $T = (\hat{\beta}_2 - \beta_2) / s(\hat{\beta}_2)$  the asymptotic p-value is

$$p = 1 - G\left(\frac{\mu_L^2}{4} + |T|\mu_L, \frac{\mu_L^2}{4}\right) + G\left(\frac{\mu_L^2}{4} - |T|\mu_L, \frac{\mu_L^2}{4}\right)$$

which is computed in MATLAB by `T1=mu2/4+abs(T)*sqrt(mu2); T2=mu2/4-abs(T)*sqrt(mu2);` and `p=1-ncx2cdf(T1, 1, mu2/4)+ncx2cdf(T2, 1, mu2/4).`

We have described an extension to the Stock-Yogo procedure for the case of one instrumental variable  $\ell_2 = 1$ . This restriction was due to the use of the analytic formula (12.89) for the asymptotic distribution, which is only available when  $\ell_2 = 1$ . In principle the procedure could be extended using simulation or bootstrap methods, but this has not been done to my knowledge.

To illustrate the Stock-Yogo and extended procedures, let us return to the Card proximity example. First, let's take the IV estimates reported in the second column of Table 12.1 which used *college* proximity as a single instrument. The reduced form estimates for the endogenous variable *education* is reported in the second column of Table 12.2. The excluded instrument *college* has a t-ratio of 4.2 which implies an  $F$  statistic of 17.8. The  $F$  statistic exceeds the rule-of thumb of 10, so the structural estimates pass the Stock-Yogo threshold. Based on the Stock-Yogo recommendation, this means that we can interpret the estimates conventionally. However, the conventional confidence interval, e.g. for the returns to education,  $0.132 \pm 0.049 * 1.96 = [0.04, 0.23]$  has an asymptotic coverage of 80%, rather than the nominal 95% rate.

Now consider the extended procedure. Given  $F = 17.8$  we can calculate the lower bound  $\mu_L^2 = 6.6$ . This implies a critical value of  $C = 2.7$ . Hence an improved confidence interval for the returns to education in this equation is  $0.132 \pm 0.049 * 2.7 = [0.01, 0.26]$ . This is a wider confidence interval, but has improved asymptotic coverage of 90%. The p-value for  $\beta_2 = 0$  is  $p = 0.012$ .

Next, let's take the 2SLS estimates reported in the fourth column of Table 11.1 which use the two instruments *public* and *private*. The reduced form equation is reported in column six of Table 12.2. An  $F$  statistic for exclusion of the two instruments is  $F = 13.9$ , which exceeds the 15% size threshold for 2SLS and all thresholds for LIML, indicating that the structural estimates pass the Stock-Yogo threshold test and can be interpreted conventionally.

The weak instrument methods described here are important for applied econometrics as they discipline researchers to assess the quality of their reduced form relationships before reporting structural estimates. The theory, however, has limitations and shortcomings. A major limitation is that the theory requires the strong assumption of conditional homoskedasticity. Despite this theoretical limitation, in practice researchers apply the Stock-Yogo recommendations to estimates computed with heteroskedasticity-robust standard errors as it is the currently the best known approach. This is an active area of research so the recommended methods may change in the years ahead.

### 12.39 Weak Instruments with $k_2 > 1$

When there are more than one endogenous regressor ( $k_2 > 1$ ) it is better to examine the reduced form as a system. Staiger and Stock (1997) and Stock and Yogo (2005) provided an analysis of this case and constructed a test for weak instruments. The theory is considerably more involved than the  $k_2 = 1$  case, so we briefly summarize it here excluding many details, emphasizing their suggested methods.

The structural equation and reduced form equations are

$$\begin{aligned} y_i &= \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + e_i \\ \mathbf{x}_{2i} &= \boldsymbol{\Gamma}'_{12} \mathbf{z}_{1i} + \boldsymbol{\Gamma}'_{22} \mathbf{z}_{2i} + \mathbf{u}_{2i}. \end{aligned}$$

As in the previous section we assume that the errors are conditionally homoskedastic.

Identification of  $\boldsymbol{\beta}_2$  requires the matrix  $\boldsymbol{\Gamma}_{22}$  to be full rank. A necessary condition is that each row of  $\boldsymbol{\Gamma}'_{22}$  is non-zero, but this is not sufficient.

We focus on the size performance of the homoskedastic Wald statistic for the 2SLS estimator of  $\boldsymbol{\beta}_2$ . For simplicity assume that the variance of  $e_i$  is known and normalized to one. Using representation (12.34), the Wald statistic can be written as

$$W = \mathbf{e}' \tilde{\mathbf{Z}}_2 \left( \tilde{\mathbf{Z}}_2' \tilde{\mathbf{Z}}_2 \right)^{-1} \tilde{\mathbf{Z}}_2' \mathbf{X}_2 \left( \mathbf{X}_2' \tilde{\mathbf{Z}}_2 \left( \tilde{\mathbf{Z}}_2' \tilde{\mathbf{Z}}_2 \right)^{-1} \tilde{\mathbf{Z}}_2' \mathbf{X}_2 \right)^{-1} \left( \mathbf{X}_2' \tilde{\mathbf{Z}}_2 \left( \tilde{\mathbf{Z}}_2' \tilde{\mathbf{Z}}_2 \right)^{-1} \tilde{\mathbf{Z}}_2' \mathbf{e} \right)$$

where  $\tilde{\mathbf{Z}}_2 = (\mathbf{I}_n - \mathbf{P}_1) \mathbf{Z}_2$  and  $\mathbf{P}_1 = \mathbf{X}_1 (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1'$ .

Recall from Section 12.36 that Stock and Staiger model the excluded instruments  $\mathbf{z}_{2i}$  as weak by setting  $\boldsymbol{\Gamma}_{22} = n^{-1/2} \mathbf{C}$  for some matrix  $\mathbf{C}$ . In this framework we have the asymptotic distribution results

$$\frac{1}{n} \tilde{\mathbf{Z}}_2' \tilde{\mathbf{Z}}_2 \xrightarrow{p} \mathbf{Q} = \mathbb{E}(\mathbf{z}_{2i} \mathbf{z}_{2i}') - \mathbb{E}(\mathbf{z}_{2i} \mathbf{z}_{1i}') (\mathbb{E}(\mathbf{z}_{1i} \mathbf{z}_{1i}'))^{-1} \mathbb{E}(\mathbf{z}_{1i} \mathbf{z}_{2i}')$$

and

$$\frac{1}{\sqrt{n}} \tilde{\mathbf{Z}}_2' \mathbf{e} \xrightarrow{d} \mathbf{Q}^{1/2} \xi_0$$

where  $\xi_0$  is a matrix normal variate whose columns are independent  $N(\mathbf{0}, \mathbf{I})$ . Furthermore, setting  $\Sigma = \mathbb{E}(\mathbf{u}_{2i} \mathbf{u}_{2i}')$  and  $\bar{\mathbf{C}} = \mathbf{Q}^{1/2} \mathbf{C} \Sigma^{-1/2}$ ,

$$\frac{1}{\sqrt{n}} \tilde{\mathbf{Z}}_2' \mathbf{X}_2 = \frac{1}{n} \tilde{\mathbf{Z}}_2' \tilde{\mathbf{Z}}_2 \mathbf{C} + \frac{1}{\sqrt{n}} \tilde{\mathbf{Z}}_2' \mathbf{U}_2 \xrightarrow{d} \mathbf{Q}^{1/2} \bar{\mathbf{C}} \Sigma^{1/2} + \mathbf{Q}^{1/2} \xi_2 \Sigma^{1/2}$$

where  $\xi_2$  is a matrix normal variate whose columns are independent  $N(\mathbf{0}, \mathbf{I})$ . The variables  $\xi_0$  and  $\xi_2$  are correlated. Together we obtain the asymptotic distribution of the Wald statistic

$$W \xrightarrow{d} S = \xi_0' (\bar{\mathbf{C}} + \xi_2) (\bar{\mathbf{C}}' \bar{\mathbf{C}})^{-1} (\bar{\mathbf{C}} + \xi_2)' \xi_0.$$

Using the spectral decomposition,  $\bar{\mathbf{C}}' \bar{\mathbf{C}} = \mathbf{H}' \Lambda \mathbf{H}$  where  $\mathbf{H}' \mathbf{H} = \mathbf{I}$  and  $\Lambda$  is diagonal. Thus we can write

$$S = \xi_0' \bar{\mathbf{C}} \Lambda^{-1} \bar{\mathbf{C}}' \xi_0$$

where  $\bar{\xi}_2 = \bar{\mathbf{C}} \mathbf{H}' + \xi_2 \mathbf{H}'$ . The matrix  $\xi^* = (\xi_0, \bar{\xi}_2)$  is multivariate normal, so  $\xi^* \xi^*$  has what is called a non-central Wishart distribution. It only depends on the matrix  $\bar{\mathbf{C}}$  through  $\mathbf{H} \bar{\mathbf{C}}' \bar{\mathbf{C}} \mathbf{H}' = \Lambda$ , which are the eigenvalues of  $\bar{\mathbf{C}}' \bar{\mathbf{C}}$ . Since  $S$  is a function of  $\xi^*$  only through  $\bar{\xi}_2' \xi_0$  we conclude that  $S$  is a function of  $\bar{\mathbf{C}}$  only through these eigenvalues.

This is a very quick derivation of a rather involved derivation, but the conclusion drawn by Stock and Yogo is that the asymptotic distribution of the Wald statistic is non-standard, and a function of the model parameters only through the eigenvalues of  $\bar{\mathbf{C}}' \bar{\mathbf{C}}$  and the correlations between the normal variates  $\xi_0$  and  $\bar{\xi}_2$ . The worst-case can be summarized by the maximal correlation between  $\xi_0$  and  $\bar{\xi}_2$  and the smallest

eigenvalue of  $\bar{\mathbf{C}}'\bar{\mathbf{C}}$ . For convenience, they rescale the latter by dividing by the number of endogenous variables. Define

$$\mathbf{G} = \bar{\mathbf{C}}'\bar{\mathbf{C}}/k_2 = \Sigma^{-1/2}\mathbf{C}'\mathbf{Q}\mathbf{C}\Sigma^{-1/2}/k_2$$

and

$$g = \lambda_{\min}(\mathbf{G}) = \lambda_{\min}(\Sigma^{-1/2}\mathbf{C}'\mathbf{Q}\mathbf{C}\Sigma^{-1/2})/k_2.$$

This can be estimated from the reduced-form regression

$$\mathbf{x}_{2i} = \hat{\Gamma}'_{12}\mathbf{z}_{1i} + \hat{\Gamma}'_{22}\mathbf{z}_{2i} + \hat{\mathbf{u}}_{2i}.$$

The estimator is

$$\begin{aligned}\hat{\mathbf{G}} &= \hat{\Sigma}^{-1/2}\hat{\Gamma}'_{22}\left(\tilde{\mathbf{Z}}'_2\tilde{\mathbf{Z}}_2\right)\hat{\Gamma}_{22}\hat{\Sigma}^{-1/2}/k_2 \\ &= \hat{\Sigma}^{-1/2}\left(\mathbf{X}'_2\tilde{\mathbf{Z}}_2\left(\tilde{\mathbf{Z}}'_2\tilde{\mathbf{Z}}_2\right)^{-1}\tilde{\mathbf{Z}}'_2\mathbf{X}_2\right)\hat{\Sigma}^{-1/2}/k_2 \\ \hat{\Sigma} &= \frac{1}{n-k}\sum_{i=1}^n \hat{\mathbf{u}}_{2i}\hat{\mathbf{u}}'_{2i} \\ \hat{g} &= \lambda_{\min}(\hat{\mathbf{G}}).\end{aligned}$$

$\hat{\mathbf{G}}$  is a matrix  $F$ -type statistic for the coefficient matrix  $\hat{\Gamma}_{22}$ .

The statistic  $\hat{g}$  was proposed by Craig and Donald (1993) as a test for underidentification. Stock and Yogo (2005) use it as a test for weak instruments. Using simulation methods, they determined critical values for  $\hat{g}$  similar to those for the  $k_2 = 1$  case. For given size  $r > 0.05$ , there is a critical value  $c$  (reported in the table below) such that if  $\hat{g} > c$ , then the 2SLS (or LIML) Wald statistic  $W$  for  $\hat{\beta}_2$  has asymptotic size bounded below  $r$ . On the other hand, if  $\hat{g} \leq c$  then we cannot bound the asymptotic size below  $r$  and we cannot reject the hypothesis of weak instruments.

The Stock-Yogo critical values for  $k_2 = 2$  are presented in Table 12.5. The methods and theory applies to the cases  $k_2 > 2$  as well, but those critical values have not been calculated. As for the  $k_2 = 1$  case, the critical values for 2SLS are dramatically increasing in  $\ell_2$ . Thus when the model is over-identified, we need quite a large value of  $\hat{g}$  to reject the hypothesis of weak instruments. This is a strong cautionary message to check the  $\hat{g}$  statistic in applications. Furthermore, the critical values for LIML are generally decreasing in  $\ell_2$  (except for  $r = 0.10$ , where the critical values are increasing for large  $\ell_2$ ). This means that for over-identified models, LIML inference is much less sensitive to weak instruments than 2SLS, and may be the preferred estimation method.

The Stock-Yogo test can be implemented in Stata using the command `estat firststage after ivregress 2sls` or `ivregres liml` if a standard (non-robust) covariance matrix has been specified (that is, without the '`,r`' option). Critical values which control for size are only available for  $k_2 \leq 2$ . For  $k_2 > 2$  critical values which control for relative bias are reported.

Robust versions of the test have been proposed by Kleibergen and Paap (2006). These can be implemented in Stata using the downloadable command `ivreg2`.

## 12.40 Example: Acemoglu, Johnson and Robinson (2001)

One particularly well-cited instrumental variable regression is in Acemoglu, Johnson and Robinson (2001) with additional details published in (2012). They are interested in the effect of political institutions on economic performance. The theory is that good institutions (rule-of-law, property rights) should result in a country having higher long-term economic output than if the same country had poor institutions. To investigate this question, they focus on a sample of 64 former European colonies. Their data is in the file `AJR2001` on the textbook website.

The authors' premise is that modern political institutions will have been influenced by the colonizing country. In particular, they argue that colonizing countries tended to set up colonies as either an "extractive state" or as a "migrant colony". An extractive state was used by the colonizer to extract resources for

Table 12.5: 5% Critical Value for Weak Instruments,  $k_2 = 2$ 

$\ell_2$	Maximal Size $r$							
	2SLS				LIML			
	0.10	0.15	0.20	0.25	0.10	0.15	0.20	0.25
2	7.0	4.6	3.9	3.6	7.0	4.6	3.9	3.6
3	13.4	8.2	6.4	5.4	5.4	3.8	3.3	3.1
4	16.9	9.9	7.5	6.3	4.7	3.4	3.0	2.8
5	19.4	11.2	8.4	6.9	4.3	3.1	2.8	2.6
6	21.7	12.3	9.1	7.4	4.1	2.9	2.6	2.5
7	23.7	13.3	9.8	7.9	3.9	2.8	2.5	2.4
8	25.6	14.3	10.4	8.4	3.8	2.7	2.4	2.3
9	27.5	15.2	11.0	8.8	3.7	2.7	2.4	2.2
10	29.3	16.2	11.6	9.3	3.6	2.6	2.3	2.1
15	38.0	20.6	14.6	11.6	3.5	2.4	2.1	2.0
20	46.6	25.0	17.6	13.8	3.6	2.4	2.0	1.9
25	55.1	29.3	20.6	16.1	3.6	2.4	1.97	1.8
30	63.5	33.6	23.5	18.3	4.1	2.4	1.95	1.7

the colonizing country, but was not largely settled by the European colonists. In this case the colonists would have had no incentive to set up good political institutions. In contrast, if a colony was set up as a “migrant colony”, then large numbers of European settlers migrated to the colony to live. These settlers would have desired institutions similar to those in their home country, and hence would have had a positive incentive to set up good political institutions. The nature of institutions is quite persistent over time, so these 19<sup>th</sup>-century foundations would affect the nature of modern institutions. The authors conclude that the 19<sup>th</sup>-century nature of the colony should be predictive of the nature of modern institutions, and hence modern economic growth.

To start the investigation they report an OLS regression of log GDP per capita in 1995 on a measure of political institutions they call “risk”, which is a measure of the protection against expropriation risk. This variable ranges from 0 to 10, with 0 the lowest protection against appropriation, and 10 the highest. For each country the authors take the average value of the index over 1985 to 1995 (the mean is 6.5 with a standard deviation of 1.5). Their reported OLS estimates (intercept omitted) are

$$\widehat{\log(GDP \text{ per Capita})} = 0.52 \text{ risk.} \quad (12.90)$$

$$(0.06)$$

These estimates imply a 52% difference in GDP between countries with a 1-unit difference in *risk*.

The authors argue that the *risk* is likely endogenous, since economic output influences political institutions, and because the variable *risk* is undoubtedly measured with error. These issues induce least-square bias in different directions and thus the overall bias effect is unclear.

To correct for the endogeneity bias the authors argue the need for an instrumental variable which does not directly affect economic performance yet is associated with political institutions. Their innovative suggestion was to use the mortality rate which faced potential European settlers in the 19<sup>th</sup> century. Colonies with high expected mortality would have been less attractive to European settlers, resulting in lower levels of European migrants. As a consequence the authors expect such colonies to have been more likely structured as an extractive state rather than a migrant colony. To measure the expected mortality rate the authors use estimates provided by historical research of the annualized deaths per 1000 soldiers, labeled *mortality*. (They used military mortality rates as the military maintained high-quality records.)

The first-stage regression is

$$risk = -0.61 \log(mortality) + \hat{u}. \quad (12.91)$$

(0.13)

These estimates confirm that 19<sup>th</sup>-century high settler mortality rates are associated with countries with lower quality modern institutions. Using  $\log(mortality)$  as an instrument for  $risk$ , they estimate the structural equation using 2SLS and report

$$\widehat{\log(GDP \text{ per Capita})} = 0.94 risk. \quad (12.92)$$

(0.16)

This estimate is much higher than the OLS estimate from (12.90). The estimate is consistent with a near doubling of GDP due to a 1-unit difference in the risk index.

These are simple regressions involving just one right-hand-side variable. The authors considered a range of other models. Included in these results are a reversal of a traditional finding. In a conventional (least-squares) regression two relevant variables for output are *latitude* (distance from the equator) and *africa* (a dummy variable for countries from Africa), both of which are difficult to interpret causally. But in the proposed instrumental variables regression the variables *latitude* and *africa* have much smaller – and statistically insignificant – coefficients.

To assess the specification, we can use the Stock-Yogo and endogeneity tests. The Stock-Yogo test is from the reduced form (12.91). The instrument has a t-ratio of 4.8 (or  $F = 23$ ) which exceeds the Stock-Yogo critical value and hence can be treated as strong. For an endogeneity test, we take the least-squares residual  $\hat{u}$  from this equation and include it in the structural equation and estimate by least-squares. We find a coefficient on  $\hat{u}$  of  $-0.57$  with a t-ratio of 4.7, which is highly significant. We conclude that the least-squares and 2SLS estimates are statistically different, and reject the hypothesis that the variable *risk* is exogenous for the GDP structural equation.

In Exercise 12.23 you will replicate and extend these results using the authors' data.

This paper is a creative and careful use of the instrumental variables method. The creativity stems from the historical analysis which lead to the focus on mortality as a potential predictor of migration choices. The care comes in the implementation, as the authors needed to gather country-level data on political institutions and mortality from distinct sources. Putting these pieces together is the art of the project.

## 12.41 Example: Angrist and Krueger (1991)

Another influential instrument variable regression is in Angrist and Krueger (1991). Their concern, similar to Card (1995), is estimation of the structural returns to education while treating educational attainment as endogenous. Like Card, their goal is to find an instrument which is exogenous for wages yet has an impact on educational attainment. A subset of their data in the file AK1991 on the textbook website.

Their creative suggestion was to focus on compulsory school attendance policies and their interaction with birthdates. Compulsory schooling laws vary across states in the United States, but typically require that youth remain in school until their sixteenth or seventeenth birthday. Angrist and Krueger argue that compulsory schooling has a causal effect on wages – youth who would have chosen to drop out of school stay in school for more years – and thus have more education which causally impacts their earnings as adults.

Angrist and Krueger next observe that these policies have differential impact on youth who are born early or late in the school year. Students who are born early in the calendar year are typically older when they enter school. Consequently when they attain the legal dropout age they have attended less school than those born near the end of the year. This means that birthdate (early in the calendar year versus late)

exogenously impacts educational attainment, and thus wages through education. Yet birthdate must be exogenous for the structural wage equation, as there is no reason to believe that birthdate itself has a causal impact on a person's ability or wages. These considerations together suggest that birthdate is a valid instrumental variable for education in a causal wage equation.

Typical wage datasets include age, but not birthdates. To obtain information on birthdate, Angrist and Krueger used U.S. Census data which includes an individual's quarter of birth (January-March, April-June, etc.). They use this variable to construct 2SLS estimates of the return to education.

Their paper carefully documents that educational attainment varies by quarter of birth (as predicted by the above discussion), and reports a large set of least-squares and 2SLS estimates. We focus on two estimates at the core of their analysis, reported in column (6) of their Tables V and VII. This involves data from the 1980 census with men born in 1930-1939, with 329,509 observations. The first equation is

$$\widehat{\log(wage)} = 0.081 \text{ } edu - 0.230 \text{ } black + 0.158 \text{ } urban + 0.244 \text{ } married \quad (12.93)$$

(0.016)	(0.026)	(0.017)	(0.005)
---------	---------	---------	---------

where *edu* years of education, and *black*, *urban*, and *married* are dummy variables indicating race (1 if black, 0 otherwise), lives in a metropolitan area, and if married. In addition to the reported coefficients, the equation also includes as regressors nine year-of-birth dummies and eight region-of-residence dummies. The equation is estimated by 2SLS. The instrumental variables are the 30 interactions of three quarter-of-birth times ten year-of-birth dummy variables.

This equation indicates an 8% increase in wages due to each year of education.

Angrist and Krueger observe that the effect of compulsory education laws are likely to vary across states, so expand the instrument set to include interactions with state-of-birth. They estimate the following equation by 2SLS

$$\widehat{\log(wage)} = 0.083 \text{ } edu - 0.233 \text{ } black + 0.151 \text{ } urban + 0.244 \text{ } married. \quad (12.94)$$

(0.009)	(0.011)	(0.009)	(0.004)
---------	---------	---------	---------

This equation also adds fifty state-of-birth dummy variables as regressors. The instrumental variables are the 180 interactions of quarter-of-birth times year-of-birth dummy variables, plus quarter-of-birth times state-of-birth interactions.

This equation shows a similar estimated causal effect of education on wages as in (12.93). More notably, the standard error is smaller in (12.94), suggesting improved precision by the expanded instrumental variable set.

However, these estimates seem excellent candidates for weak instruments and many instruments. Indeed, this paper (published in 1991) helped spark these two literatures. We can use the Stock-Yogo tools to explore the instrument strength and the implications for the Angrist-Krueger estimates.

We first take equation (12.93). Using the original Angrist-Krueger data, we estimate the corresponding reduced form, and calculate the *F* statistic for the 30 excluded instruments. We find *F* = 4.8. It has an asymptotic p-value of 0.000, suggesting that we can reject (at any significance level) the hypothesis that the coefficients on the excluded instruments are zero. Thus Angrist and Krueger appear to be correct that quarter of birth helps to explain educational attainment and are thus a valid instrumental variable set. However, using the Stock-Yogo test, *F* = 4.8 is not high enough to reject the hypothesis that the instruments are weak. Specifically, for  $\ell_2 = 30$  the critical value for the *F* statistic is 45 (if we want to bound size below 15%). The actual value of 4.8 is far below 45. Since we cannot reject that the instruments are weak, this indicates that we cannot interpret the 2SLS estimates and test statistics in (12.93) as reliable.

Second, take (12.94) with the expanded regressor and instrument set. Estimating the corresponding reduced form, we find the *F* statistic for the 180 excluded instruments is *F* = 2.43 which also has an asymptotic p-value of 0.000 indicating that we can reject at any significance level the hypothesis that the excluded instruments have no effect on educational attainment. However, using the Stock-Yogo test we also cannot reject the hypothesis that the instruments are weak. While Stock and Yogo did not calculate

the critical values for  $\ell_2 = 180$ , the 2SLS critical values are increasing in  $\ell_2$  so we can use those for  $\ell_2 = 30$  as a lower bound. Hence the observed value of  $F = 2.43$  is far below the level needed for significance. Consequently the results in (12.94) cannot be viewed as reliable. In particular, the observation that the standard errors in (12.94) are smaller than those in (12.93) should not be interpreted as evidence of greater precision. Rather, they should be viewed as evidence of unreliability due to weak instruments.

When instruments are weak, one constructive suggestion is to use LIML estimation rather than 2SLS. Another constructive suggestion is to alter the instrument set. While Angrist and Krueger used a large number of instrumental variables, we can consider using a smaller set. Take equation (12.93). Rather than estimating it using the 30 interaction instruments, consider using only the three quarter-of-birth dummy variables. We report the reduced form estimates here:

$$\widehat{edu} = -1.57 \text{ black} + 1.05 \text{ urban} + 0.225 \text{ married} + 0.050 Q_2 + 0.101 Q_3 + 0.142 Q_4$$

(0.02)	(0.01)	(0.016)	(0.016)	(0.016)	(0.016)
--------	--------	---------	---------	---------	---------

(12.95)

where  $Q_2$ ,  $Q_3$  and  $Q_4$  are dummy variables for birth in the 2<sup>nd</sup>, 3<sup>rd</sup>, and 4<sup>th</sup> quarter. The regression also includes nine year-of-birth and eight region-of-residence dummy variables.

The reduced form coefficients in (12.95) on the quarter-of-birth dummies are quite instructive. The coefficients are positive and increasing, consistent with the Angrist-Krueger hypothesis that individuals born later in the year achieve higher average education. Focusing on the weak instrument problem, the  $F$  test for exclusion of these three variables is  $F = 31$ . The Stock-Yogo critical value is 12.8 for  $\ell_2 = 3$  and a size of 15%, and is 22.3 for a size of 10%. Since  $F = 31$  exceeds both these thresholds we can reject the hypothesis that this reduced form is weak. Estimating the model by 2SLS with these three instruments we find

$$\widehat{\log(wage)} = 0.099 \text{ edu} - 0.201 \text{ black} + 0.139 \text{ urban} + 0.240 \text{ married}. \quad (12.96)$$

(0.021)	(0.033)	(0.022)	(0.006)
---------	---------	---------	---------

These estimates indicate a slightly larger (10%) causal impact of education on wages, but with a larger standard error. The Stock-Yogo analysis indicates that we can interpret the confidence intervals from these estimates as having asymptotic coverage 85%.

While the original Angrist-Krueger estimates suffer due to weak instruments, their paper is a very creative and thoughtful application of the **natural experiment** methodology. They discovered a completely exogenous variation present in the world – birthdate – and showed how this has a small but measurable effect on educational attainment, and thereby on earnings. Their crafting of this natural experiment regression is extremely clever and demonstrates a style of analysis which can successfully underlie an effective instrumental variables empirical analysis.

## 12.42 Programming

We now present Stata code for some of the empirical work reported in this chapter.

**Stata do File for Card Example**

```

use Card1995.dta, clear
set more off
gen exp = age76 - ed76 - 6
gen exp2 = (exp^2)/100
* Drop observations with missing wage
drop if lwage76==.
* Table 12.1 regressions
reg lwage76 ed76 exp exp2 black reg76r smsa76r, r
ivregress 2sls lwage76 exp exp2 black reg76r smsa76r (ed76=nearc4), r
ivregress 2sls lwage76 black reg76r smsa76r (ed76 exp exp2 = nearc4 age76
age2), r perfect
ivregress 2sls lwage76 exp exp2 black reg76r smsa76r (ed76=nearc4a nearc4b),
r
ivregress 2sls lwage76 black reg76r smsa76r (ed76 exp exp2 = nearc4a nearc4b
age76 age2), r perfect
ivregress liml lwage76 exp exp2 black reg76r smsa76r (ed76=nearc4a nearc4b),
r
* Table 12.2 regressions
reg lwage76 exp exp2 black reg76r smsa76r nearc4, r
reg ed76 exp exp2 black reg76r smsa76r nearc4, r
reg ed76 black reg76r smsa76r nearc4 age76 age2, r
reg exp black reg76r smsa76r nearc4 age76 age2, r
reg exp2 black reg76r smsa76r nearc4 age76 age2, r
reg ed76 exp exp2 black reg76r smsa76r nearc4a nearc4b, r
reg lwage76 ed76 exp exp2 smsa76r reg76r, r
reg lwage76 nearc4 exp exp2 smsa76r reg76r, r
reg ed76 nearc4 exp exp2 smsa76r reg76r, r

```

**Stata do File for Acemoglu-Johnson-Robinson Example**

```

use AJR2001.dta, clear
reg loggdp risk
reg risk logmort0
predict u, residual
ivregress 2sls loggdp (risk=logmort0)
reg loggdp risk u

```

**Stata do File for Angrist-Krueger Example**

```
use AK1991.dta, clear
ivregress 2sls logwage black smsa married i.yob i.region (edu = i.qob#i.yob)
ivregress 2sls logwage black smsa married i.yob i.region i.state (edu =
i.qob#i.yob i.qob#i.state)
reg edu black smsa married i.yob i.region i.qob#i.yob
testparm i.qob#i.yob
reg edu black smsa married i.yob i.region i.state i.qob#i.yob i.qob#i.state
testparm i.qob#i.yob i.qob#i.state
reg edu black smsa married i.yob i.region i.qob
testparm i.qob
ivregress 2sls logwage black smsa married i.yob i.region (edu = i.qob)
```

## Exercises

**Exercise 12.1** Consider the single equation model

$$y_i = z_i \beta + e_i,$$

where  $y_i$  and  $z_i$  are both real-valued ( $1 \times 1$ ). Let  $\hat{\beta}$  denote the IV estimator of  $\beta$  using as an instrument a dummy variable  $d_i$  (takes only the values 0 and 1). Find a simple expression for the IV estimator in this context.

**Exercise 12.2** In the linear model

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(e_i | \mathbf{x}_i) &= 0 \end{aligned}$$

suppose  $\sigma^2_i = \mathbb{E}(e_i^2 | x_i)$  is known. Show that the GLS estimator of  $\boldsymbol{\beta}$  can be written as an IV estimator using some instrument  $z_i$ . (Find an expression for  $z_i$ .)

**Exercise 12.3** Take the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}.$$

Let the OLS estimator for  $\boldsymbol{\beta}$  be  $\hat{\beta}$  and the OLS residual be  $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\beta}$ .

Let the IV estimator for  $\boldsymbol{\beta}$  using some instrument  $\mathbf{Z}$  be  $\tilde{\beta}$  and the IV residual be  $\tilde{\mathbf{e}} = \mathbf{y} - \mathbf{X}\tilde{\beta}$ . If  $\mathbf{X}$  is indeed endogenous, will IV “fit” better than OLS, in the sense that  $\tilde{\mathbf{e}}'\tilde{\mathbf{e}} < \hat{\mathbf{e}}'\hat{\mathbf{e}}$ , at least in large samples?

**Exercise 12.4** The reduced form between the regressors  $\mathbf{x}_i$  and instruments  $\mathbf{z}_i$  takes the form

$$\mathbf{x}_i = \boldsymbol{\Gamma}' \mathbf{z}_i + \mathbf{u}_i$$

or

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{U}$$

where  $\mathbf{x}_i$  is  $k \times 1$ ,  $\mathbf{z}_i$  is  $l \times 1$ ,  $\mathbf{X}$  is  $n \times k$ ,  $\mathbf{Z}$  is  $n \times l$ ,  $\mathbf{U}$  is  $n \times k$ , and  $\boldsymbol{\Gamma}$  is  $l \times k$ . The parameter  $\boldsymbol{\Gamma}$  is defined by the population moment condition

$$\mathbb{E}(\mathbf{z}_i \mathbf{u}'_i) = \mathbf{0}.$$

Show that the method of moments estimator for  $\boldsymbol{\Gamma}$  is  $\hat{\boldsymbol{\Gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X})$ .

**Exercise 12.5** In the structural model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{U}$$

with  $\boldsymbol{\Gamma} l \times k$ ,  $l \geq k$ , we claim that  $\boldsymbol{\beta}$  is identified (can be recovered from the reduced form) if  $\text{rank}(\boldsymbol{\Gamma}) = k$ . Explain why this is true. That is, show that if  $\text{rank}(\boldsymbol{\Gamma}) < k$  then  $\boldsymbol{\beta}$  cannot be identified.

**Exercise 12.6** For Theorem 12.3, establish that  $\hat{V}_{\boldsymbol{\beta}} \xrightarrow{P} V_{\boldsymbol{\beta}}$ .

**Exercise 12.7** Take the linear model

$$\begin{aligned} y_i &= x_i \beta + e_i \\ \mathbb{E}(e_i | x_i) &= 0. \end{aligned}$$

where  $x_i$  and  $\beta$  are  $1 \times 1$ .

- (a) Show that  $\mathbb{E}(x_i e_i) = 0$  and  $\mathbb{E}(x_i^2 e_i) = 0$ . Is  $\mathbf{z}_i = (x_i \quad x_i^2)'$  a valid instrumental variable for estimation of  $\beta$ ?

(b) Define the 2SLS estimator of  $\beta$ , using  $z_i$  as an instrument for  $x_i$ . How does this differ from OLS?

**Exercise 12.8** Suppose that price and quantity are determined by the intersection of the linear demand and supply curves

$$\begin{aligned}\text{Demand: } Q &= a_0 + a_1 P + a_2 Y + \epsilon_1 \\ \text{Supply: } Q &= b_0 + b_1 P + b_2 W + \epsilon_2\end{aligned}$$

where income ( $Y$ ) and wage ( $W$ ) are determined outside the market. In this model, are the parameters identified?

**Exercise 12.9** Consider the model

$$\begin{aligned}y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(e_i | \mathbf{z}_i) &= 0\end{aligned}$$

with  $y_i$  scalar and  $\mathbf{x}_i$  and  $\mathbf{z}_i$  each a  $k$  vector. You have a random sample  $(y_i, \mathbf{x}_i, \mathbf{z}_i : i = 1, \dots, n)$ .

- (a) Suppose that  $\mathbf{x}_i$  is exogenous in the sense that  $E(e_i | \mathbf{z}_i, \mathbf{x}_i) = 0$ . Is the IV estimator  $\hat{\boldsymbol{\beta}}_{\text{iv}}$  unbiased for  $\boldsymbol{\beta}$ ?
- (b) Continuing to assume that  $\mathbf{x}_i$  is exogenous, find the variance matrix for  $\hat{\boldsymbol{\beta}}_{\text{iv}}$ ,  $\text{var}(\hat{\boldsymbol{\beta}}_{\text{iv}} | \mathbf{X}, \mathbf{Z})$ .

**Exercise 12.10** Consider the model

$$\begin{aligned}y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbf{x}_i &= \Gamma' \mathbf{z}_i + \mathbf{u}_i \\ \mathbb{E}(\mathbf{z}_i e_i) &= \mathbf{0} \\ \mathbb{E}(\mathbf{z}_i \mathbf{u}'_i) &= \mathbf{0}\end{aligned}$$

with  $y_i$  scalar and  $\mathbf{x}_i$  and  $\mathbf{z}_i$  each a  $k$  vector. You have a random sample  $(y_i, \mathbf{x}_i, \mathbf{z}_i : i = 1, \dots, n)$ . Take the control function equation

$$\begin{aligned}e_i &= \mathbf{u}'_i \boldsymbol{\gamma} + \varepsilon_i \\ \mathbb{E}(\mathbf{u}_i \varepsilon_i) &= \mathbf{0}\end{aligned}$$

and assume for simplicity that  $\mathbf{u}_i$  is observed. Inserting into the structural equation we find

$$y_i = \mathbf{z}'_i \boldsymbol{\beta} + \mathbf{u}'_i \boldsymbol{\gamma} + \varepsilon_i.$$

The control function estimator  $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$  is OLS estimation of this equation.

- (a) Show that  $\mathbb{E}(\mathbf{x}_i \varepsilon_i) = \mathbf{0}$  (algebraically).
- (b) Derive the asymptotic distribution of  $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ .

**Exercise 12.11** Consider the structural equation

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i \tag{12.97}$$

with  $x_i$  treated as endogenous so that  $E(x_i e_i) \neq 0$ . Assume  $y_i$  and  $x_i$  are scalar. Suppose we also have a scalar instrument  $z_i$  which satisfies

$$\mathbb{E}(e_i | z_i) = 0$$

so in particular  $\mathbb{E}(e_i) = 0$ ,  $\mathbb{E}(z_i e_i) = 0$  and  $\mathbb{E}(z_i^2 e_i) = 0$ .

- (a) Should  $x_i^2$  be treated as endogenous or exogenous?
- (b) Suppose we have a scalar instrument  $z_i$  which satisfies

$$x_i = \gamma_0 + \gamma_1 z_i + u_i \quad (12.98)$$

with  $u_i$  independent of  $z_i$  and mean zero.

Consider using  $(1, z_i, z_i^2)$  as instruments. Is this a sufficient number of instruments? (Would this be just-identified, over-identified, or under-identified)?

- (c) Write out the reduced form equation for  $x_i^2$ . Under what condition on the reduced form parameters (12.98) are the parameters in (12.97) identified?

**Exercise 12.12** Consider the structural equation and reduced form

$$\begin{aligned} y_i &= \beta x_i^2 + e_i \\ x_i &= \gamma z_i + u_i \\ \mathbb{E}(z_i e_i) &= 0 \\ \mathbb{E}(z_i u_i) &= 0 \end{aligned}$$

with  $x_i^2$  treated as endogenous so that  $\mathbb{E}(x_i^2 e_i) \neq 0$ . For simplicity assume no intercepts.  $y_i$ ,  $z_i$ , and  $x_i$  are scalar. Assume  $\gamma \neq 0$ . Consider the following estimator. First, estimate  $\gamma$  by OLS of  $x_i$  on  $z_i$  and construct the fitted values  $\hat{x}_i = \hat{\gamma} z_i$ . Second, estimate  $\beta$  by OLS of  $y_i$  on  $\hat{x}_i^2$ .

- (a) Write out this estimator  $\hat{\beta}$  explicitly as a function of the sample.
- (b) Find its probability limit as  $n \rightarrow \infty$
- (c) In general, is  $\hat{\beta}$  consistent for  $\beta$ ? Is there a reasonable condition under which  $\hat{\beta}$  is consistent?

**Exercise 12.13** Consider the structural equation

$$\begin{aligned} y_i &= \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + e_i \\ \mathbb{E}(\mathbf{z}_i e_i) &= 0 \end{aligned}$$

where  $\mathbf{x}_{2i}$  is  $k_2 \times 1$  and treated as endogenous. The variables  $\mathbf{z}_i = (\mathbf{x}_{1i}, \mathbf{x}_{2i})$  are treated as exogenous, where  $\mathbf{z}_{2i}$  is  $\ell_2 \times 1$  and  $\ell_2 \geq k_2$ . You are interested in testing the hypothesis

$$\mathbb{H}_0 : \boldsymbol{\beta}_2 = 0.$$

Consider the reduced form equation for  $y_i$

$$y_i = \mathbf{x}'_{1i} \boldsymbol{\lambda}_1 + \mathbf{z}'_{2i} \boldsymbol{\lambda}_2 + v_i. \quad (12.99)$$

Show how to test  $\mathbb{H}_0$  using only the OLS estimates of (12.99).

Hint: This will require an analysis of the reduced form equations and their relation to the structural equation.

**Exercise 12.14** Take the linear instrumental variables equation

$$\begin{aligned} y_i &= \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + e_i \\ \mathbb{E}(\mathbf{z}_i e_i) &= 0 \end{aligned}$$

where  $\mathbf{x}_{1i}$  is  $k_1 \times 1$ ,  $\mathbf{x}_{2i}$  is  $k_2 \times 1$ , and  $\mathbf{z}_i$  is  $\ell \times 1$ , with  $\ell \geq k = k_1 + k_2$ . The sample size is  $n$ . Assume that  $\mathbf{Q}_{zz} = \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i) > 0$  and  $\mathbf{Q}_{zx} = \mathbb{E}(\mathbf{z}_i \mathbf{x}'_i)$  has full rank  $k$ .

Suppose that only  $(y_i, \mathbf{x}_{1i}, \mathbf{z}_i)$  are available, and  $\mathbf{x}_{2i}$  is missing from the dataset.

Consider the 2SLS estimator  $\hat{\boldsymbol{\beta}}_1$  of  $\boldsymbol{\beta}_1$  obtained from the misspecified IV regression, by regressing  $y_i$  on  $\mathbf{x}_{1i}$  only, using  $\mathbf{z}_i$  as an instrument for  $\mathbf{x}_{1i}$ .

- (a) Find a stochastic decomposition  $\hat{\beta}_1 = \beta_1 + \mathbf{b}_{1n} + \mathbf{r}_{1n}$  where  $\mathbf{r}_{1n}$  depends on the error  $e_i$ , and  $\mathbf{b}_{1n}$  does not depend on the error  $e_i$ .
- (b) Show that  $\mathbf{r}_{1n} \rightarrow_p 0$  as  $n \rightarrow \infty$ .
- (c) Find the probability limit of  $\mathbf{b}_{1n}$  and  $\hat{\beta}_1$  as  $n \rightarrow \infty$ .
- (d) Does  $\hat{\beta}_1$  suffer from “omitted variables bias”? Explain. Under what conditions is there no omitted variables bias?
- (e) Find the asymptotic distribution as  $n \rightarrow \infty$  of

$$\sqrt{n}(\hat{\beta}_1 - \beta_1 - \mathbf{b}_{1n}).$$

**Exercise 12.15** Take the linear instrumental variables equation

$$\begin{aligned} y_i &= x_i \beta_1 + z_i \beta_2 + e_i \\ \mathbb{E}(e_i | z_i) &= 0 \end{aligned}$$

where for simplicity both  $x_i$  and  $z_i$  are scalar  $1 \times 1$ .

- (a) Can the coefficients  $(\beta_1, \beta_2)$  be estimated by 2SLS using  $z_i$  as an instrument for  $x_i$ ? Why or why not?
- (b) Can the coefficients  $(\beta_1, \beta_2)$  be estimated by 2SLS using  $z_i$  and  $z_i^2$  as instruments?
- (c) For the 2SLS estimator suggested in (b), what is the implicit exclusion restriction?
- (d) In (b), what is the implicit assumption about instrument relevance?  
[Hint: Write down the implied reduced form equation for  $x_i$ .]
- (e) In a generic application, would you be comfortable with the assumptions in (c) and (d)?

**Exercise 12.16** Take a linear equation with endogeneity and a just-identified linear reduced form

$$\begin{aligned} y_i &= x_i \beta + e_i \\ x_i &= \gamma z_i + u_i \end{aligned}$$

where both  $x_i$  and  $z_i$  are scalar  $1 \times 1$ . Assume that

$$\begin{aligned} \mathbb{E}(z_i e_i) &= 0 \\ \mathbb{E}(z_i u_i) &= 0. \end{aligned}$$

- (a) Derive the reduced form equation

$$y_i = z_i \lambda + v_i.$$

Show that  $\beta = \lambda/\gamma$  if  $\gamma \neq 0$ , and that  $\mathbb{E}(z_i v_i) = 0$ .

- (b) Let  $\hat{\lambda}$  denote the OLS estimate from linear regression of  $Y$  on  $Z$ , and let  $\hat{\gamma}$  denote the OLS estimate from linear regression of  $X$  on  $Z$ . Write  $\theta = (\lambda, \gamma)'$  and let  $\hat{\theta} = (\hat{\lambda}, \hat{\gamma})'$ . Define the error vector  $\xi_i = \begin{pmatrix} v_i \\ u_i \end{pmatrix}$ . Write  $\sqrt{n}(\hat{\theta} - \theta)$  using a single expression as a function of the error  $\xi_i$ .
- (c) Show that  $\mathbb{E}(z_i \xi_i) = 0$ .
- (d) Derive the joint asymptotic distribution of  $\sqrt{n}(\hat{\theta} - \theta)$  as  $n \rightarrow \infty$ . Hint: Define  $\Omega_\xi = \mathbb{E}(z_i^2 \xi_i \xi_i')$

- (e) Using the previous result and the Delta Method, find the asymptotic distribution of the Indirect Least Squares estimator  $\hat{\beta} = \hat{\lambda}/\hat{\gamma}$ .
- (f) Is the answer in (e) the same as the asymptotic distribution of the 2SLS estimator in Theorem 12.2?

Hint: Show that  $\begin{pmatrix} 1 & -\beta \end{pmatrix} \xi_i = e_i$  and  $\begin{pmatrix} 1 & -\beta \end{pmatrix} \Omega_\xi \begin{pmatrix} 1 \\ -\beta \end{pmatrix} = \mathbb{E}(z_i^2 e_i^2)$ .

**Exercise 12.17** Take the model

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(z_i e_i) &= 0 \end{aligned}$$

and consider the two-stage least-squares estimator. The first-stage estimate is

$$\begin{aligned} \hat{\mathbf{X}} &= \mathbf{Z}\hat{\Gamma} \\ \hat{\Gamma} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} \end{aligned}$$

and the second-stage is least-squares of  $y_i$  on  $\hat{\mathbf{x}}_i$ :

$$\hat{\boldsymbol{\beta}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'y$$

with least-squares residuals

$$\hat{\mathbf{e}} = y - \hat{\mathbf{X}}\hat{\boldsymbol{\beta}}.$$

Consider  $\hat{\sigma}^2 = \frac{1}{n}\hat{\mathbf{e}}'\hat{\mathbf{e}}$  as an estimator for  $\sigma^2 = \mathbb{E}(e_i^2)$ . Is this appropriate? If not, propose an alternative estimator.

**Exercise 12.18** You have two independent iid samples  $(y_{1i}, \mathbf{x}_{1i}, \mathbf{z}_{1i} : i = 1, \dots, n)$  and  $(y_{2i}, \mathbf{x}_{2i}, \mathbf{z}_{2i} : i = 1, \dots, n)$ . The dependent variables  $y_{1i}$  and  $y_{2i}$  are real-valued. The regressors  $\mathbf{x}_{1i}$  and  $\mathbf{x}_{2i}$  and instruments  $\mathbf{z}_{1i}$  and  $\mathbf{z}_{2i}$  are  $k$ -vectors. The model is standard just-identified linear instrumental variables

$$\begin{aligned} y_{1i} &= \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + e_{1i} \\ \mathbb{E}(\mathbf{z}_{1i} e_{1i}) &= \mathbf{0} \\ y_{2i} &= \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + e_{2i} \\ \mathbb{E}(\mathbf{z}_{2i} e_{2i}) &= \mathbf{0}. \end{aligned}$$

For concreteness, sample 1 are women and sample 2 are men. You want to test  $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ , that the two samples have the same coefficients.

- (a) Develop a test statistic for  $H_0$ .
- (b) Derive the asymptotic distribution of the test statistic.
- (c) Describe (in brief) the testing procedure.

**Exercise 12.19** To estimate  $\beta$  in the model  $y_i = x_i\beta + e_i$  with  $x_i$  scalar and endogenous, with household level data, you want to use as an instrument the state of residence.

- (a) What are the assumptions needed to justify this choice of instrument?
- (b) Is the model just identified or overidentified?

**Exercise 12.20** The model is

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + e_i$$

$$\mathbb{E}(\mathbf{z}_i e_i) = \mathbf{0}.$$

An economist wants to obtain the 2SLS estimates and standard errors for  $\boldsymbol{\beta}$ . He uses the following steps

- Regresses  $\mathbf{x}_i$  on  $\mathbf{z}_i$ , obtains the predicted values  $\hat{\mathbf{x}}_i$ .
- Regresses  $y_i$  on  $\hat{\mathbf{x}}_i$ , obtains the coefficient estimate  $\hat{\boldsymbol{\beta}}$  and standard error  $s(\hat{\boldsymbol{\beta}})$  from this regression.

Is this correct? Does this produce the 2SLS estimates and standard errors?

**Exercise 12.21** Let

$$y_i = \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + e_i.$$

Let  $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$  denote the 2SLS estimates of  $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$  when  $\mathbf{z}_{2i}$  is used as an instrument for  $\mathbf{x}_{2i}$  and they are the same dimension (so the model is just identified). Let  $(\hat{\boldsymbol{\lambda}}_1, \hat{\boldsymbol{\lambda}}_2)$  be the OLS estimates from the regression

$$y_i = \mathbf{x}'_{1i} \hat{\boldsymbol{\lambda}}_1 + \mathbf{z}'_{2i} \hat{\boldsymbol{\lambda}}_2 + e_i.$$

Show that  $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\lambda}}_1$ .

**Exercise 12.22** In the linear model

$$y_i = x_i \beta + e_i$$

suppose  $\sigma^2_i = E(e_i^2 | x_i)$  is known. Show that the GLS estimator of  $\beta$  can be written as an instrumental variables estimator using some instrument  $z_i$ . (Find an expression for  $z_i$ .)

**Exercise 12.23** You will replicate and extend the work reported in Acemoglu, Johnson and Robinson (2001). The authors provided an expanded set of controls when they published their 2012 extension and posted the data on the AER website. This dataset is AJR2001 on the textbook website.

- Estimate the OLS regression (12.90), the reduced form regression (12.91) and the 2SLS regression (12.92). (Which point estimate is different by 0.01 from the reported values? This is a common phenomenon in empirical replication).
- For the above estimates, calculate both homoskedastic and heteroskedastic-robust standard errors. Which were used by the authors (as reported in (12.90)-(12.91)-(12.92)?)
- Calculate the 2SLS estimates by the Indirect Least Squares formula. Are they the same?
- Calculate the 2SLS estimates by the two-stage approach. Are they the same?
- Calculate the 2SLS estimates by the control variable approach. Are they the same?
- Acemoglu, Johnson and Robinson (2001) reported many specifications including alternative regressor controls, for example *latitude* and *africa*. Estimate by least-squares the equation for log-GDP adding *latitude* and *africa* as regressors. Does this regression suggest that *latitude* and *africa* are predictive of the level of GDP?
- Now estimate the same equation as in (f) but by 2SLS using log mortality as an instrument for *risk*. How does the interpretation of the effect of *latitude* and *africa* change?
- Return to our baseline model (without including *latitude* and *africa*). The authors' reduced form equation uses log(mortality) as the instrument, rather than, say, the level of mortality. Estimate the reduced form for risk with *mortality* as the instrument. (This variable is not provided in the dataset, so you need to take the exponential of the mortality variable.) Can you explain why the authors preferred the equation with log(mortality)?

- (i) Try an alternative reduced form, including both log(mortality) and the square of log(mortality). Interpret the results. Re-estimate the structural equation by 2SLS using both log(mortality) and its square as instruments. How do the results change?
- (j) For the estimates in (i), are the instruments strong or weak using the Stock-Yogo test?
- (k) Calculate and interpret a test for exogeneity of the instruments.
- (l) Estimate the equation by LIML, using the instruments log(mortality) and the square of log(mortality).

**Exercise 12.24** In Exercise 12.23 you extended the reported in Acemoglu, Johnson and Robinson (2001). Consider the 2SLS regression (12.92). Compute the standard errors both by the asymptotic formula and by the bootstrap using a large number (10,000) of bootstrap replications. Re-calculate the bootstrap standard errors. Comment on the reliability of bootstrap standard errors for IV regression.

**Exercise 12.25** You will replicate and extend the work reported in the chapter relating to Card (1995). The data is from the author's website, and is posted as *Card1995*. The model we focus on is labeled 2SLS(a) in Table 12.1, which uses *public* and *private* as instruments for *Edu*. The variables you will need for this exercise include *lwage76*, *ed76*, *age76*, *smsa76r*, *reg76r*, *nearc2*, *nearc4*, *nearc4a*, *nearc4b*. See the description file for definitions.

$$\log(Wage) = \beta_0 + \beta_1 Edu + \beta_2 Exp + \beta_3 Exp^2/100 + \beta_4 South + \beta_5 Black + \epsilon$$

where *Edu* = *Education* (Years), *Exp* = *Experience* (Years), and *South* and *Black* are regional and racial dummy variables. The variables *Exp* = *Age* − *Edu* − 6 and *Exp*<sup>2</sup>/100 are not in the dataset, they need to be generated.

- (a) First, replicate the reduced form regression presented in the final column of Table 12.2, and the 2SLS regression described above (using *public* and *private* as instruments for *Edu*) to verify that you have the same variable definitions.
- (b) Now try a different reduced form model. The variable *nearc2* means "grew up near a 2-year college". See if adding it to the reduced form equation is useful.
- (c) Now try more interactions in the reduced form. Create the interactions *nearc4a\*age76* and *nearc4a\*age76<sup>2</sup>/100*, and add them to the reduced form equation. Estimate this by least-squares. Interpret the coefficients on the two new variables.
- (d) Estimate the structural equation by 2SLS using the expanded instrument set  $\{nearc4a, nearc4b, nearc4a*age76, nearc4a*age76^2/100\}$ .  
What is the impact on the structural estimate of the return to schooling?
- (e) Using the Stock-Yogo test, are the instruments strong or weak?
- (f) Test the hypothesis that *Edu* is exogenous for the structural return to schooling.
- (g) Re-estimate the last equation by LIML. Do the results change meaningfully?

**Exercise 12.26** In Exercise 12.25 you extended the work reported in Card (1995). Now, estimate the IV equation corresponding to the IV(a) column of Table 12.1, which is the baseline specification considered in Card. Use the bootstrap to calculate a BC percentile confidence interval. In this example, should we also report the bootstrap standard error?

**Exercise 12.27** You will extend Angrist and Krueger (1991). In their Table VIII, they report their estimates of an analog of (12.94) for the subsample of 26,913 black men. Use this sub-sample for the following analysis.

- (a) Start by considering estimation of an equation which is identical in form to (12.94), with the same additional regressors (year-of-birth, region-of-residence, and state-of-birth dummy variables) and 180 excluded instrumental variables (the interactions of quarter-of-birth times year-of-birth dummy variables, and quarter-of-birth times state-of-birth interactions). But now, it is estimated on the subsample of black men. One regressor must be omitted to achieve identification. Which variable is this?
- (b) Estimate the reduced form for the above equation by least-squares. Calculate the  $F$  statistic for the excluded instruments. What do you conclude about the strength of the instruments?
- (c) Repeat, now estimating the reduced form for the analog of (12.93) which has 30 excluded instrumental variables, and does not include the state-of-birth dummy variables in the regression. What do you conclude about the strength of the instruments?
- (d) Repeat, now estimating the reduced form for the analog of (12.96) which has only 3 excluded instrumental variables. Are the instruments sufficiently strong for 2SLS estimation? For LIML estimation?
- (e) Estimate the structural wage equation using what you believe is the most appropriate set of regressors, instruments, and the most appropriate estimation method. What is the estimated return to education (for the subsample of black men) and its standard error? Without doing a formal hypothesis test, do these results (or in which way?) appear meaningfully different from the results for the full sample?

**Exercise 12.28** In Exercise 12.27 you extended the work reported in Angrist and Krueger (1991) by estimating wage equations for the subsample of black men. Re-estimate equation (12.96) for this group, which uses as instruments only the three quarter-of-birth dummy variables. Calculate the standard error for the return to education by asymptotic and bootstrap methods, and a BC percentile interval. In this application of 2SLS, is it appropriate to report a bootstrap standard error?

# Chapter 13

## Generalized Method of Moments

### 13.1 Introduction

One of the most popular estimation methods in applied econometrics is the Generalized Method of Moments (GMM). GMM generalizes the classical method of moments estimator by allowing for models that have more equations than unknown parameters and are thus overidentified. GMM includes as special cases OLS, IV, multivariate regression, and 2SLS. It includes both linear and nonlinear models. In this chapter we focus primarily on linear models.

The GMM label and methods were introduced to econometrics in a seminal paper by Lars Hansen (1982). The ideas and methods build on the work of Amemiya (1974, 1977), Gallant (1977), and Gallant and Jorgenson (1979). The ideas are closely related to the contemporaneous work of Halbert White (1980, 1982) and White and Domowitz (1984). The methods are also related to what are called **estimating equations** in the statistics literature. For a review of the later see Godambe (1991).

### 13.2 Moment Equation Models

All of the models that have been introduced so far can be written as **moment equation models**, where the population parameters solve a system of moment equations. Moment equation models are much broader than the models so far considered, and understanding their common structure opens up straightforward techniques to handle new econometric models.

Moment equation models take the following form. Let  $\mathbf{g}_i(\boldsymbol{\beta})$  be a known  $\ell \times 1$  function of the  $i^{th}$  observation and a  $k \times 1$  parameter  $\boldsymbol{\beta}$ . A moment equation model is summarized by the moment equations

$$\mathbb{E}(\mathbf{g}_i(\boldsymbol{\beta})) = \mathbf{0} \quad (13.1)$$

and a parameter space  $\boldsymbol{\beta} \in \mathcal{B}$ . For example, in the instrumental variables model  $\mathbf{g}_i(\boldsymbol{\beta}) = \mathbf{z}_i(y_i - \mathbf{x}'_i \boldsymbol{\beta})$ .

In general, we say that a parameter  $\boldsymbol{\beta}$  is **identified** if there is a unique mapping from the data distribution to  $\boldsymbol{\beta}$ . In the context of the model (13.1) this means that there is a unique  $\boldsymbol{\beta}$  satisfying (13.1). Since (13.1) is a system of  $\ell$  equations with  $k$  unknowns, then it is necessary that  $\ell \geq k$  for there to be a unique solution. If  $\ell = k$  we say that the model is **just identified**, meaning that there is just enough information to identify the parameters. If  $\ell > k$  we say that the model is **overidentified**, meaning that there is excess information (which can improve estimation efficiency). If  $\ell < k$  we say that the model is **underidentified**, meaning that there is insufficient information to identify the parameters. In general, we assume that  $\ell \geq k$  so the model is either just identified or overidentified.

### 13.3 Method of Moments Estimators

In this section we consider the just-identified case  $\ell = k$ .

Define the sample analog of (13.5)

$$\bar{\mathbf{g}}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{g}_i(\boldsymbol{\beta}). \quad (13.2)$$

The **method of moments estimator (MME)**  $\hat{\boldsymbol{\beta}}_{\text{mm}}$  for  $\boldsymbol{\beta}$  is defined as the parameter value which sets  $\bar{\mathbf{g}}_n(\boldsymbol{\beta}) = \mathbf{0}$ . Thus

$$\bar{\mathbf{g}}_n(\hat{\boldsymbol{\beta}}_{\text{mm}}) = \frac{1}{n} \sum_{i=1}^n \mathbf{g}_i(\hat{\boldsymbol{\beta}}_{\text{mm}}) = \mathbf{0}. \quad (13.3)$$

The equations (13.3) are known as the **estimating equations** as they are the equations which determine the estimator  $\hat{\boldsymbol{\beta}}_{\text{mm}}$ .

In some contexts (such as those discussed in the examples below), there is an explicit solution for  $\hat{\boldsymbol{\beta}}_{\text{mm}}$ . In other cases the solution must be found numerically.

We now show how most of the estimators discussed so far in the textbook can be written as method of moments estimators.

**Mean:** Set  $g_i(\mu) = y_i - \mu$ . The MME is  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i$ .

**Mean and Variance:** Set

$$\mathbf{g}_i(\mu, \sigma^2) = \begin{pmatrix} y_i - \mu \\ (y_i - \mu)^2 - \sigma^2 \end{pmatrix}.$$

The MME are  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i$  and  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\mu})^2$ .

**OLS:** Set  $\mathbf{g}_i(\boldsymbol{\beta}) = \mathbf{x}_i(y_i - \mathbf{x}'_i \boldsymbol{\beta})$ . The MME is  $\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{y})$ .

**OLS and Variance:** Set

$$\mathbf{g}_i(\boldsymbol{\beta}, \sigma^2) = \begin{pmatrix} \mathbf{x}_i(y_i - \mathbf{x}'_i \boldsymbol{\beta}) \\ (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 - \sigma^2 \end{pmatrix}.$$

The MME is  $\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{y})$  and  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}})^2$ .

**Multivariate Least Squares, vector form:** Set  $\mathbf{g}_i(\boldsymbol{\beta}) = \mathbf{X}_i(\mathbf{y}_i - \mathbf{X}'_i \boldsymbol{\beta})$ . The MME is  $\hat{\boldsymbol{\beta}} = (\sum_{i=1}^n \mathbf{X}_i \mathbf{X}'_i)^{-1} (\sum_{i=1}^n \mathbf{X}_i \mathbf{y}_i)$  which is (11.4).

**Multivariate Least Squares, matrix form:** Set  $\mathbf{g}_i(\mathbf{B}) = \text{vec}(\mathbf{x}_i(\mathbf{y}'_i - \mathbf{x}'_i \mathbf{B}))$ . The MME is  $\hat{\mathbf{B}} = (\sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i)^{-1} (\sum_{i=1}^n \mathbf{x}_i \mathbf{y}'_i)$  which is (11.6).

**Seemingly Unrelated Regression:** Set

$$\mathbf{g}_i(\boldsymbol{\beta}, \Sigma) = \begin{pmatrix} \mathbf{X}_i \Sigma^{-1} (\mathbf{y}_i - \mathbf{X}'_i \boldsymbol{\beta}) \\ \text{vec}(\Sigma - (\mathbf{y}_i - \mathbf{X}'_i \boldsymbol{\beta})(\mathbf{y}_i - \mathbf{X}'_i \boldsymbol{\beta})') \end{pmatrix}.$$

The MME is  $\hat{\boldsymbol{\beta}} = (\sum_{i=1}^n \mathbf{X}_i \hat{\Sigma}^{-1} \mathbf{X}'_i)^{-1} (\sum_{i=1}^n \mathbf{X}_i \hat{\Sigma}^{-1} \mathbf{y}_i)$  and  $\hat{\Sigma} = n^{-1} \sum_{i=1}^n (\mathbf{y}_i - \mathbf{X}'_i \hat{\boldsymbol{\beta}})(\mathbf{y}_i - \mathbf{X}'_i \hat{\boldsymbol{\beta}})'$ .

**IV:** Set  $\mathbf{g}_i(\boldsymbol{\beta}) = \mathbf{z}_i(y_i - \mathbf{x}'_i \boldsymbol{\beta})$ . The MME is  $\hat{\boldsymbol{\beta}} = (\sum_{i=1}^n \mathbf{z}_i \mathbf{x}'_i)^{-1} (\sum_{i=1}^n \mathbf{z}_i y_i)$ .

**Generated Regressors:** Set

$$\mathbf{g}_i(\boldsymbol{\beta}, \mathbf{A}) = \begin{pmatrix} \mathbf{A}' \mathbf{z}_i(y_i - \mathbf{z}'_i \mathbf{A} \boldsymbol{\beta}) \\ \text{vec}(\mathbf{z}_i(\mathbf{x}'_i - \mathbf{z}'_i \mathbf{A})) \end{pmatrix}.$$

The MME is  $\hat{\mathbf{A}} = (\sum_{i=1}^n \mathbf{z}_i \mathbf{z}'_i)^{-1} (\sum_{i=1}^n \mathbf{z}_i \mathbf{x}'_i)$  and  $\hat{\boldsymbol{\beta}} = (\hat{\mathbf{A}}' \sum_{i=1}^n \mathbf{z}_i \mathbf{z}'_i \hat{\mathbf{A}})^{-1} (\hat{\mathbf{A}}' \sum_{i=1}^n \mathbf{z}_i y_i)$ .

A common feature unifying these examples is that the estimator can be written as the solution to a set of estimating equations (13.3). This provides a common framework which enables a convenient development of a unified distribution theory.

## 13.4 Overidentified Moment Equations

In the instrumental variables model  $g_i(\boldsymbol{\beta}) = \mathbf{z}_i(y_i - \mathbf{x}'_i \boldsymbol{\beta})$ . Thus (13.2) is

$$\bar{\mathbf{g}}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{g}_i(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i(y_i - \mathbf{x}'_i \boldsymbol{\beta}) = \frac{1}{n} (\mathbf{Z}' \mathbf{y} - \mathbf{Z}' \mathbf{X} \boldsymbol{\beta}). \quad (13.4)$$

We have defined the method of moments estimator for  $\boldsymbol{\beta}$  as the parameter value which sets  $\bar{\mathbf{g}}_n(\boldsymbol{\beta}) = \mathbf{0}$ . However, when the model is overidentified ( $\ell > k$ ) then this is generally impossible as there are more equations than free parameters. Equivalently, there is no choice of  $\boldsymbol{\beta}$  which sets (13.4) to zero. Thus the method of moments estimator is not defined for the overidentified case.

While we cannot find an estimator which sets  $\bar{\mathbf{g}}_n(\boldsymbol{\beta})$  equal to zero, we can try to find an estimator which makes  $\bar{\mathbf{g}}_n(\boldsymbol{\beta})$  as close to zero as possible.

One way to think about this is to define the vector  $\boldsymbol{\mu} = \mathbf{Z}' \mathbf{y}$ , the matrix  $\mathbf{G} = \mathbf{Z}' \mathbf{X}$  and the “error”  $\boldsymbol{\eta} = \boldsymbol{\mu} - \mathbf{G} \boldsymbol{\beta}$ . Then we can write (13.4) as

$$\boldsymbol{\mu} = \mathbf{G} \boldsymbol{\beta} + \boldsymbol{\eta}.$$

This looks like a regression equation with the  $\ell \times 1$  dependent variable  $\boldsymbol{\mu}$ , the  $\ell \times k$  regressor matrix  $\mathbf{G}$ , and the  $\ell \times 1$  error vector  $\boldsymbol{\eta}$ . Recall, the goal is to make the error vector  $\boldsymbol{\eta}$  as small as possible. Recalling our knowledge about least-squares, we know that a simple method is to use least-squares regression of  $\boldsymbol{\mu}$  on  $\mathbf{G}$ , which minimizes the sum-of-squares  $\boldsymbol{\eta}' \boldsymbol{\eta}$ . This is certainly one way to make  $\boldsymbol{\eta}$  “small”. This least-squares solution is  $\hat{\boldsymbol{\beta}} = (\mathbf{G}' \mathbf{G})^{-1} (\mathbf{G}' \boldsymbol{\mu})$ .

More generally, we know that when errors are non-homogeneous it can be more efficient to estimate by weighted least squares. Thus for some weight matrix  $\mathbf{W}$ , consider the estimator

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{G}' \mathbf{W} \mathbf{G})^{-1} (\mathbf{G}' \mathbf{W} \boldsymbol{\mu}) \\ &= (\mathbf{X}' \mathbf{Z} \mathbf{W} \mathbf{Z}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{Z} \mathbf{W} \mathbf{Z}' \mathbf{y}).\end{aligned}$$

This minimizes the weighted sum of squares  $\boldsymbol{\eta}' \mathbf{W} \boldsymbol{\eta}$ . This solution is known as the generalized method of moments (GMM).

The estimator is typically defined as follows. Given a set of moment equations (13.2) and an  $\ell \times \ell$  weight matrix  $\mathbf{W} > 0$ , the GMM criterion function is defined as

$$J(\boldsymbol{\beta}) = n \cdot \bar{\mathbf{g}}_n(\boldsymbol{\beta})' \mathbf{W} \bar{\mathbf{g}}_n(\boldsymbol{\beta}).$$

The factor “ $n$ ” is not important for the definition of the estimator, but is convenient for the distribution theory. The criterion  $J(\boldsymbol{\beta})$  is the weighted sum of squared moment equation errors. When  $\mathbf{W} = \mathbf{I}_\ell$ , then  $J(\boldsymbol{\beta}) = n \cdot \bar{\mathbf{g}}_n(\boldsymbol{\beta})' \bar{\mathbf{g}}_n(\boldsymbol{\beta}) = n \cdot \|\bar{\mathbf{g}}_n(\boldsymbol{\beta})\|^2$ , the square of the Euclidean length. Since we restrict attention to positive definite weight matrices  $\mathbf{W}$ , the criterion  $J(\boldsymbol{\beta})$  is always non-negative.

The **Generalized Method of Moments (GMM)** estimator is defined as the minimizer of the GMM criterion  $J(\boldsymbol{\beta})$ .

**Definition 13.1** The Generalized Method of Moments estimator is

$$\hat{\boldsymbol{\beta}}_{\text{gmm}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} J_n(\boldsymbol{\beta}).$$

Recall that in the just-identified case  $k = \ell$ , the method of moments estimator  $\hat{\boldsymbol{\beta}}_{\text{mm}}$  solves  $\bar{\mathbf{g}}_n(\hat{\boldsymbol{\beta}}_{\text{mm}}) = \mathbf{0}$ . Hence in this case  $J_n(\hat{\boldsymbol{\beta}}_{\text{mm}}) = 0$  which means that  $\hat{\boldsymbol{\beta}}_{\text{mm}}$  minimizes  $J_n(\boldsymbol{\beta})$  and equals  $\hat{\boldsymbol{\beta}}_{\text{gmm}} = \hat{\boldsymbol{\beta}}_{\text{mm}}$ . This means that GMM includes MME as a special case. This implies that all of our results for GMM will apply to any method of moments estimators.

In the over-identified case the GMM estimator will depend on the choice of weight matrix  $\mathbf{W}$  and so this is an important focus of the theory. In the just-identified case, the GMM estimator simplifies to the MME which does not depend on  $\mathbf{W}$ .

The method and theory of the generalized method of moments was developed in an influential paper by Lars Hansen (1982). This paper introduced the method, its asymptotic distribution, the form of the efficient weight matrix, and tests for overidentification.

## 13.5 Linear Moment Models

One of the great advantages of the moment equation framework is that it allows both linear and non-linear models. However, when the moment equations are linear in the parameters then we have explicit solutions for the estimates and a straightforward asymptotic distribution theory. Hence we start by confining attention to linear moment equations, and return to nonlinear moment equations later. In the examples listed earlier, the estimators which have linear moment equations include the sample mean, OLS, multivariate least squares, IV, and 2SLS. The estimates which have non-linear moment equations include the sample variance, SUR, and generated regressors.

In particular, we focus on the overidentified IV model

$$\mathbf{g}_i(\boldsymbol{\beta}) = \mathbf{z}_i(y_i - \mathbf{x}'_i \boldsymbol{\beta}) \quad (13.5)$$

where  $\mathbf{z}_i$  is  $\ell \times 1$  and  $\mathbf{x}_i$  is  $k \times 1$ .

## 13.6 GMM Estimator

Given (13.5) the sample moment equations are (13.4). The GMM criterion can be written as

$$J(\boldsymbol{\beta}) = n(\mathbf{Z}'\mathbf{y} - \mathbf{Z}'\mathbf{X}\boldsymbol{\beta})' \mathbf{W} (\mathbf{Z}'\mathbf{y} - \mathbf{Z}'\mathbf{X}\boldsymbol{\beta}).$$

The GMM estimator minimizes  $J(\boldsymbol{\beta})$ . The first order conditions are

$$\begin{aligned} \mathbf{0} &= \frac{\partial}{\partial \boldsymbol{\beta}} J(\hat{\boldsymbol{\beta}}) \\ &= 2 \frac{\partial}{\partial \boldsymbol{\beta}} \bar{\mathbf{g}}_n(\hat{\boldsymbol{\beta}})' \mathbf{W} \bar{\mathbf{g}}_n(\hat{\boldsymbol{\beta}}) \\ &= -2 \left( \frac{1}{n} \mathbf{X}' \mathbf{Z} \right) \mathbf{W} \left( \frac{1}{n} \mathbf{Z}' (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \right). \end{aligned}$$

The solution is given as follows.

**Theorem 13.1** For the overidentified IV model

$$\hat{\boldsymbol{\beta}}_{\text{gmm}} = (\mathbf{X}' \mathbf{Z} \mathbf{W} \mathbf{Z}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{Z} \mathbf{W} \mathbf{Z}' \mathbf{y}). \quad (13.6)$$

While the estimator depends on  $\mathbf{W}$ , the dependence is only up to scale. This is because if  $\mathbf{W}$  is replaced by  $c\mathbf{W}$  for some  $c > 0$ ,  $\hat{\boldsymbol{\beta}}_{\text{gmm}}$  does not change.

When  $\mathbf{W}$  is fixed by the user, we call  $\hat{\boldsymbol{\beta}}_{\text{gmm}}$  a **one-step GMM** estimator.

The GMM estimator (13.6) resembles the 2SLS estimator (12.31). In fact they are equal when  $\mathbf{W} = (\mathbf{Z}' \mathbf{Z})^{-1}$ . This means that the 2SLS estimator is a one-step GMM estimator for the linear model. In the just-identified case it also simplifies to the IV estimator (12.26).

**Theorem 13.2** If  $\mathbf{W} = (\mathbf{Z}' \mathbf{Z})^{-1}$  then  $\hat{\boldsymbol{\beta}}_{\text{gmm}} = \hat{\boldsymbol{\beta}}_{\text{2sls}}$ . Furthermore, if  $k = \ell$  then  $\hat{\boldsymbol{\beta}}_{\text{gmm}} = \hat{\boldsymbol{\beta}}_{\text{iv}}$ .

## 13.7 Distribution of GMM Estimator

Let

$$\mathbf{Q} = \mathbb{E}(\mathbf{z}_i \mathbf{x}'_i)$$

and

$$\boldsymbol{\Omega} = \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i e_i^2) = \mathbb{E}(\mathbf{g}_i \mathbf{g}'_i)$$

where  $\mathbf{g}_i = \mathbf{z}_i e_i$ . Then

$$\left( \frac{1}{n} \mathbf{X}' \mathbf{Z} \right) \mathbf{W} \left( \frac{1}{n} \mathbf{Z}' \mathbf{X} \right) \xrightarrow{p} \mathbf{Q}' \mathbf{W} \mathbf{Q}$$

and

$$\left( \frac{1}{n} \mathbf{X}' \mathbf{Z} \right) \mathbf{W} \left( \frac{1}{\sqrt{n}} \mathbf{Z}' \mathbf{e} \right) \xrightarrow{d} \mathbf{Q}' \mathbf{W} \cdot \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}).$$

We conclude:

**Theorem 13.3 Asymptotic Distribution of GMM Estimator.** Under Assumption 12.2, as  $n \rightarrow \infty$

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}})$$

where

$$\mathbf{V}_{\boldsymbol{\beta}} = (\mathbf{Q}' \mathbf{W} \mathbf{Q})^{-1} (\mathbf{Q}' \mathbf{W} \boldsymbol{\Omega} \mathbf{W} \mathbf{Q}) (\mathbf{Q}' \mathbf{W} \mathbf{Q})^{-1}. \quad (13.7)$$

We find that the GMM estimator is asymptotically normal with a “sandwich form” asymptotic variance.

Our derivation treated the weight matrix  $\mathbf{W}$  as if it is non-random, but Theorem 13.3 carries over to the case where the weight matrix  $\widehat{\mathbf{W}}$  is random so long as it converges in probability to some positive definite limit  $\mathbf{W}$ . This may require scaling the weight matrix, for example replacing  $\widehat{\mathbf{W}} = (\mathbf{Z}' \mathbf{Z})^{-1}$  with  $\widehat{\mathbf{W}} = (n^{-1} \mathbf{Z}' \mathbf{Z})^{-1}$ . Since rescaling the weight matrix does not affect the estimator this is ignored in implementation.

## 13.8 Efficient GMM

The asymptotic distribution of the GMM estimator  $\hat{\boldsymbol{\beta}}_{\text{gmm}}$  depends on the weight matrix  $\mathbf{W}$  through the asymptotic variance  $\mathbf{V}_{\boldsymbol{\beta}}$ . The asymptotically optimal weight matrix  $\mathbf{W}_0$  is one which minimizes  $\mathbf{V}_{\boldsymbol{\beta}}$ . This turns out to be  $\mathbf{W}_0 = \boldsymbol{\Omega}^{-1}$ . The proof is left to Exercise 13.4.

When the GMM estimator  $\hat{\boldsymbol{\beta}}$  is constructed with  $\mathbf{W} = \mathbf{W}_0 = \boldsymbol{\Omega}^{-1}$  (or a weight matrix which is a consistent estimator of  $\mathbf{W}_0$ ) we call it the **Efficient GMM** estimator:

$$\hat{\boldsymbol{\beta}}_{\text{gmm}} = (\mathbf{X}' \mathbf{Z} \boldsymbol{\Omega}^{-1} \mathbf{Z}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{Z} \boldsymbol{\Omega}^{-1} \mathbf{Z}' \mathbf{y}).$$

Its asymptotic distribution takes a simpler form than in Theorem 13.3. By substituting  $\mathbf{W} = \mathbf{W}_0 = \boldsymbol{\Omega}^{-1}$  into (13.7) we find

$$\mathbf{V}_{\boldsymbol{\beta}} = (\mathbf{Q}' \boldsymbol{\Omega}^{-1} \mathbf{Q})^{-1} (\mathbf{Q}' \boldsymbol{\Omega}^{-1} \boldsymbol{\Omega} \boldsymbol{\Omega}^{-1} \mathbf{Q}) (\mathbf{Q}' \boldsymbol{\Omega}^{-1} \mathbf{Q})^{-1} = (\mathbf{Q}' \boldsymbol{\Omega}^{-1} \mathbf{Q})^{-1}.$$

This is the asymptotic variance of the efficient GMM estimator.

**Theorem 13.4 Asymptotic Distribution of GMM with Efficient Weight Matrix.** Under Assumption 12.2 and  $\mathbf{W} = \boldsymbol{\Omega}^{-1}$ , as  $n \rightarrow \infty$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, V_{\boldsymbol{\beta}})$$

where

$$V_{\boldsymbol{\beta}} = (\mathbf{Q}' \boldsymbol{\Omega}^{-1} \mathbf{Q})^{-1}.$$

**Theorem 13.5 Efficient GMM.** Under Assumption 12.2, for any  $\mathbf{W} > 0$ ,

$$(\mathbf{Q}' \mathbf{W} \mathbf{Q})^{-1} (\mathbf{Q}' \mathbf{W} \boldsymbol{\Omega} \mathbf{W} \mathbf{Q}) (\mathbf{Q}' \mathbf{W} \mathbf{Q})^{-1} - (\mathbf{Q}' \boldsymbol{\Omega}^{-1} \mathbf{Q})^{-1} > 0.$$

Thus if  $\hat{\boldsymbol{\beta}}_{\text{gmm}}$  is the efficient GMM estimator and  $\tilde{\boldsymbol{\beta}}_{\text{gmm}}$  is another GMM estimator, then

$$\text{avar}(\hat{\boldsymbol{\beta}}_{\text{gmm}}) \leq \text{avar}(\tilde{\boldsymbol{\beta}}_{\text{gmm}}).$$

For a proof, see Exercise 13.4.

This means that the smallest possible GMM covariance matrix (in the positive definite sense) is achieved by the efficient GMM weight matrix.

$\mathbf{W}_0 = \boldsymbol{\Omega}^{-1}$  is not known in practice but it can be estimated consistently as we discuss in Section 13.10. For any  $\widehat{\mathbf{W}} \xrightarrow{p} \mathbf{W}_0$ , the asymptotic distribution in Theorem 13.4 is unaffected. Consequently we still call any  $\hat{\boldsymbol{\beta}}_{\text{gmm}}$  constructed with an estimate of the efficient weight matrix an efficient GMM estimator.

By “efficient”, we mean that this estimator has the smallest asymptotic variance in the class of GMM estimators with this set of moment conditions. This is a weak concept of optimality, as we are only considering alternative weight matrices  $\widehat{\mathbf{W}}$ . However, it turns out that the GMM estimator is semiparametrically efficient as shown by Gary Chamberlain (1987). If it is known that  $\mathbb{E}(\mathbf{g}(\mathbf{w}_i, \boldsymbol{\beta})) = \mathbf{0}$ , and this is all that is known, this is a semi-parametric problem as the distribution of the data is unknown. Chamberlain showed that in this context no semiparametric estimator (one which is consistent globally for the class of models considered) can have a smaller asymptotic variance than  $(\mathbf{G}' \boldsymbol{\Omega}^{-1} \mathbf{G})^{-1}$  where  $\mathbf{G} = \mathbb{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}'} \mathbf{g}_i(\boldsymbol{\beta})\right)$ . Since the GMM estimator has this asymptotic variance, it is semiparametrically efficient.

The results in this section show that in the linear model no estimator has better asymptotic efficiency than the efficient linear GMM estimator. No estimator can do better (in this first-order asymptotic sense), without imposing additional assumptions.

### 13.9 Efficient GMM versus 2SLS

For the linear model we introduced the 2SLS estimator as a standard estimator for  $\boldsymbol{\beta}$ . Now we have introduced the GMM estimator which includes 2SLS as a special case. Is there a context where 2SLS is efficient?

To answer this question, recall that the 2SLS estimator is GMM given the weight matrix  $\widehat{\mathbf{W}} = (\mathbf{Z}' \mathbf{Z})^{-1}$  or equivalently  $\widehat{\mathbf{W}} = (n^{-1} \mathbf{Z}' \mathbf{Z})^{-1}$  since scaling doesn't matter. Since  $\widehat{\mathbf{W}} \xrightarrow{p} (\mathbb{E}(\mathbf{z}_i \mathbf{z}_i'))^{-1}$ , this is asymptotically equivalent to using the weight matrix  $\mathbf{W} = (\mathbb{E}(\mathbf{z}_i \mathbf{z}_i'))^{-1}$ . In contrast, the efficient weight matrix takes the form  $(\mathbb{E}(\mathbf{z}_i \mathbf{z}_i' e_i^2))^{-1}$ . Now suppose that the structural equation error  $e_i$  is conditionally homoskedastic in the sense that  $\mathbb{E}(e_i^2 | \mathbf{z}_i) = \sigma^2$ . Then the efficient weight matrix equals  $\mathbf{W} = (\mathbb{E}(\mathbf{z}_i \mathbf{z}_i'))^{-1} \sigma^{-2}$ , or equivalently  $\mathbf{W} = (\mathbb{E}(\mathbf{z}_i \mathbf{z}_i'))^{-1}$  since scaling doesn't matter. The latter weight matrix is the same as the

2SLS asymptotic weight matrix. This shows that the 2SLS weight matrix is the efficient weight matrix under conditional homoskedasticity.

**Theorem 13.6** Under Assumption 12.2 and  $\mathbb{E}(e_i^2 | \mathbf{z}_i) = \sigma^2$  then  $\hat{\beta}_{\text{2sls}}$  is efficient GMM.

This shows that 2SLS is efficient under homoskedasticity. When homoskedasticity holds, there is no reason to use efficient GMM over 2SLS. More broadly, when homoskedasticity is a reasonable approximation then 2SLS will be a reasonable estimator. However, this result also shows that in the general case where the error is conditionally heteroskedastic, then 2SLS is generically inefficient relative to efficient GMM.

### 13.10 Estimation of the Efficient Weight Matrix

To construct the efficient GMM estimator we need a consistent estimator  $\widehat{\mathbf{W}}$  of  $\mathbf{W}_0 = \boldsymbol{\Omega}^{-1}$ . The convention is to form an estimate  $\widehat{\boldsymbol{\Omega}}$  of  $\boldsymbol{\Omega}$  and then set  $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Omega}}^{-1}$ .

The **two-step GMM estimator** proceeds by using a one-step consistent estimate of  $\boldsymbol{\beta}$  to construct the weight matrix estimator  $\widehat{\mathbf{W}}$ . In the linear model the natural one-step estimator for  $\boldsymbol{\beta}$  is the 2SLS estimator  $\hat{\beta}_{\text{2sls}}$ . Set  $\tilde{e}_i = y_i - \mathbf{x}'_i \hat{\beta}_{\text{2sls}}$ ,  $\tilde{\mathbf{g}}_i = \mathbf{g}_i(\hat{\boldsymbol{\beta}}) = \mathbf{z}_i \tilde{e}_i$  and  $\bar{\mathbf{g}}_n = n^{-1} \sum_{i=1}^n \tilde{\mathbf{g}}_i$ . Two moment estimators of  $\boldsymbol{\Omega}$  are then

$$\widehat{\boldsymbol{\Omega}} = \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{g}}_i \tilde{\mathbf{g}}'_i \quad (13.8)$$

and

$$\widehat{\boldsymbol{\Omega}}^* = \frac{1}{n} \sum_{i=1}^n (\tilde{\mathbf{g}}_i - \bar{\mathbf{g}}_n)(\tilde{\mathbf{g}}_i - \bar{\mathbf{g}}_n)'. \quad (13.9)$$

The estimator (13.8) is an uncentered covariance matrix estimator while the estimator (13.9) is a centered version. Either estimator is consistent when  $\mathbb{E}(\mathbf{z}_i e_i) = \mathbf{0}$  which holds under correct specification. However under misspecification we may have  $\mathbb{E}(\mathbf{z}_i e_i) \neq \mathbf{0}$ . In the latter context  $\widehat{\boldsymbol{\Omega}}^*$  may be viewed as a robust estimator. For some testing problems it turns out to be preferable to use a covariance matrix estimator which is robust to the alternative hypothesis. For these reasons estimator (13.9) is generally preferred. Unfortunately, estimator (13.8) is more commonly seen in practice since it is the default choice by most packages. It is also worth observing that when the model is just identified then  $\bar{\mathbf{g}}_n = \mathbf{0}$  so the two are algebraically identical.

Given the choice of covariance matrix estimator we set  $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Omega}}^{-1}$  or  $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Omega}}^{*-1}$ . Given this weight matrix, we then construct the **two-step GMM estimator** as (13.6) using the weight matrix  $\widehat{\mathbf{W}}$ .

Since the 2SLS estimator is consistent for  $\boldsymbol{\beta}$ , by arguments nearly identical to those used for covariance matrix estimation, we can show that  $\widehat{\boldsymbol{\Omega}}$  and  $\widehat{\boldsymbol{\Omega}}^*$  are consistent for  $\boldsymbol{\Omega}$  and thus  $\widehat{\mathbf{W}}$  is consistent for  $\boldsymbol{\Omega}^{-1}$ . See Exercise 13.3.

This also means that the two-step GMM estimator satisfies the conditions for Theorem 13.4. We have established.

**Theorem 13.7** Under Assumption 12.2 and  $\boldsymbol{\Omega} > 0$ , if  $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Omega}}^{-1}$  or  $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Omega}}^{*-1}$  where the latter are defined in (13.8) and (13.9) then as  $n \rightarrow \infty$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, V_{\boldsymbol{\beta}})$$

where

$$V_{\boldsymbol{\beta}} = (\mathbf{Q}' \boldsymbol{\Omega}^{-1} \mathbf{Q})^{-1}.$$

This shows that the two-step GMM estimator is asymptotically efficient.

The two-step GMM estimator of the IV regression equation can be computed in Stata using the `ivregress gmm` command. By default it uses formula (13.8). The centered version (13.9) may be selected using the `center` option.

## 13.11 Iterated GMM

The asymptotic distribution of the two-step GMM estimator does not depend on the choice of the preliminary one-step estimator. However, the actual value of the estimator depends on this choice, and so will the finite sample distribution. This is undesirable and likely inefficient. To remove this dependence we can iterate the estimation sequence. Specifically, given  $\hat{\beta}_{\text{gmm}}$  we can construct an updated weight matrix estimate  $\hat{W}$  and then re-estimate  $\hat{\beta}_{\text{gmm}}$ . This updating can be iterated until convergence<sup>1</sup>. The result is called the **iterated GMM estimator** and is a common implementation of efficient GMM.

Interestingly, B. Hansen and Lee (2018) show that the iterated GMM estimator is unaffected if the weight matrix is computed with or without centering. Standard errors and test statistics, however, will be affected by the choice.

The iterated GMM estimator of the IV regression equation can be computed in Stata using the `ivregress gmm` command using the `igmm` option.

## 13.12 Covariance Matrix Estimation

An estimator of the asymptotic variance of  $\hat{\beta}_{\text{gmm}}$  can be obtained by replacing the matrices in the asymptotic variance formula by consistent estimates.

For the one-step or two-step GMM estimator the covariance matrix estimator is

$$\hat{V}_{\beta} = (\hat{Q}' \hat{W} \hat{Q})^{-1} (\hat{Q}' \hat{W} \hat{\Omega} \hat{W} \hat{Q}) (\hat{Q}' \hat{W} \hat{Q})^{-1} \quad (13.10)$$

where

$$\hat{Q} = \frac{1}{n} \sum_{i=1}^n z_i x_i'$$

and using either the uncentered estimator (13.8) or centered estimator (13.9) with the residuals  $\hat{e}_i = y_i - x_i' \hat{\beta}_{\text{gmm}}$ .

For the efficient iterated gmm estimator the covariance matrix estimator is

$$\hat{V}_{\beta} = (\hat{Q}' \hat{\Omega}^{-1} \hat{Q})^{-1} = \left( \left( \frac{1}{n} X' Z \right) \hat{\Omega}^{-1} \left( \frac{1}{n} Z' X \right) \right)^{-1}. \quad (13.11)$$

$\hat{\Omega}$  can be computed using either the uncentered estimator (13.8) or centered estimator (13.9). Based on the asymptotic approximation the estimator (13.11) can be used as well for the two-step estimator but should use the final residuals  $\hat{e}_i = y_i - x_i' \hat{\beta}_{\text{gmm}}$ .

Asymptotic standard errors are given by the square roots of the diagonal elements of  $n^{-1} \hat{V}_{\beta}$ .

In Stata, the default covariance matrix estimation method is determined by the choice of weight matrix. Thus if the centered estimator (13.9) is used for the weight matrix, it is also used for the covariance matrix estimator.

---

<sup>1</sup>In practice, “convergence” obtains when the difference between the estimates obtained at subsequent steps is smaller than a pre-specified tolerance. A sufficient condition for convergence is that the sequence is a contraction mapping. Indeed, B. Hansen and Lee (2018) have shown that the iterated GMM estimator generally satisfies this condition in large samples.

### 13.13 Clustered Dependence

In Section 4.21 we introduced clustered dependence and in Section 12.25 described covariance matrix estimation for 2SLS. The methods extend naturally to GMM, but with the additional complication of potentially altering weight matrix calculation.

As before, the structural equation for the  $g^{th}$  cluster can be written as the matrix system

$$\mathbf{y}_g = \mathbf{X}_g \boldsymbol{\beta} + \mathbf{e}_g.$$

Using this notation the centered GMM estimator with weight matrix  $\mathbf{W}$  can be written as

$$\hat{\boldsymbol{\beta}}_{\text{gmm}} = (\mathbf{Z}' \mathbf{Z} \mathbf{W} \mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{Z} \mathbf{W} \left( \sum_{g=1}^G \mathbf{Z}'_g \mathbf{e}_g \right).$$

The cluster-robust covariance matrix estimator for  $\hat{\boldsymbol{\beta}}_{\text{gmm}}$  is then

$$\hat{\mathbf{V}}_{\boldsymbol{\beta}} = (\mathbf{Z}' \mathbf{Z} \mathbf{W} \mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{Z} \hat{\mathbf{S}} \mathbf{W} \mathbf{Z}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z} \mathbf{W} \mathbf{Z}' \mathbf{Z})^{-1} \quad (13.12)$$

with

$$\hat{\mathbf{S}} = \sum_{g=1}^G \mathbf{Z}'_g \hat{\mathbf{e}}_g \hat{\mathbf{e}}'_g \mathbf{Z}_g \quad (13.13)$$

and the clustered residuals

$$\hat{\mathbf{e}}_g = \mathbf{y}_g - \mathbf{X}_g \hat{\boldsymbol{\beta}}_{\text{gmm}}. \quad (13.14)$$

The cluster-robust estimator (13.12) is appropriate for the one-step or two-step GMM estimator. It is also appropriate for the iterated estimator when the latter uses a conventional (non-clustered) efficient weight matrix. However in the clustering context it is more natural to use a cluster-robust weight matrix such as  $\mathbf{W} = \hat{\mathbf{S}}^{-1}$  where  $\hat{\mathbf{S}}$  is a cluster-robust covariance estimator as in (13.13) based on a one-step or iterated residual. This gives rise to the cluster-robust GMM estimator

$$\hat{\boldsymbol{\beta}}_{\text{gmm}} = (\mathbf{Z}' \mathbf{Z} \hat{\mathbf{S}}^{-1} \mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{Z} \hat{\mathbf{S}}^{-1} \mathbf{Z}' \mathbf{y}. \quad (13.15)$$

For this estimator (especially when iterated) an appropriate cluster-robust covariance matrix estimator is

$$\hat{\mathbf{V}}_{\boldsymbol{\beta}} = (\mathbf{Z}' \mathbf{Z} \hat{\mathbf{S}}^{-1} \mathbf{Z}' \mathbf{Z})^{-1}$$

where  $\hat{\mathbf{S}}$  is calculated using the final residuals.

To implement a cluster-robust weight matrix, use the 2SLS estimator for first step. Compute the cluster residuals (13.14) and covariance matrix (13.13). Then (13.15) is the two-step GMM estimator. Updating the residuals and covariance matrix, we can iterate the sequence to obtain the iterated GMM estimator.

In Stata, using the `ivregress gmm` command with the `cluster` option implements the two-step GMM estimator using the cluster-robust weight matrix and cluster-robust covariance matrix estimator. To use the centered covariance matrix use the `center` option, and to implement the iterated GMM estimator use the `igmm` option. Alternatively, you can use the `wmatrix` and `vce` options to separately specify the weight matrix and covariance matrix estimation methods.

### 13.14 Wald Test

For a given function  $\mathbf{r}(\boldsymbol{\beta}) : \mathbb{R}^k \rightarrow \Theta \subset \mathbb{R}^q$  we define the parameter  $\boldsymbol{\theta} = \mathbf{r}(\boldsymbol{\beta})$ . The GMM estimator of  $\boldsymbol{\theta}$  is  $\hat{\boldsymbol{\theta}}_{\text{gmm}} = \mathbf{r}(\hat{\boldsymbol{\beta}}_{\text{gmm}})$ . By the delta method it is asymptotically normal with covariance matrix

$$\begin{aligned} \mathbf{V}_{\boldsymbol{\theta}} &= \mathbf{R}' \mathbf{V}_{\boldsymbol{\beta}} \mathbf{R} \\ \mathbf{R} &= \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{r}(\boldsymbol{\beta})'. \end{aligned}$$

An estimator of the asymptotic covariance matrix is

$$\begin{aligned}\widehat{V}_{\boldsymbol{\theta}} &= \widehat{\mathbf{R}}' \widehat{V}_{\boldsymbol{\beta}} \widehat{\mathbf{R}} \\ \widehat{\mathbf{R}} &= \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{r}(\widehat{\boldsymbol{\beta}}_{\text{gmm}})'.\end{aligned}$$

When  $\theta$  is scalar then an asymptotic standard error for  $\widehat{\theta}_{\text{gmm}}$  is formed as  $\sqrt{n^{-1} \widehat{V}_{\boldsymbol{\theta}}}$ .

A standard test of the hypothesis

$$\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$$

against

$$\mathbb{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$$

is based on the Wald statistic

$$W = n(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \widehat{V}_{\boldsymbol{\theta}}^{-1} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

Let  $G_q(u)$  denote the  $\chi^2_q$  distribution function.

**Theorem 13.8** Under Assumptions 12.2 and 7.3, and  $\mathbb{H}_0$  holds, as  $n \rightarrow \infty$ ,

$$W \xrightarrow{d} \chi^2_q.$$

For  $c$  satisfying  $\alpha = 1 - G_q(c)$ ,

$$\mathbb{P}(W > c | \mathbb{H}_0) \longrightarrow \alpha$$

so the test “Reject  $\mathbb{H}_0$  if  $W > c$ ” has asymptotic size  $\alpha$ .

For a proof see Exercise 13.5.

In Stata, the commands `test` and `testparm` can be used after `ivregress gmm` to implement Wald tests of linear hypotheses. The commands `nlcom` and `testnl` can be used after `ivregress gmm` to implement Wald tests of nonlinear hypotheses.

### 13.15 Restricted GMM

It is often desirable to impose restrictions on the coefficients. In this section we consider estimation subject to the linear constraints  $\mathbf{R}' \boldsymbol{\beta} = \mathbf{c}$ . In the following section we consider nonlinear constraints.

The constrained GMM estimator minimizes the GMM criterion subject to the constraint. It is defined as

$$\widehat{\boldsymbol{\beta}}_{\text{cgmm}} = \underset{\mathbf{R}' \boldsymbol{\beta} = \mathbf{c}}{\operatorname{argmin}} J(\boldsymbol{\beta}).$$

This is the parameter vector which makes the estimating equations as close to zero as possible with respect to the weighted quadratic distance while imposing the restriction on the parameters.

Suppose the weight matrix  $\mathbf{W}$  is fixed. Using the methods of Chapter 8 it is straightforward to derive that the constrained GMM estimator is

$$\widehat{\boldsymbol{\beta}}_{\text{cgmm}} = \widehat{\boldsymbol{\beta}}_{\text{gmm}} - (\mathbf{X}' \mathbf{Z} \mathbf{W} \mathbf{Z}' \mathbf{X})^{-1} \mathbf{R} \left( \mathbf{R}' (\mathbf{X}' \mathbf{Z} \mathbf{W} \mathbf{Z}' \mathbf{X})^{-1} \mathbf{R} \right)^{-1} \left( \mathbf{R}' \widehat{\boldsymbol{\beta}}_{\text{gmm}} - \mathbf{c} \right). \quad (13.16)$$

(For details, see Exercise 13.6.)

We derive the asymptotic distribution under the assumption that the restriction is true. Make the substitution  $\mathbf{c} = \mathbf{R}'\boldsymbol{\beta}$  in (13.16) and reorganize to find

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{cgmm}} - \boldsymbol{\beta}) = \left( \mathbf{I}_k - (\mathbf{X}' \mathbf{Z} \mathbf{W} \mathbf{Z}' \mathbf{X})^{-1} \mathbf{R} \left( \mathbf{R}' (\mathbf{X}' \mathbf{Z} \mathbf{W} \mathbf{Z}' \mathbf{X})^{-1} \mathbf{R} \right)^{-1} \mathbf{R}' \right) \sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta}). \quad (13.17)$$

This is a linear function of  $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta})$ . Since the asymptotic distribution of the latter is known, the asymptotic distribution of  $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{cgmm}} - \boldsymbol{\beta})$  is a linear function of the former.

**Theorem 13.9** Under Assumptions 12.2 and 8.3, for the constrained GMM estimator (13.16),

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{cgmm}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}_{\text{cgmm}})$$

as  $n \rightarrow \infty$ , where

$$\begin{aligned} \mathbf{V}_{\text{cgmm}} &= \mathbf{V}_{\boldsymbol{\beta}} - (\mathbf{Q}' \mathbf{W} \mathbf{Q})^{-1} \mathbf{R} \left( \mathbf{R}' (\mathbf{Q}' \mathbf{W} \mathbf{Q})^{-1} \mathbf{R} \right)^{-1} \mathbf{R}' \mathbf{V}_{\boldsymbol{\beta}} \\ &\quad - \mathbf{V}_{\boldsymbol{\beta}} \mathbf{R} \left( \mathbf{R}' (\mathbf{Q}' \mathbf{W} \mathbf{Q})^{-1} \mathbf{R} \right)^{-1} \mathbf{R}' (\mathbf{Q}' \mathbf{W} \mathbf{Q})^{-1} \\ &\quad + (\mathbf{Q}' \mathbf{W} \mathbf{Q})^{-1} \mathbf{R} \left( \mathbf{R}' (\mathbf{Q}' \mathbf{W} \mathbf{Q})^{-1} \mathbf{R} \right)^{-1} \mathbf{R}' \mathbf{V}_{\boldsymbol{\beta}} \mathbf{R} \left( \mathbf{R}' (\mathbf{Q}' \mathbf{W} \mathbf{Q})^{-1} \mathbf{R} \right)^{-1} \mathbf{R}' (\mathbf{Q}' \mathbf{W} \mathbf{Q})^{-1} \end{aligned} \quad (13.18)$$

For a proof, see Exercise 13.8. Unfortunately the asymptotic covariance matrix formula (13.18) is quite tedious!

Now suppose that the weight matrix is set as  $\mathbf{W} = \tilde{\boldsymbol{\Omega}}^{-1}$ , the efficient weight matrix from unconstrained estimation. In this case the constrained GMM estimator can be written as

$$\hat{\boldsymbol{\beta}}_{\text{cgmm}} = \hat{\boldsymbol{\beta}}_{\text{gmm}} - \tilde{\mathbf{V}}_{\boldsymbol{\beta}} \mathbf{R} (\mathbf{R}' \tilde{\mathbf{V}}_{\boldsymbol{\beta}} \mathbf{R})^{-1} (\mathbf{R}' \hat{\boldsymbol{\beta}}_{\text{gmm}} - \mathbf{c}) \quad (13.19)$$

which is the same formula (8.25) as efficient minimum distance. (For details, see Exercise 13.7.) We also find that the asymptotic covariance matrix simplifies considerably.

**Theorem 13.10** Under Assumptions 12.2 and 8.3, for the efficient constrained GMM estimator (13.19),

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{cgmm}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}_{\text{cgmm}})$$

as  $n \rightarrow \infty$ , where

$$\mathbf{V}_{\text{cgmm}} = \mathbf{V}_{\boldsymbol{\beta}} - \mathbf{V}_{\boldsymbol{\beta}} \mathbf{R} (\mathbf{R}' \mathbf{V}_{\boldsymbol{\beta}} \mathbf{R})^{-1} \mathbf{R}' \mathbf{V}_{\boldsymbol{\beta}}. \quad (13.20)$$

For a proof, see Exercise 13.9.

The asymptotic covariance matrix (13.20) can be estimated by

$$\tilde{\mathbf{V}}_{\text{cgmm}} = \tilde{\mathbf{V}}_{\boldsymbol{\beta}} - \tilde{\mathbf{V}}_{\boldsymbol{\beta}} \mathbf{R} (\mathbf{R}' \tilde{\mathbf{V}}_{\boldsymbol{\beta}} \mathbf{R})^{-1} \mathbf{R}' \tilde{\mathbf{V}}_{\boldsymbol{\beta}}. \quad (13.21)$$

$$\begin{aligned} \tilde{\mathbf{V}}_{\boldsymbol{\beta}} &= (\tilde{\boldsymbol{\Omega}}' \tilde{\boldsymbol{\Omega}}^{-1} \tilde{\boldsymbol{\Omega}})^{-1} \\ \tilde{\boldsymbol{\Omega}} &= \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i' \tilde{e}_i^2 \\ \tilde{e}_i &= y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{\text{cgmm}}. \end{aligned} \quad (13.22)$$

The covariance matrix (13.18) can be estimated similarly, though using (13.10) to estimate  $\mathbf{V}_\beta$ . The covariance matrix estimator  $\tilde{\Omega}$  can also be replaced with a centered version.

A constrained iterated GMM estimator can be implemented by setting  $\mathbf{W} = \tilde{\Omega}^{-1}$  where  $\tilde{\Omega}$  is defined in (13.22), and then iterating until convergence. This is a natural estimator, as it is the appropriate implementation of the idea of iterated GMM.

Since both  $\hat{\Omega}$  and  $\tilde{\Omega}$  converge to the same limit  $\Omega$  (under the assumption that the constraint is true), the constrained iterated GMM estimator has the same asymptotic distribution as given in Theorem 13.10.

### 13.16 Nonlinear Restricted GMM

Nonlinear constraints on the parameters can be written as  $\mathbf{r}(\beta) = \mathbf{0}$  for some function where  $\mathbf{r} : \mathbb{R}^k \rightarrow \mathbb{R}^q$ . Least-squares estimation subject to nonlinear constraints was explored in Section 8.14. In this section we introduce GMM estimation subject to nonlinear constraints. The constraint is nonlinear if  $\mathbf{r}(\beta)$  cannot be written as a linear function of  $\beta$ .

The constrained GMM estimator minimizes the GMM criterion subject to the constraint. It is defined as

$$\hat{\beta}_{\text{cgmm}} = \underset{\mathbf{r}(\beta)=\mathbf{0}}{\operatorname{argmin}} J(\beta). \quad (13.23)$$

This is the parameter vector which makes the estimating equations as close to zero as possible with respect to the weighted quadratic distance while imposing the restriction on the parameters.

In general there is no explicit solution for  $\hat{\beta}_{\text{cgmm}}$ . Instead, the solution needs to be found numerically. Fortunately there are excellent nonlinear constrained optimization solvers which make the task quite feasible. We do not review these here, but can be found in any numerical software system.

For the asymptotic distribution assume that the restriction  $\mathbf{r}(\beta) = \mathbf{0}$  is true. Then, using the same methods as in the proof of Theorem 8.10 we can show that (13.17) approximately holds, in the sense that

$$\sqrt{n}(\hat{\beta}_{\text{cgmm}} - \beta) = \left( \mathbf{I}_k - (\mathbf{X}' \mathbf{Z} \mathbf{W} \mathbf{Z}' \mathbf{X})^{-1} \mathbf{R} \left( \mathbf{R}' (\mathbf{X}' \mathbf{Z} \mathbf{W} \mathbf{Z}' \mathbf{X})^{-1} \mathbf{R} \right)^{-1} \mathbf{R}' \right) \sqrt{n}(\hat{\beta}_{\text{gmm}} - \beta) + o_p(1)$$

where  $\mathbf{R} = \frac{\partial}{\partial \beta} \mathbf{r}(\beta)'$ . Thus the asymptotic distribution of the constrained estimator takes the same form as in the linear case.

**Theorem 13.11** Under Assumptions 12.2 and 8.3, for the constrained GMM estimator (13.23)

$$\sqrt{n}(\hat{\beta}_{\text{cgmm}} - \beta) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_{\text{cgmm}})$$

as  $n \rightarrow \infty$ , where  $\mathbf{V}_{\text{cgmm}}$  equals (13.18). If  $\mathbf{W} = \hat{\Omega}^{-1}$ , then  $\mathbf{V}_{\text{cgmm}}$  equals (13.20).

$$\mathbf{V}_{\text{cgmm}} = \mathbf{V}_\beta - \mathbf{V}_\beta \mathbf{R} (\mathbf{R}' \mathbf{V}_\beta \mathbf{R})^{-1} \mathbf{R}' \mathbf{V}_\beta.$$

The asymptotic covariance matrix in the efficient case is estimated by (13.21) with  $\mathbf{R}$  replaced with

$$\hat{\mathbf{R}} = \frac{\partial}{\partial \beta} \mathbf{r}(\hat{\beta}_{\text{cgmm}})'.$$

The asymptotic covariance matrix (13.18) in the general case is estimated similarly.

To implement an iterated restricted GMM estimator, the weight matrix may be set as  $\mathbf{W} = \tilde{\Omega}^{-1}$  where  $\tilde{\Omega}$  is defined in (13.22), and then iterated until convergence.

### 13.17 Constrained Regression

Take the conventional projection model

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) &= \mathbf{0}. \end{aligned}$$

We can view this as a very special case of GMM. It is model (13.5) with  $\mathbf{z}_i = \mathbf{x}_i$ . This is just-identified GMM and the estimator is least-squares  $\hat{\boldsymbol{\beta}}_{\text{gmm}} = \hat{\boldsymbol{\beta}}_{\text{ols}}$ .

In Chapter 8 we discussed estimation of the projection model subject to linear constraints  $\mathbf{R}'\boldsymbol{\beta} = \mathbf{c}$ , which includes exclusion restrictions. Since the projection model is a special case of GMM, the constrained projection model is also constrained GMM. From the results of Section 13.15 we find that the efficient constrained GMM estimator is

$$\hat{\boldsymbol{\beta}}_{\text{cgmm}} = \hat{\boldsymbol{\beta}}_{\text{ols}} - \hat{V}_{\boldsymbol{\beta}} \mathbf{R} (\mathbf{R}' \hat{V}_{\boldsymbol{\beta}} \mathbf{R})^{-1} (\mathbf{R}' \hat{\boldsymbol{\beta}}_{\text{ols}} - \mathbf{c}) = \hat{\boldsymbol{\beta}}_{\text{emd}},$$

the efficient minimum distance estimator. Thus for linear constraints on the linear projection model, efficient GMM equals efficient minimum distance. Thus one convenient method to implement efficient minimum distance is by using GMM.

### 13.18 Multivariate Regression

GMM methods can simplify estimation and inference for multivariate regressions such as those introduced in Chapter 11.

The general multivariate regression (projection) model is

$$\begin{aligned} y_{ji} &= \mathbf{x}'_{ji} \boldsymbol{\beta}_j + e_{ji} \\ \mathbb{E}(\mathbf{x}_{ji} e_{ji}) &= \mathbf{0} \end{aligned}$$

for  $j = 1, \dots, m$ . Using the notation from Section 11.2 the equations can be written jointly as

$$\mathbf{y}_i = \bar{\mathbf{X}}_i \boldsymbol{\beta} + \mathbf{e}_i$$

and for the full sample as

$$\mathbf{y} = \bar{\mathbf{X}} \boldsymbol{\beta} + \mathbf{e}.$$

The  $\bar{k}$  moment conditions are

$$\mathbb{E}\left(\bar{\mathbf{X}}'_i (\mathbf{y}_i - \bar{\mathbf{X}}_i \boldsymbol{\beta})\right) = \mathbf{0}. \quad (13.24)$$

Given a  $\bar{k} \times \bar{k}$  weight matrix  $\mathbf{W}$  the GMM criterion is

$$J(\boldsymbol{\beta}) = n \left( \mathbf{y} - \bar{\mathbf{X}} \boldsymbol{\beta} \right)' \bar{\mathbf{X}} \mathbf{W} \bar{\mathbf{X}}' \left( \mathbf{y} - \bar{\mathbf{X}} \boldsymbol{\beta} \right).$$

The GMM estimator  $\hat{\boldsymbol{\beta}}_{\text{gmm}}$  minimizes  $J(\boldsymbol{\beta})$ . Since this is a just-identified model the estimator solves the sample equations

$$\bar{\mathbf{X}}' \left( \mathbf{y} - \bar{\mathbf{X}} \hat{\boldsymbol{\beta}}_{\text{gmm}} \right) = \mathbf{0}.$$

The solution is

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{gmm}} &= \left( \sum_{i=1}^n \bar{\mathbf{X}}'_i \bar{\mathbf{X}}_i \right)^{-1} \left( \sum_{i=1}^n \bar{\mathbf{X}}'_i \mathbf{y}_i \right) \\ &= (\bar{\mathbf{X}}' \bar{\mathbf{X}})^{-1} (\bar{\mathbf{X}}' \mathbf{y}) \\ &= \hat{\boldsymbol{\beta}}_{\text{ols}}, \end{aligned}$$

the least-squares estimator.

Thus the unconstrained GMM estimator of the multivariate regression model is least-squares. The estimator does not depend on the weight matrix since the model is just-identified.

A important advantage of the GMM framework is to incorporate constraints. Consider the class of restrictions  $\mathbf{R}'\boldsymbol{\beta} = \mathbf{c}$ . Minimization of the GMM criterion subject to this restriction has solutions as described in (13.15). The restricted GMM estimator is

$$\hat{\boldsymbol{\beta}}_{\text{gmm}} = \hat{\boldsymbol{\beta}}_{\text{ols}} - \left( \bar{\mathbf{X}}' \bar{\mathbf{X}} \mathbf{W} \bar{\mathbf{X}}' \bar{\mathbf{X}} \right)^{-1} \mathbf{R} \left( \mathbf{R}' \left( \bar{\mathbf{X}}' \bar{\mathbf{X}} \mathbf{W} \bar{\mathbf{X}}' \bar{\mathbf{X}} \right)^{-1} \mathbf{R} \right)^{-1} (\mathbf{R}' \hat{\boldsymbol{\beta}}_{\text{ols}} - \mathbf{c}).$$

This estimator depends on the weight matrix because it is over-identified.

A simple choice for weight matrix is  $\mathbf{W} = \bar{\mathbf{X}}' \bar{\mathbf{X}}$ . This leads to the one-step estimator

$$\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_{\text{ols}} - \left( \bar{\mathbf{X}}' \bar{\mathbf{X}} \right)^{-1} \mathbf{R} \left( \mathbf{R}' \left( \bar{\mathbf{X}}' \bar{\mathbf{X}} \right)^{-1} \mathbf{R} \right)^{-1} (\mathbf{R}' \hat{\boldsymbol{\beta}}_{\text{ols}} - \mathbf{c}).$$

The asymptotically efficient choice sets  $\mathbf{W} = \hat{\boldsymbol{\Omega}}^{-1}$  where  $\hat{\boldsymbol{\Omega}} = n^{-1} \sum_{i=1}^n \bar{\mathbf{X}}_i' \hat{\mathbf{e}}_i \hat{\mathbf{e}}_i' \bar{\mathbf{X}}_i$  and  $\hat{\mathbf{e}}_i = \mathbf{y}_i - \bar{\mathbf{X}}_i \hat{\boldsymbol{\beta}}_1$ . This leads to the two-step estimator

$$\hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_{\text{ols}} - \left( \bar{\mathbf{X}}' \bar{\mathbf{X}} \hat{\boldsymbol{\Omega}}^{-1} \bar{\mathbf{X}}' \bar{\mathbf{X}} \right)^{-1} \mathbf{R} \left( \mathbf{R}' \left( \bar{\mathbf{X}}' \bar{\mathbf{X}} \hat{\boldsymbol{\Omega}}^{-1} \bar{\mathbf{X}}' \bar{\mathbf{X}} \right)^{-1} \mathbf{R} \right)^{-1} (\mathbf{R}' \hat{\boldsymbol{\beta}}_{\text{ols}} - \mathbf{c}).$$

When the regressors  $\mathbf{x}_i$  are common across all equations, then the multivariate regression model can be written conveniently as in (11.3)

$$\begin{aligned} \mathbf{y}_i &= \mathbf{B}' \mathbf{x}_i + \mathbf{e}_i \\ \mathbb{E}(\mathbf{x}_i \mathbf{e}_i') &= \mathbf{0}. \end{aligned}$$

The moment restrictions can be written as the matrix system

$$\mathbb{E}(\mathbf{x}_i (\mathbf{y}_i' - \mathbf{x}_i' \mathbf{B})) = \mathbf{0}.$$

Written as a vector system this is (13.24) and thus leads to the same restricted GMM estimators.

These are general formula for imposing restrictions. In specific cases (such as an exclusion restriction) direct methods may be more convenient. In all cases, the solution is found by minimization of the GMM criterion  $J(\boldsymbol{\beta})$  subject to the restriction.

### 13.19 Distance Test

In Section 13.14 we introduced Wald tests of the hypothesis  $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$  where  $\boldsymbol{\theta} = \mathbf{r}(\boldsymbol{\beta})$  for a given function  $\mathbf{r}(\boldsymbol{\beta}) : \mathbb{R}^k \rightarrow \Theta \subset \mathbb{R}^q$ . When  $\mathbf{r}(\boldsymbol{\beta})$  is non-linear, an alternative is to use a criterion-based statistic. This is sometimes called the GMM Distance statistic and sometimes called a LR-like statistic (the LR is for likelihood-ratio). The idea was first put forward by Newey and West (1987a).

The idea is to compare the unrestricted and restricted estimators by contrasting the criterion functions. The unrestricted estimator takes the form

$$\hat{\boldsymbol{\beta}}_{\text{gmm}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} J(\boldsymbol{\beta})$$

where

$$\hat{J}(\boldsymbol{\beta}) = n \cdot \bar{\mathbf{g}}_n(\boldsymbol{\beta})' \hat{\boldsymbol{\Omega}}^{-1} \bar{\mathbf{g}}_n(\boldsymbol{\beta})$$

is the unrestricted GMM criterion with an efficient weight matrix estimate  $\hat{\boldsymbol{\Omega}}$ . The minimized value of the criterion is

$$\hat{J} = \hat{J}(\hat{\boldsymbol{\beta}}_{\text{gmm}}).$$

As in Section 13.15, the estimator subject to  $\mathbf{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$  is

$$\hat{\boldsymbol{\beta}}_{\text{cgmm}} = \underset{\mathbf{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0}{\operatorname{argmin}} \tilde{J}(\boldsymbol{\beta})$$

where

$$\tilde{J}(\boldsymbol{\beta}) = n \cdot \bar{\mathbf{g}}_n(\boldsymbol{\beta})' \tilde{\boldsymbol{\Omega}}^{-1} \bar{\mathbf{g}}_n(\boldsymbol{\beta})$$

which depends on an efficient weight matrix estimate, either  $\hat{\boldsymbol{\Omega}}$  (the same as the unrestricted estimator), or  $\tilde{\boldsymbol{\Omega}}$  (the iterated weight matrix from constrained estimation). The minimized value of the criterion is

$$\tilde{J} = \tilde{J}(\hat{\boldsymbol{\beta}}_{\text{cgmm}}).$$

The GMM distance (or LR-like) statistic is the difference in the criterion functions.

$$D = \tilde{J} - \hat{J}.$$

The distance test shares the useful feature of LR tests in that it is a natural by-product of the computation of alternative models.

The test has the following large sample distribution.

**Theorem 13.12** Under Assumptions 12.2 and 7.3, and  $\mathbb{H}_0$  holds, then as  $n \rightarrow \infty$ ,

$$D \xrightarrow{d} \chi_q^2.$$

For  $c$  satisfying  $\alpha = 1 - G_q(c)$ ,

$$\mathbb{P}(D > c | \mathbb{H}_0) \longrightarrow \alpha$$

so the test “Reject  $\mathbb{H}_0$  if  $D > c$ ” has asymptotic size  $\alpha$ .

The proof is given in Section 13.28.

Theorem 13.12 shows that the distance statistic has the same asymptotic distribution as Wald and likelihood ratio statistics, and can be interpreted similarly. Small values of  $D$  mean that imposing the restriction does not result in a large value of the moment equations. Hence the restriction appears to be compatible with the data. On the other hand, large values of  $D$  mean that imposing the restriction results in a much larger value of the moment equations, implying that the restriction is not compatible with the data. The finding that the asymptotic distribution is chi-squared allows the calculation of asymptotic critical values and p-values.

We now discuss the choice of weight matrix. As mentioned above, one simple choice is to set  $\tilde{\boldsymbol{\Omega}} = \hat{\boldsymbol{\Omega}}$ . In this case we have the following result.

**Theorem 13.13** If  $\tilde{\boldsymbol{\Omega}} = \hat{\boldsymbol{\Omega}}$  then  $D \geq 0$ . Furthermore, if  $\mathbf{r}$  is linear in  $\boldsymbol{\beta}$ , then  $D$  equals the Wald statistic.

The statement that  $\tilde{\boldsymbol{\Omega}} = \hat{\boldsymbol{\Omega}}$  implies  $D \geq 0$  follows from the fact that in this case the criterion functions  $\hat{J}(\boldsymbol{\beta}) = \tilde{J}(\boldsymbol{\beta})$  are identical, so the constrained minimum cannot be smaller than the unconstrained. The statement that linear hypotheses and  $\tilde{\boldsymbol{\Omega}} = \hat{\boldsymbol{\Omega}}$  implies  $D = W$  follows from applying the expression for the constrained GMM estimator (13.19) and using the variance matrix formula (13.11).

This result shows some advantages to using the same weight matrix to estimate both  $\hat{\beta}_{\text{gmm}}$  and  $\hat{\beta}_{\text{cgmm}}$ . In particular, the non-negativity finding motivated Newey and West (1987a) to recommend using  $\tilde{\Omega} = \hat{\Omega}$ . However, this is not an important advantage. Setting  $\tilde{\Omega}$  to be the constrained efficient weight matrix is natural for efficient estimation of  $\hat{\beta}_{\text{cgmm}}$ . In the event that  $D < 0$  the test simply fails to reject  $H_0$  at any significance level.

As discussed in Section 9.17, for tests of nonlinear hypotheses the Wald statistic can work quite poorly. In particular, the Wald statistic is affected by how the hypothesis  $r(\beta)$  is formulated. In contrast, the distance statistic  $D$  is not affected by the algebraic formulation of the hypothesis. Current evidence suggests that the  $D$  statistic appears to have good sampling properties, and is a preferred test statistic relative to the Wald statistic for nonlinear hypotheses. (See B. Hansen (2006).)

In Stata, the command `estat overid` after `ivregress gmm` can be used to report the value of the GMM criterion  $J$ . By estimating the two nested GMM regressions the values  $\hat{J}$  and  $\tilde{J}$  can be obtained and  $D$  computed.

## 13.20 Continuously-Updated GMM

An alternative to the two-step GMM estimator can be constructed by letting the weight matrix be an explicit function of  $\beta$ . These leads to the criterion function

$$J(\beta) = n \cdot \bar{g}_n(\beta)' \left( \frac{1}{n} \sum_{i=1}^n g(w_i, \beta) g(w_i, \beta)' \right)^{-1} \bar{g}_n(\beta).$$

The  $\hat{\beta}$  which minimizes this function is called the **continuously-updated GMM (CU-GMM) estimator**, and was introduced by L. Hansen, Heaton and Yaron (1996).

A complication is that the continuously-updated criterion  $J(\beta)$  is not quadratic in  $\beta$ . This means that minimization requires numerical methods. It may appear that the CU-GMM estimator is the same as the iterated GMM estimator, but this is not the case at all. They solve distinct first-order conditions, and can be quite different in applications.

Relative to traditional GMM, the CU-GMM estimator has lower bias but thicker distributional tails. While it has received considerable theoretical attention, it is not used commonly in applications.

## 13.21 OverIdentification Test

In Section 12.31 we introduced the Sargan (1958) overidentification test for the 2SLS estimator under the assumption of homoskedasticity. L. Hansen (1982) generalized the test to cover the GMM estimator allowing for general heteroskedasticity.

Recall, overidentified models ( $\ell > k$ ) are special in the sense that there may not be a parameter value  $\beta$  such that the moment condition

$$H_0 : \mathbb{E}(z_i e_i) = \mathbf{0}$$

holds. Thus the model – the overidentifying restrictions – are testable.

For example, take the linear model  $y_i = \beta_1' x_{1i} + \beta_2' x_{2i} + e_i$  with  $\mathbb{E}(x_{1i} e_i) = \mathbf{0}$  and  $\mathbb{E}(x_{2i} e_i) = \mathbf{0}$ . It is possible that  $\beta_2 = \mathbf{0}$ , so that the linear equation may be written as  $y_i = \beta_1' x_{1i} + e_i$ . However, it is possible that  $\beta_2 \neq \mathbf{0}$ , and in this case it would be impossible to find a value of  $\beta_1$  so that both  $\mathbb{E}(x_{1i} (y_i - x_{1i}' \beta_1)) = \mathbf{0}$  and  $\mathbb{E}(x_{2i} (y_i - x_{1i}' \beta_1)) = \mathbf{0}$  hold simultaneously. In this sense an exclusion restriction can be seen as an overidentifying restriction.

Note that  $\bar{g}_n \xrightarrow{P} \mathbb{E}(z_i e_i)$ , and thus  $\bar{g}_n$  can be used to assess whether or not the hypothesis that  $\mathbb{E}(z_i e_i) = \mathbf{0}$  is true or not. Assuming that an efficient weight matrix estimate is used, the criterion function at the parameter estimates is

$$J = J(\hat{\beta}_{\text{gmm}}) = n \bar{g}_n' \hat{\Omega}^{-1} \bar{g}_n.$$

This is a quadratic form in  $\bar{\mathbf{g}}_n$ , and is thus a natural test statistic for  $\mathbb{H}_0 : \mathbb{E}(\mathbf{z}_i e_i) = \mathbf{0}$ . Note that we assume that the criterion function is constructed with an efficient weight matrix estimate. This is important for the distribution theory.

**Theorem 13.14** Under Assumption 12.2, then as  $n \rightarrow \infty$ ,

$$J = J(\hat{\boldsymbol{\beta}}_{\text{gmm}}) \xrightarrow{d} \chi^2_{\ell-k}.$$

For  $c$  satisfying  $\alpha = 1 - G_{\ell-k}(c)$ ,

$$\mathbb{P}(J > c | \mathbb{H}_0) \longrightarrow \alpha$$

so the test “Reject  $\mathbb{H}_0$  if  $J > c$ ” has asymptotic size  $\alpha$ .

The proof of the theorem is left to Exercise 13.13.

The degrees of freedom of the asymptotic distribution are the number of overidentifying restrictions. If the statistic  $J$  exceeds the chi-square critical value, we can reject the model. Based on this information alone it is unclear what is wrong, but it is typically cause for concern. The GMM overidentification test is a very useful by-product of the GMM methodology, and it is advisable to report the statistic  $J$  whenever GMM is the estimation method. When over-identified models are estimated by GMM, it is customary to report the  $J$  statistic as a general test of model adequacy.

In Stata, the command `estat overid` after `ivregress gmm` can be used to implement the overidentification test. The GMM criterion  $J$  and its asymptotic p-value using the  $\chi^2_{\ell-k}$  distribution are reported.

## 13.22 Subset OverIdentification Tests

In Section 12.32 we introduced subset overidentification tests for the 2SLS estimator under the assumption of homoskedasticity. In this section we describe how to construct analogous tests for the GMM estimator under general heteroskedasticity.

Recall, subset overidentification tests are used when it is desired to focus attention on a subset of instruments whose validity is questioned. Partition  $\mathbf{z}_i = (\mathbf{z}_{ai}, \mathbf{z}_{bi})$  with dimensions  $\ell_a$  and  $\ell_b$ , respectively, where  $\mathbf{z}_{ai}$  contains the instruments which are believed to be uncorrelated with  $e_i$ , and  $\mathbf{z}_{bi}$  contains the instruments which may be correlated with  $e_i$ . It is necessary to select this partition so that  $\ell_a > k$ , so that the instruments  $\mathbf{z}_{ai}$  alone identify the parameters. The instruments  $\mathbf{z}_{bi}$  are potentially valid additional instruments.

Given this partition, the maintained hypothesis is that  $\mathbb{E}(\mathbf{z}_{ai} e_i) = \mathbf{0}$ . The null and alternative hypotheses are

$$\begin{aligned}\mathbb{H}_0 : \mathbb{E}(\mathbf{z}_{bi} e_i) &= \mathbf{0} \\ \mathbb{H}_1 : \mathbb{E}(\mathbf{z}_{bi} e_i) &\neq \mathbf{0}.\end{aligned}$$

The GMM test is constructed as follows. First, estimate the model by efficient GMM with only the smaller set  $\mathbf{z}_{ai}$  of instruments. Let  $\tilde{J}$  denote the resulting GMM criterion. Second, estimate the model by efficient GMM with the full set  $\mathbf{z}_i = (\mathbf{z}_{ai}, \mathbf{z}_{bi})$  of instruments. Let  $\hat{J}$  denote the resulting GMM criterion. The test statistic is the difference in the criterion functions:

$$C = \hat{J} - \tilde{J}.$$

This is similar in form to the GMM distance statistic presented in Section 13.19. The difference is that the distance statistic compares models which differ based on the parameter restrictions, while the  $C$  statistic compares models based on different instrument sets.

Typically, the model with the greater instrument set will produce a larger value for  $J$  so that  $C \geq 0$ . However negative values can algebraically occur. That is okay for this simply leads to a non-rejection of  $\mathbb{H}_0$ .

If the smaller instrument set  $\mathbf{z}_{ai}$  is just-identified so that  $\ell_a = k$  then  $\tilde{J} = 0$  so  $C = \hat{J}$  is simply the standard overidentification test. This is why we have restricted attention to the case  $\ell_a > k$ .

The test has the following large sample distribution.

**Theorem 13.15** Under Assumption 12.2 and  $\mathbb{E}(\mathbf{z}_{ai}\mathbf{x}'_i)$  has full rank  $k$ , then as  $n \rightarrow \infty$ ,

$$C \xrightarrow{d} \chi^2_{\ell_b}.$$

For  $c$  satisfying  $\alpha = 1 - G_{\ell_b}(c)$ ,

$$\mathbb{P}(C > c | \mathbb{H}_0) \longrightarrow \alpha$$

so the test “Reject  $\mathbb{H}_0$  if  $C > c$ ” has asymptotic size  $\alpha$ .

The proof of Theorem 13.15 is presented in Section 13.28.

In Stata, the command `estat overid zb afer ivregress gmm` can be used to implement a subset overidentification test, where `zb` is the name(s) of the instruments(s) tested for validity. The statistic  $C$  and its asymptotic p-value using the  $\chi^2_{\ell_2}$  distribution are reported.

### 13.23 Endogeneity Test

In Section 12.29 we introduced tests for endogeneity in the context of 2SLS estimation. Endogeneity tests are simple to implement in the GMM framework as a subset overidentification test. The model is

$$y_i = \mathbf{x}'_{1i}\boldsymbol{\beta}_1 + \mathbf{x}'_{2i}\boldsymbol{\beta}_2 + e_i$$

where the maintained assumption is that the regressors  $\mathbf{x}_{1i}$  and excluded instruments  $\mathbf{z}_{2i}$  are exogenous so that  $\mathbb{E}(\mathbf{x}_{1i}e_i) = \mathbf{0}$  and  $\mathbb{E}(\mathbf{z}_{2i}e_i) = \mathbf{0}$ . The question is whether or not  $\mathbf{x}_{2i}$  is endogenous. Thus the null hypothesis is

$$\mathbb{H}_0 : \mathbb{E}(\mathbf{x}_{2i}e_i) = \mathbf{0}$$

with the alternative

$$\mathbb{H}_1 : \mathbb{E}(\mathbf{x}_{2i}e_i) \neq \mathbf{0}.$$

The GMM test is constructed as follows. First, estimate the model by efficient GMM using  $(\mathbf{x}_{1i}, \mathbf{z}_{2i})$  as instruments for  $(\mathbf{x}_{1i}, \mathbf{x}_{2i})$ . Let  $\tilde{J}$  denote the resulting GMM criterion. Second, estimate the model by efficient GMM using  $(\mathbf{x}_{1i}, \mathbf{x}_{2i}, \mathbf{z}_{2i})$  as instruments for  $(\mathbf{x}_{1i}, \mathbf{x}_{2i})$ . Let  $\hat{J}$  denote the resulting GMM criterion. The test statistic is the difference in the criterion functions:

$$C = \hat{J} - \tilde{J}.$$

The distribution theory for the test is a special case of the theory of overidentification testing.

**Theorem 13.16** Under Assumption 12.2 and  $\mathbb{E}(z_{2i}x'_{2i})$  has full rank  $k_2$ , then as  $n \rightarrow \infty$ ,

$$C \xrightarrow{d} \chi^2_{k_2}.$$

For  $c$  satisfying  $\alpha = 1 - G_{k_2}(c)$ ,

$$\mathbb{P}(C > c | \mathbb{H}_0) \longrightarrow \alpha$$

so the test “Reject  $\mathbb{H}_0$  if  $C > c$ ” has asymptotic size  $\alpha$ .

In Stata, the command `estat endogenous` after `ivregress gmm` can be used to implement the test for endogeneity. The statistic  $C$  and its asymptotic p-value using the  $\chi^2_{k_2}$  distribution are reported.

### 13.24 Subset Endogeneity Test

In Section 12.30 we introduced subset endogeneity tests for 2SLS estimation. GMM tests are simple to implement as subset overidentification tests. The model is

$$\begin{aligned} y_i &= \mathbf{x}'_{1i}\boldsymbol{\beta}_1 + \mathbf{x}'_{2i}\boldsymbol{\beta}_2 + \mathbf{x}'_{3i}\boldsymbol{\beta}_3 + e_i \\ \mathbb{E}(z_i e_i) &= \mathbf{0} \end{aligned}$$

where the instrument vector is  $\mathbf{z}_i = (\mathbf{x}_{1i}, \mathbf{x}_{2i})$ . The  $k_3 \times 1$  variables  $\mathbf{x}_{3i}$  are treated as endogenous, and the  $k_2 \times 1$  variables  $\mathbf{x}_{2i}$  are treated as potentially endogenous. The hypothesis to test is that  $\mathbf{x}_{2i}$  is exogenous, or

$$\mathbb{H}_0 : \mathbb{E}(\mathbf{x}_{2i}e_i) = \mathbf{0}$$

against

$$\mathbb{H}_1 : \mathbb{E}(\mathbf{x}_{2i}e_i) \neq \mathbf{0}.$$

The test requires that  $\ell_2 \geq (k_2 + k_3)$  so that the model can be estimated under  $\mathbb{H}_1$ .

The GMM test is constructed as follows. First, estimate the model by efficient GMM using  $(\mathbf{x}_{1i}, \mathbf{z}_{2i})$  as instruments for  $(\mathbf{x}_{1i}, \mathbf{x}_{2i}, \mathbf{x}_{3i})$ . Let  $\tilde{J}$  denote the resulting GMM criterion. Second, estimate the model by efficient GMM using  $(\mathbf{x}_{1i}, \mathbf{x}_{2i}, \mathbf{z}_{2i})$  as instruments for  $(\mathbf{x}_{1i}, \mathbf{x}_{2i}, \mathbf{x}_{3i})$ . Let  $\hat{J}$  denote the resulting GMM criterion. The test statistic is the difference in the criterion functions:

$$C = \hat{J} - \tilde{J}.$$

The distribution theory for the test is a special case of the theory of overidentification testing.

**Theorem 13.17** Under Assumption 12.2 and  $\mathbb{E}(z_{2i}(\mathbf{x}'_{2i}, \mathbf{x}'_{3i}))$  has full rank  $k_2 + k_3$ , then as  $n \rightarrow \infty$ ,

$$C \xrightarrow{d} \chi^2_{k_2}.$$

For  $c$  satisfying  $\alpha = 1 - G_{k_2}(c)$ ,

$$\mathbb{P}(C > c | \mathbb{H}_0) \longrightarrow \alpha$$

so the test “Reject  $\mathbb{H}_0$  if  $C > c$ ” has asymptotic size  $\alpha$ .

In Stata, the command `estat endogenous x2` after `ivregress gmm` can be used to implement the test for endogeneity, where `x2` is the name(s) of the variable(s) tested for endogeneity. The statistic  $C$  and its asymptotic p-value using the  $\chi^2_{k_2}$  distribution are reported.

## 13.25 Nonlinear GMM

GMM applies whenever an economic or statistical model implies the  $\ell \times 1$  moment condition

$$\mathbb{E}(\mathbf{g}_i(\boldsymbol{\beta})) = \mathbf{0}.$$

where  $\mathbf{g}_i(\boldsymbol{\beta})$  is a possibly nonlinear function of the parameters  $\boldsymbol{\beta}$ . Often, this is all that is known. Identification requires  $\ell \geq k = \dim(\boldsymbol{\beta})$ . The GMM estimator minimizes

$$J(\boldsymbol{\beta}) = n \cdot \bar{\mathbf{g}}_n(\boldsymbol{\beta})' \hat{\mathbf{W}} \bar{\mathbf{g}}_n(\boldsymbol{\beta})$$

for some weight matrix  $\hat{\mathbf{W}}$ , where

$$\bar{\mathbf{g}}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{g}_i(\boldsymbol{\beta}).$$

The efficient GMM estimator can be constructed by setting

$$\hat{\mathbf{W}} = \left( \frac{1}{n} \sum_{i=1}^n \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i' - \bar{\mathbf{g}}_n \bar{\mathbf{g}}_n' \right)^{-1},$$

with  $\hat{\mathbf{g}}_i = \mathbf{g}(\mathbf{w}_i, \tilde{\boldsymbol{\beta}})$  constructed using a preliminary consistent estimator  $\tilde{\boldsymbol{\beta}}$ , perhaps obtained by first setting  $\hat{\mathbf{W}} = \mathbf{I}_\ell$ .

As in the case of the linear model, the weight matrix can be iterated until convergence to obtain the iterated GMM estimator.

### Proposition 13.1 Distribution of Nonlinear GMM Estimator

Under general regularity conditions,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, V_{\boldsymbol{\beta}})$$

where

$$V_{\boldsymbol{\beta}} = (\mathbf{Q}' \mathbf{W} \mathbf{Q})^{-1} (\mathbf{Q}' \mathbf{W} \boldsymbol{\Omega} \mathbf{W} \mathbf{Q}) (\mathbf{Q}' \mathbf{W} \mathbf{Q})^{-1}$$

with

$$\boldsymbol{\Omega} = \mathbb{E}(\mathbf{g}_i \mathbf{g}_i')$$

and

$$\mathbf{Q} = \mathbb{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}'} \mathbf{g}_i(\boldsymbol{\beta})\right).$$

If the efficient weight matrix is used then

$$V_{\boldsymbol{\beta}} = (\mathbf{Q}' \boldsymbol{\Omega}^{-1} \mathbf{Q})^{-1}.$$

The proof of this result is omitted as it uses more advanced techniques.

The asymptotic covariance matrices can be estimated by sample counterparts of the population matrices. For the case of a general weight matrix,

$$\hat{V}_{\boldsymbol{\beta}} = (\hat{\mathbf{Q}}' \hat{\mathbf{W}} \hat{\mathbf{Q}})^{-1} (\hat{\mathbf{Q}}' \hat{\mathbf{W}} \hat{\boldsymbol{\Omega}} \hat{\mathbf{W}} \hat{\mathbf{Q}}) (\hat{\mathbf{Q}}' \hat{\mathbf{W}} \hat{\mathbf{Q}})^{-1}$$

where

$$\hat{\boldsymbol{\Omega}} = \frac{1}{n} \sum_{i=1}^n (\mathbf{g}_i(\hat{\boldsymbol{\beta}}) - \bar{\mathbf{g}})(\mathbf{g}_i(\hat{\boldsymbol{\beta}}) - \bar{\mathbf{g}})'$$

$$\bar{\mathbf{g}} = n^{-1} \sum_{i=1}^n \mathbf{g}_i(\hat{\boldsymbol{\beta}})$$

and

$$\hat{\mathbf{Q}} = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\beta}'} \mathbf{g}_i(\hat{\boldsymbol{\beta}}).$$

For the case of the iterated efficient weight matrix,

$$\hat{V}_{\boldsymbol{\beta}} = (\hat{\mathbf{Q}}' \hat{\boldsymbol{\Omega}}^{-1} \hat{\mathbf{Q}})^{-1}.$$

All of the methods discussed in this chapter – Wald tests, constrained estimation, Distance tests, overidentification tests, endogeneity tests – apply similarly to the nonlinear GMM estimator.

### 13.26 Bootstrap for GMM

The bootstrap for 2SLS (Section 12.23) can be used for GMM estimation. The standard bootstrap algorithm generates bootstrap samples by sampling the triplets  $(y_i^*, \mathbf{x}_i^*, \mathbf{z}_i^*)$  independently and with replacement from the original sample. The GMM estimator is applied to the bootstrap sample to obtain the bootstrap estimates  $\hat{\boldsymbol{\beta}}_{\text{gmm}}^*$ . This is repeated  $B$  times to create a sample of  $B$  bootstrap draws. Given these draws, bootstrap confidence intervals, including percentile, BC percentile,  $\text{BC}_a$  and percentile-t, are calculated conventionally.

For variance and standard error estimation, the same cautions apply as for 2SLS. It is difficult to know if the GMM estimator has a finite variance in a given application. It is best to avoid using the bootstrap to calculate standard errors. Instead, use the bootstrap for percentile-type and percentile-t confidence intervals.

When the model is overidentified, as discussed for 2SLS, bootstrap GMM inference will not achieve an asymptotic refinement unless the bootstrap estimator is recentered to satisfy the orthogonality condition. We now describe the recentering recommended by Hall and Horowitz (1996).

For linear GMM wth weight matrix  $\mathbf{W}$ , the recentered GMM bootstrap estimator is

$$\hat{\boldsymbol{\beta}}_{\text{gmm}}^{**} = (\mathbf{X}' \mathbf{Z}^* \mathbf{W}^* \mathbf{Z}' \mathbf{X}^*)^{-1} (\mathbf{X}' \mathbf{Z}^* \mathbf{W}^* (\mathbf{Z}' \mathbf{y}^* - \mathbf{Z}' \hat{\mathbf{e}}))$$

where  $\mathbf{W}^*$  is the bootstrap version of  $\mathbf{W}$  and  $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}_{\text{gmm}}$ . For efficient GMM,

$$\mathbf{W}^* = \left( \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i^* \mathbf{z}_i^{*\prime} (\mathbf{y}_i - \mathbf{x}_i^{*\prime} \tilde{\boldsymbol{\beta}}^*)^2 \right)^{-1}$$

for preliminary estimator  $\tilde{\boldsymbol{\beta}}^*$ .

For nonlinear GMM (Section 13.25), the bootstrap criterion function is modified. The recentered bootstrap criterion is

$$\begin{aligned} J^{**}(\boldsymbol{\beta}) &= n \left( \bar{\mathbf{g}}_n^*(\boldsymbol{\beta}) - \bar{\mathbf{g}}_n(\hat{\boldsymbol{\beta}}_{\text{gmm}}) \right)' \mathbf{W}^* \left( \bar{\mathbf{g}}_n^*(\boldsymbol{\beta}) - \bar{\mathbf{g}}_n(\hat{\boldsymbol{\beta}}_{\text{gmm}}) \right) \\ \bar{\mathbf{g}}_n^*(\boldsymbol{\beta}) &= \frac{1}{n} \sum_{i=1}^n \mathbf{g}_i^*(\boldsymbol{\beta}) \end{aligned}$$

where  $\bar{\mathbf{g}}_n(\hat{\boldsymbol{\beta}}_{\text{gmm}})$  is from the sample, not from the bootstrap data. The bootstrap estimator is

$$\hat{\boldsymbol{\beta}}_{\text{gmm}}^{**} = \operatorname{argmin}_{\boldsymbol{\beta}} J^{**}(\boldsymbol{\beta}).$$

The bootstrap can also be used to calculate the p-value of the GMM overidentification test. For the GMM estimator with an efficient weight matrix the standard overidentification test is the Hansen  $J$  statistic

$$J = n \bar{\mathbf{g}}_n(\hat{\boldsymbol{\beta}}_{\text{gmm}})' \hat{\boldsymbol{\Omega}}^{-1} \bar{\mathbf{g}}_n(\hat{\boldsymbol{\beta}}_{\text{gmm}}).$$

The recentered bootstrap analog is

$$J^{**} = n \left( \bar{\mathbf{g}}_n^*(\hat{\boldsymbol{\beta}}_{\text{gmm}}) - \bar{\mathbf{g}}_n(\hat{\boldsymbol{\beta}}_{\text{gmm}}) \right)' \hat{\boldsymbol{\Omega}}^{*-1} \left( \bar{\mathbf{g}}_n^*(\hat{\boldsymbol{\beta}}_{\text{gmm}}) - \bar{\mathbf{g}}_n(\hat{\boldsymbol{\beta}}_{\text{gmm}}) \right).$$

On each bootstrap sample  $J^{**}(b)$  is calculated and stored. The bootstrap p-value is

$$p^* = \frac{1}{B} \sum_{b=1}^B \mathbb{1}(J^{**}(b) > S).$$

This bootstrap p-value is asymptotically valid since the statistic  $J^{**}$  satisfies the overidentified moment conditions.

### 13.27 Conditional Moment Equation Models

In many contexts, an economic model implies more than an unconditional moment restriction of the form  $\mathbb{E}(\mathbf{g}(\mathbf{w}_i, \boldsymbol{\beta})) = \mathbf{0}$ . It implies a conditional moment restriction of the form

$$\mathbb{E}(\mathbf{e}_i(\boldsymbol{\beta}) | \mathbf{z}_i) = \mathbf{0}$$

where  $\mathbf{e}_i(\boldsymbol{\beta})$  is some  $s \times 1$  function of the observation and the parameters. In many cases,  $s = 1$ . The variable  $\mathbf{z}_i$  is often called an **instrument**.

It turns out that this conditional moment restriction is much more powerful, and restrictive, than the unconditional moment equation model discussed throughout this chapter.

For example, the linear model  $y_i = \mathbf{x}'_i \boldsymbol{\beta} + e_i$  with instruments  $\mathbf{z}_i$  falls into this class under the assumption  $\mathbb{E}(e_i | \mathbf{z}_i) = 0$ . In this case,  $e_i(\boldsymbol{\beta}) = y_i - \mathbf{x}'_i \boldsymbol{\beta}$ .

It is also helpful to realize that conventional regression models also fall into this class, except that in this case  $\mathbf{x}_i = \mathbf{z}_i$ . For example, in linear regression,  $e_i(\boldsymbol{\beta}) = y_i - \mathbf{x}'_i \boldsymbol{\beta}$ , while in a nonlinear regression model  $e_i(\boldsymbol{\beta}) = y_i - \mathbf{g}(\mathbf{x}_i, \boldsymbol{\beta})$ . In a joint model of the conditional mean  $\mathbb{E}(y | \mathbf{x}) = \mathbf{x}' \boldsymbol{\beta}$  and variance  $\text{var}(y | \mathbf{x}) = f(\mathbf{x})' \boldsymbol{\gamma}$ , then

$$\mathbf{e}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \begin{cases} y_i - \mathbf{x}'_i \boldsymbol{\beta} \\ (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 - f(\mathbf{x}_i)' \boldsymbol{\gamma} \end{cases}.$$

Here  $s = 2$ .

Given a conditional moment restriction, an unconditional moment restriction can always be constructed. That is for any  $\ell \times 1$  function  $\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\beta})$ , we can set  $\mathbf{g}_i(\boldsymbol{\beta}) = \boldsymbol{\phi}(\mathbf{z}_i, \boldsymbol{\beta}) e_i(\boldsymbol{\beta})$  which satisfies  $\mathbb{E}(\mathbf{g}_i(\boldsymbol{\beta})) = \mathbf{0}$  and hence defines an unconditional moment equation model. The obvious problem is that the class of functions  $\boldsymbol{\phi}$  is infinite. Which should be selected?

This is equivalent to the problem of selection of the best instruments. If  $z_i \in \mathbb{R}$  is a valid instrument satisfying  $\mathbb{E}(e_i | z_i) = 0$ , then  $z_i, z_i^2, z_i^3, \dots$ , etc., are all valid instruments. Which should be used?

One solution is to construct an infinite list of potent instruments, and then use the first  $k$  instruments. How is  $k$  to be determined? This is an area of theory still under development. A recent study of this problem is Donald and Newey (2001).

Another approach is to construct the **optimal instrument**. The form was uncovered by Chamberlain (1987). Take the case  $s = 1$ . Let

$$\mathbf{R}_i = \mathbb{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}} e_i(\boldsymbol{\beta}) | \mathbf{z}_i\right)$$

and

$$\sigma_i^2 = \mathbb{E}(e_i(\boldsymbol{\beta})^2 | \mathbf{z}_i).$$

Then the “optimal instrument” is

$$\mathbf{A}_i = -\sigma_i^{-2} \mathbf{R}_i$$

so the optimal moment is

$$\mathbf{g}_i(\boldsymbol{\beta}) = \mathbf{A}_i e_i(\boldsymbol{\beta}).$$

Setting  $\mathbf{g}_i(\boldsymbol{\beta})$  to be this choice (which is  $k \times 1$ , so is just-identified) yields the best GMM estimator possible.

In practice,  $\mathbf{A}_i$  is unknown, but its form does help us think about construction of optimal instruments.

In the linear model  $e_i(\boldsymbol{\beta}) = y_i - \mathbf{x}'_i \boldsymbol{\beta}$ , note that

$$\mathbf{R}_i = -\mathbb{E}(\mathbf{x}_i | \mathbf{z}_i)$$

and

$$\sigma_i^2 = \mathbb{E}(e_i^2 | \mathbf{z}_i),$$

so

$$\mathbf{A}_i = \sigma_i^{-2} \mathbb{E}(\mathbf{x}_i | \mathbf{z}_i).$$

In the case of linear regression,  $\mathbf{x}_i = \mathbf{z}_i$ , so  $\mathbf{A}_i = \sigma_i^{-2} \mathbf{z}_i$ . Hence efficient GMM is equivalently to optimal GLS.

In the case of endogenous variables, note that the efficient instrument  $\mathbf{A}_i$  involves the estimation of the conditional mean of  $\mathbf{x}_i$  given  $\mathbf{z}_i$ . In other words, to get the best instrument for  $\mathbf{x}_i$ , we need the best conditional mean model for  $\mathbf{x}_i$  given  $\mathbf{z}_i$ , not just an arbitrary linear projection. The efficient instrument is also inversely proportional to the conditional variance of  $e_i$ . This is the same as the GLS estimator; namely that improved efficiency can be obtained if the observations are weighted inversely to the conditional variance of the errors.

### 13.28 Technical Proofs\*

**Proof of Theorem 13.12.** Set  $\tilde{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{cgmm}}$  and  $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{gmm}}$ . By standard covariance matrix analysis  $\hat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$  and  $\tilde{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$ . Thus we can replace  $\hat{\boldsymbol{\Omega}}$  and  $\tilde{\boldsymbol{\Omega}}$  in the criteria without affecting the asymptotic distribution. In particular

$$\begin{aligned} \tilde{J}(\hat{\boldsymbol{\beta}}_{\text{cgmm}}) &= \frac{1}{n} \tilde{\mathbf{e}}' \mathbf{Z} \tilde{\boldsymbol{\Omega}}^{-1} \mathbf{Z}' \tilde{\mathbf{e}} \\ &= \frac{1}{n} \tilde{\mathbf{e}}' \mathbf{Z} \hat{\boldsymbol{\Omega}}^{-1} \mathbf{Z}' \tilde{\mathbf{e}} + o_p(1). \end{aligned} \quad (13.25)$$

Now observe that

$$\mathbf{Z}' \tilde{\mathbf{e}} = \mathbf{Z}' \hat{\mathbf{e}} - \mathbf{Z}' \mathbf{X} (\hat{\boldsymbol{\beta}}_{\text{cgmm}} - \hat{\boldsymbol{\beta}}_{\text{gmm}}).$$

Thus

$$\begin{aligned} \frac{1}{n} \tilde{\mathbf{e}}' \mathbf{Z} \hat{\boldsymbol{\Omega}}^{-1} \mathbf{Z}' \tilde{\mathbf{e}} &= \frac{1}{n} \hat{\mathbf{e}}' \mathbf{Z} \hat{\boldsymbol{\Omega}}^{-1} \mathbf{Z}' \hat{\mathbf{e}} - \frac{2}{n} (\hat{\boldsymbol{\beta}}_{\text{cgmm}} - \hat{\boldsymbol{\beta}}_{\text{gmm}})' \mathbf{X}' \mathbf{Z} \hat{\boldsymbol{\Omega}}^{-1} \mathbf{Z}' \hat{\mathbf{e}} \\ &\quad + \frac{1}{n} (\hat{\boldsymbol{\beta}}_{\text{cgmm}} - \hat{\boldsymbol{\beta}}_{\text{gmm}})' \mathbf{X}' \mathbf{Z} \hat{\boldsymbol{\Omega}}^{-1} \mathbf{Z}' \mathbf{X} (\hat{\boldsymbol{\beta}}_{\text{cgmm}} - \hat{\boldsymbol{\beta}}_{\text{gmm}}) \\ &= \hat{J}(\hat{\boldsymbol{\beta}}_{\text{gmm}}) + \frac{1}{n} (\hat{\boldsymbol{\beta}}_{\text{cgmm}} - \hat{\boldsymbol{\beta}}_{\text{gmm}})' \mathbf{X}' \mathbf{Z} \hat{\boldsymbol{\Omega}}^{-1} \mathbf{Z}' \mathbf{X} (\hat{\boldsymbol{\beta}}_{\text{cgmm}} - \hat{\boldsymbol{\beta}}_{\text{gmm}}) \end{aligned} \quad (13.26)$$

where the second equality holds since  $\mathbf{X}' \mathbf{Z} \hat{\boldsymbol{\Omega}}^{-1} \mathbf{Z}' \hat{\mathbf{e}} = \mathbf{0}$  is the first-order condition for  $\hat{\boldsymbol{\beta}}_{\text{gmm}}$ . By (13.16) and Theorem 13.4, under  $\mathbb{H}_0$

$$\begin{aligned} \sqrt{n} (\hat{\boldsymbol{\beta}}_{\text{cgmm}} - \hat{\boldsymbol{\beta}}_{\text{gmm}}) &= -(\mathbf{X}' \mathbf{Z} \hat{\boldsymbol{\Omega}}^{-1} \mathbf{Z}' \mathbf{X})^{-1} \mathbf{R} \left( \mathbf{R}' (\mathbf{X}' \mathbf{Z} \hat{\boldsymbol{\Omega}}^{-1} \mathbf{Z}' \mathbf{X})^{-1} \mathbf{R} \right)^{-1} \mathbf{R}' \sqrt{n} (\hat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta}) + o_p(1) \\ &\xrightarrow{d} (\mathbf{Q}' \boldsymbol{\Omega}^{-1} \mathbf{Q})^{-1} \mathbf{R} \mathbf{Z} \end{aligned} \quad (13.27)$$

where

$$\begin{aligned} Z &\sim N(\mathbf{0}, V_R) \\ V_R &= \left( R V' (Q' \Omega^{-1} Q)^{-1} R \right)^{-1}. \end{aligned} \quad (13.28)$$

Putting together (13.25), (13.26), (13.27) and (13.28),

$$\begin{aligned} D &= \tilde{J}(\hat{\beta}_{\text{cgmm}}) - \hat{J}(\hat{\beta}_{\text{gmm}}) \\ &= \sqrt{n} (\hat{\beta}_{\text{cgmm}} - \hat{\beta}_{\text{gmm}})' \frac{1}{n} X' Z \hat{\Omega}^{-1} \frac{1}{n} Z' X \sqrt{n} (\hat{\beta}_{\text{cgmm}} - \hat{\beta}_{\text{gmm}}) \\ &\xrightarrow{d} Z' V_R^{-1} Z \\ &\sim \chi_q^2 \end{aligned}$$

since  $V_R > 0$  and  $Z$  is  $q \times 1$ . ■

**Proof of Theorem 13.15.** Let  $\tilde{\beta}$  denote the GMM estimate obtained with the instrument set  $z_{ai}$  and let  $\hat{\beta}$  denote the GMM estimates obtained with the instrument set  $z_i$ . Set

$$\begin{aligned} \tilde{e} &= y - X \tilde{\beta} \\ \hat{e} &= y - X \hat{\beta} \\ \tilde{\Omega} &= n^{-1} \sum_{i=1}^n z_{ai} z_{ai}' \tilde{e}_i^2 \\ \hat{\Omega} &= n^{-1} \sum_{i=1}^n z_i z_i' \tilde{e}_i^2. \end{aligned}$$

Let  $R$  be the  $\ell \times \ell_a$  selector matrix so that  $z_{ai} = R' z_i$ . Note that

$$\tilde{\Omega} = R' n^{-1} \sum_{i=1}^n z_i z_i' \tilde{e}_i^2 R.$$

By standard covariance matrix analysis,  $\hat{\Omega} \xrightarrow{p} \Omega$  and  $\tilde{\Omega} \xrightarrow{p} R' \Omega R$ . Also,  $\frac{1}{n} Z' X \xrightarrow{p} Q$ , say. By the CLT,  $n^{-1/2} Z' e \xrightarrow{d} Z$  where  $Z \sim N(\mathbf{0}, \Omega)$ . Then

$$\begin{aligned} n^{-1/2} Z' \hat{e} &= \left( I_\ell - \left( \frac{1}{n} Z' X \right) \left( \frac{1}{n} X' Z \hat{\Omega}^{-1} \frac{1}{n} Z' X \right)^{-1} \left( \frac{1}{n} X' Z \right) \hat{\Omega}^{-1} \right) n^{-1/2} Z' e \\ &\xrightarrow{d} \left( I_\ell - Q (Q' \Omega^{-1} Q)^{-1} Q' \Omega^{-1} \right) Z \end{aligned}$$

and

$$\begin{aligned} n^{-1/2} Z'_a \tilde{e} &= R' \left( I_\ell - \left( \frac{1}{n} Z' X \right) \left( \frac{1}{n} X' Z R \tilde{\Omega}^{-1} R' \frac{1}{n} Z' X \right)^{-1} \left( \frac{1}{n} X' Z \right) R \tilde{\Omega}^{-1} R' \right) n^{-1/2} Z' e \\ &\xrightarrow{d} R' \left( I_\ell - Q \left( Q' R (R' \Omega R)^{-1} R' Q \right)^{-1} Q' R (R' \Omega R)^{-1} R' \right) Z \end{aligned}$$

jointly.

By linear rotations of  $Z$  and  $R$  we can set  $\Omega = I_\ell$  to simplify the notation. Thus setting  $P_Q = Q (Q' Q)^{-1} Q'$ ,  $P_R = R (R' R)^{-1} R'$  and  $Z \sim N(\mathbf{0}, I_\ell)$  we have

$$\hat{J} \xrightarrow{d} Z' (I_\ell - P_Q) Z$$

and

$$\tilde{J} \xrightarrow{d} Z' \left( P_R - P_R Q (Q' P_R Q)^{-1} Q' P_R \right) Z.$$

It follows that

$$C = \widehat{J} - \widetilde{J} \xrightarrow{d} Z' A Z$$

where

$$A = \left( I_\ell - P_Q - P_R + P_R Q (Q' P_R Q)^{-1} Q' P_R \right).$$

This is a quadratic form in a standard normal vector, and the matrix  $A$  is idempotent (this is straightforward to check).  $Z' A Z$  is thus distributed as  $\chi_d^2$  with degrees of freedom  $d$  equal to the rank of  $A$ . This is

$$\begin{aligned} \text{rank}(A) &= \text{tr} \left( I_\ell - P_Q - P_R + P_R Q (Q' P_R Q)^{-1} Q' P_R \right) \\ &= \ell - k - \ell_a + k \\ &= \ell_b. \end{aligned}$$

Thus the asymptotic distribution of  $C$  is  $\chi_{\ell_b}^2$  as claimed. ■

## Exercises

**Exercise 13.1** Take the model

$$\begin{aligned}y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) &= \mathbf{0} \\ e_i^2 &= \mathbf{z}'_i \boldsymbol{\gamma} + \eta_i \\ \mathbb{E}(\mathbf{z}_i \eta_i) &= \mathbf{0}.\end{aligned}$$

Find the method of moments estimators  $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$  for  $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ .

**Exercise 13.2** Take the single equation

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \\ \mathbb{E}(\mathbf{e} | \mathbf{Z}) &= \mathbf{0}\end{aligned}$$

Assume  $\mathbb{E}(e_i^2 | \mathbf{z}_i) = \sigma^2$ . Show that if  $\hat{\boldsymbol{\beta}}_{\text{gmm}}$  is the GMM estimated by GMM with weight matrix  $\mathbf{W}_n = (\mathbf{Z}' \mathbf{Z})^{-1}$ , then

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \sigma^2 (\mathbf{Q}' \mathbf{M}^{-1} \mathbf{Q})^{-1})$$

where  $\mathbf{Q} = \mathbb{E}(\mathbf{z}_i \mathbf{x}'_i)$  and  $\mathbf{M} = \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i)$ .

**Exercise 13.3** Take the model  $y_i = \mathbf{x}'_i \boldsymbol{\beta} + e_i$  with  $\mathbb{E}(\mathbf{z}_i e_i) = \mathbf{0}$ . Let  $\tilde{e}_i = y_i - \mathbf{x}'_i \tilde{\boldsymbol{\beta}}$  where  $\tilde{\boldsymbol{\beta}}$  is consistent for  $\boldsymbol{\beta}$  (e.g. a GMM estimator with arbitrary weight matrix). Define an estimate of the optimal GMM weight matrix

$$\widehat{\mathbf{W}} = \left( \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{z}'_i \tilde{e}_i^2 \right)^{-1}.$$

Show that  $\widehat{\mathbf{W}} \xrightarrow{P} \boldsymbol{\Omega}^{-1}$  where  $\boldsymbol{\Omega} = \mathbb{E}(\mathbf{z}_i \mathbf{z}'_i e_i^2)$ .

**Exercise 13.4** In the linear model estimated by GMM with general weight matrix  $\mathbf{W}$ , the asymptotic variance of  $\hat{\boldsymbol{\beta}}_{GMM}$  is

$$\mathbf{V} = (\mathbf{Q}' \mathbf{W} \mathbf{Q})^{-1} \mathbf{Q}' \mathbf{W} \boldsymbol{\Omega} \mathbf{W} \mathbf{Q} (\mathbf{Q}' \mathbf{W} \mathbf{Q})^{-1}.$$

- (a) Let  $\mathbf{V}_0$  be this matrix when  $\mathbf{W} = \boldsymbol{\Omega}^{-1}$ . Show that  $\mathbf{V}_0 = (\mathbf{Q}' \boldsymbol{\Omega}^{-1} \mathbf{Q})^{-1}$ .
- (b) We want to show that for any  $\mathbf{W}$ ,  $\mathbf{V} - \mathbf{V}_0$  is positive semi-definite (for then  $\mathbf{V}_0$  is the smaller possible covariance matrix and  $\mathbf{W} = \boldsymbol{\Omega}^{-1}$  is the efficient weight matrix). To do this, start by finding matrices  $\mathbf{A}$  and  $\mathbf{B}$  such that  $\mathbf{V} = \mathbf{A}' \boldsymbol{\Omega} \mathbf{A}$  and  $\mathbf{V}_0 = \mathbf{B}' \boldsymbol{\Omega} \mathbf{B}$ .
- (c) Show that  $\mathbf{B}' \boldsymbol{\Omega} \mathbf{A} = \mathbf{B}' \boldsymbol{\Omega} \mathbf{B}$  and therefore that  $\mathbf{B}' \boldsymbol{\Omega} (\mathbf{A} - \mathbf{B}) = \mathbf{0}$ .
- (d) Use the expressions  $\mathbf{V} = \mathbf{A}' \boldsymbol{\Omega} \mathbf{A}$ ,  $\mathbf{A} = \mathbf{B} + (\mathbf{A} - \mathbf{B})$ , and  $\mathbf{B}' \boldsymbol{\Omega} (\mathbf{A} - \mathbf{B}) = \mathbf{0}$  to show that  $\mathbf{V} \geq \mathbf{V}_0$ .

**Exercise 13.5** Prove Theorem 13.8.

**Exercise 13.6** Derive the constrained GMM estimator (13.16).

**Exercise 13.7** Show that the constrained GMM estimator (13.16) with the efficient weight matrix is (13.19).

**Exercise 13.8** Prove Theorem 13.9.

**Exercise 13.9** Prove Theorem 13.10.

**Exercise 13.10** The equation of interest is

$$\begin{aligned} y_i &= \mathbf{m}(\mathbf{x}_i, \boldsymbol{\beta}) + e_i \\ \mathbb{E}(\mathbf{z}_i e_i) &= \mathbf{0}. \end{aligned}$$

The observed data is  $(y_i, \mathbf{z}_i, \mathbf{x}_i)$ .  $\mathbf{z}_i$  is  $\ell \times 1$  and  $\boldsymbol{\beta}$  is  $k \times 1$ ,  $\ell \geq k$ . Show how to construct an efficient GMM estimator for  $\boldsymbol{\beta}$ .

**Exercise 13.11** As a continuation of Exercise 12.7, derive the efficient GMM estimator using the instrument  $\mathbf{z}_i = (x_i \ x_i^2)'$ . Does this differ from 2SLS and/or OLS?

**Exercise 13.12** In the linear model  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$  with  $\mathbb{E}(\mathbf{x}_i e_i) = \mathbf{0}$ , a Generalized Method of Moments (GMM) criterion function for  $\boldsymbol{\beta}$  is defined as

$$J(\boldsymbol{\beta}) = \frac{1}{n} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{X} \widehat{\boldsymbol{\Omega}}^{-1} \mathbf{X}' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (13.29)$$

where  $\widehat{\boldsymbol{\Omega}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \widehat{e}_i^2$ ,  $\widehat{e}_i = y_i - \mathbf{x}_i' \widehat{\boldsymbol{\beta}}$  are the OLS residuals, and  $\widehat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}$  is least-squares. The GMM estimator of  $\boldsymbol{\beta}$ , subject to the restriction  $\mathbf{r}(\boldsymbol{\beta}) = \mathbf{0}$ , is defined as

$$\widetilde{\boldsymbol{\beta}} = \underset{\mathbf{r}(\boldsymbol{\beta}) = \mathbf{0}}{\operatorname{argmin}} J_n(\boldsymbol{\beta}).$$

The GMM test statistic (the distance statistic) of the hypothesis  $\mathbf{r}(\boldsymbol{\beta}) = \mathbf{0}$  is

$$D = J(\widetilde{\boldsymbol{\beta}}) = \min_{\mathbf{r}(\boldsymbol{\beta}) = \mathbf{0}} J(\boldsymbol{\beta}). \quad (13.30)$$

(a) Show that you can rewrite  $J(\boldsymbol{\beta})$  in (13.29) as

$$J(\boldsymbol{\beta}) = n(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})' \widehat{V}_{\boldsymbol{\beta}}^{-1} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})$$

and thus  $\widetilde{\boldsymbol{\beta}}$  is the same as the minimum distance estimator.

(b) Show that under linear hypotheses the distance statistic  $D$  in (13.30) equals the Wald statistic.

**Exercise 13.13** Take the linear model

$$\begin{aligned} y_i &= \mathbf{x}_i' \boldsymbol{\beta} + e_i \\ \mathbb{E}(\mathbf{z}_i e_i) &= \mathbf{0}. \end{aligned}$$

and consider the GMM estimator  $\widehat{\boldsymbol{\beta}}$  of  $\boldsymbol{\beta}$ . Let

$$J = n \bar{\mathbf{g}}_n(\widehat{\boldsymbol{\beta}})' \widehat{\boldsymbol{\Omega}}^{-1} \bar{\mathbf{g}}_n(\widehat{\boldsymbol{\beta}})$$

denote the test of overidentifying restrictions. Show that  $J \xrightarrow{d} \chi_{\ell-k}^2$  as  $n \rightarrow \infty$  by demonstrating each of the following:

(a) Since  $\boldsymbol{\Omega} > 0$ , we can write  $\boldsymbol{\Omega}^{-1} = \mathbf{C}\mathbf{C}'$  and  $\boldsymbol{\Omega} = \mathbf{C}'^{-1}\mathbf{C}^{-1}$ .

(b)  $J = n(\mathbf{C}' \bar{\mathbf{g}}_n(\widehat{\boldsymbol{\beta}}))' (\mathbf{C}' \widehat{\boldsymbol{\Omega}} \mathbf{C})^{-1} \mathbf{C}' \bar{\mathbf{g}}_n(\widehat{\boldsymbol{\beta}})$ .

(c)  $\mathbf{C}' \bar{\mathbf{g}}_n(\widehat{\boldsymbol{\beta}}) = \mathbf{D}_n \mathbf{C}' \bar{\mathbf{g}}_n(\boldsymbol{\beta})$  where

$$\begin{aligned} \mathbf{D}_n &= \mathbf{I}_{\ell} - \mathbf{C}' \left( \frac{1}{n} \mathbf{Z}' \mathbf{X} \right) \left( \left( \frac{1}{n} \mathbf{X}' \mathbf{Z} \right) \widehat{\boldsymbol{\Omega}}^{-1} \left( \frac{1}{n} \mathbf{Z}' \mathbf{X} \right) \right)^{-1} \left( \frac{1}{n} \mathbf{X}' \mathbf{Z} \right) \widehat{\boldsymbol{\Omega}}^{-1} \mathbf{C}'^{-1} \\ \bar{\mathbf{g}}_n(\boldsymbol{\beta}) &= \frac{1}{n} \mathbf{Z}' \mathbf{e}. \end{aligned}$$

(d)  $\mathbf{D}_n \xrightarrow{p} \mathbf{I}_\ell - \mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'$  where  $\mathbf{R} = \mathbf{C}'\mathbb{E}(\mathbf{z}_i\mathbf{x}_i')$ .

(e)  $n^{1/2}\mathbf{C}'\bar{\mathbf{g}}_n(\boldsymbol{\beta}) \xrightarrow{d} \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_\ell)$ .

(f)  $J \xrightarrow{d} \mathbf{u}'(\mathbf{I}_\ell - \mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}')\mathbf{u}$ .

(g)  $\mathbf{u}'(\mathbf{I}_\ell - \mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}')\mathbf{u} \sim \chi_{\ell-k}^2$ .

Hint:  $\mathbf{I}_\ell - \mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'$  is a projection matrix.

**Exercise 13.14** Take the model

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}(\mathbf{z}_i e_i) &= \mathbf{0} \end{aligned}$$

$y_i$  scalar,  $\mathbf{x}_i$  a  $k$  vector and  $\mathbf{z}_i$  an  $\ell$  vector,  $\ell \geq k$ . Assume i.i.d. observations. Consider the statistic

$$\begin{aligned} J_n(\boldsymbol{\beta}) &= n\bar{\mathbf{m}}_n(\boldsymbol{\beta})'\mathbf{W}\bar{\mathbf{m}}_n(\boldsymbol{\beta}) \\ \bar{\mathbf{m}}_n(\boldsymbol{\beta}) &= \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}) \end{aligned}$$

for some weight matrix  $\mathbf{W} > 0$ .

(a) Take the hypothesis

$$\mathbb{H}_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$$

Derive the asymptotic distribution of  $J_n(\boldsymbol{\beta}_0)$  under  $\mathbb{H}_0$  as  $n \rightarrow \infty$ .

- (b) What choice for  $\mathbf{W}$  yields a known asymptotic distribution in part (a)? (Be specific about degrees of freedom.)
- (c) Write down an appropriate estimator  $\widehat{\mathbf{W}}$  for  $\mathbf{W}$  which takes advantage of  $\mathbb{H}_0$ . (You do not need to demonstrate consistency or unbiasedness.)
- (d) Describe an asymptotic test of  $\mathbb{H}_0$  against  $\mathbb{H}_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$  based on this statistic.
- (e) Use the result in part (d) to construct a confidence region for  $\boldsymbol{\beta}$ . What can you say about the form of this region? For example, does the confidence region take the form of an ellipse, similar to conventional confidence regions?

**Exercise 13.15** Consider the model

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}(\mathbf{z}_i e_i) &= \mathbf{0} \end{aligned} \tag{13.31}$$

$$\mathbf{R}'\boldsymbol{\beta} = \mathbf{0} \tag{13.32}$$

with  $y_i$  scalar,  $\mathbf{x}_i$  a  $k$  vector and  $\mathbf{z}_i$  an  $\ell$  vector with  $\ell > k$ . The matrix  $\mathbf{R}$  is  $k \times q$  with  $1 \leq q < k$ . You have a random sample  $(y_i, \mathbf{x}_i, \mathbf{z}_i : i = 1, \dots, n)$ .

For simplicity, assume the efficient weight matrix  $\mathbf{W} = (\mathbb{E}(\mathbf{z}_i \mathbf{z}_i' e_i^2))^{-1}$  is known.

- (a) Write out the GMM estimator  $\widehat{\boldsymbol{\beta}}$  of  $\boldsymbol{\beta}$  given the moment conditions (13.31) but ignoring constraint (13.32).
- (b) Write out the GMM estimator  $\widetilde{\boldsymbol{\beta}}$  of  $\boldsymbol{\beta}$  given the moment conditions (13.31) and constraint (13.32).

- (c) Find the asymptotic distribution of  $\sqrt{n}(\tilde{\beta} - \beta)$  as  $n \rightarrow \infty$  under the assumption that (13.31) and (13.32) are correct.

**Exercise 13.16** The observed data is  $\{y_i, x_i, z_i\} \in \mathbb{R} \times \mathbb{R}^k \times \mathbb{R}^\ell$ ,  $k > 1$  and  $\ell > k > 1$ ,  $i = 1, \dots, n$ . The model is

$$\begin{aligned} y_i &= \mathbf{x}'_i \beta + e_i \\ \mathbb{E}(z_i e_i) &= 0. \end{aligned} \tag{13.33}$$

- (a) Given a weight matrix  $W > 0$ , write down the GMM estimator  $\hat{\beta}$  for  $\beta$ .

- (b) Suppose the model is misspecified in that

$$\begin{aligned} e_i &= \delta n^{-1/2} + u_i \\ \mathbb{E}(u_i | z_i) &= 0 \end{aligned} \tag{13.34}$$

with  $\mu_z = \mathbb{E}(z_i) \neq 0$  and  $\delta \neq 0$ . Show that (13.34) implies (13.33) is false.

- (c) Express  $\sqrt{n}(\hat{\beta} - \beta)$  as a function of  $W$ ,  $n$ ,  $\delta$ , and the variables  $(x_i, z_i, u_i)$ .

- (d) Find the asymptotic distribution of  $\sqrt{n}(\hat{\beta} - \beta)$  under Assumption (13.34).

**Exercise 13.17** The model is

$$\begin{aligned} y_i &= z_i \beta + x_i \gamma + e_i \\ \mathbb{E}(e_i | x_i) &= 0. \end{aligned}$$

Thus  $z_i$  is potentially endogenous and  $x_i$  is exogenous. Assume that  $z_i$  and  $x_i$  are scalar. Someone suggests estimating  $(\beta, \gamma)$  by GMM, using the pair  $(x_i, x_i^2)$  as the instruments. Is this feasible? Under what conditions, if any, (in addition to those described above) is this a valid estimator?

**Exercise 13.18** The observations are i.i.d.,  $(y_i, \mathbf{x}_i, \mathbf{q}_i : i = 1, \dots, n)$ , where  $\mathbf{x}_i$  is  $k \times 1$  and  $\mathbf{q}_i$  is  $m \times 1$ . The model is

$$\begin{aligned} y_i &= \mathbf{x}'_i \beta + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) &= 0 \\ \mathbb{E}(\mathbf{q}_i e_i) &= 0. \end{aligned}$$

Find the efficient GMM estimator for  $\beta$ .

**Exercise 13.19** You want to estimate  $\mu = \mathbb{E}(y_i)$  under the assumption that  $\mathbb{E}(x_i) = 0$ , where  $y_i$  and  $x_i$  are scalar and observed from a random sample. Find an efficient GMM estimator for  $\mu$ .

**Exercise 13.20** Consider the model

$$\begin{aligned} y_i &= \mathbf{x}'_i \beta + e_i \\ \mathbb{E}(z_i e_i) &= 0 \\ \mathbf{R}' \beta &= 0. \end{aligned}$$

The dimensions are  $\mathbf{x} \in \mathbb{R}^k$ ,  $\mathbf{z} \in \mathbb{R}^\ell$ ,  $\ell > k$ . The matrix  $\mathbf{R}$  is  $k \times q$ ,  $1 \leq q < k$ . Derive an efficient GMM estimator for  $\beta$  for this model.

**Exercise 13.21** Take the linear equation  $y_i = \mathbf{x}'_i \beta + e$ , and consider the following estimators of  $\beta$ .

1.  $\hat{\beta}$ : 2SLS using the instruments  $z_{1i}$ .

2.  $\tilde{\beta}$ : 2SLS using the instruments  $\mathbf{z}_{1i}$ .
3.  $\bar{\beta}$ : GMM using the instruments  $\mathbf{z}_i = (\mathbf{z}_{1i}, \mathbf{z}_{2i})$  and the weight matrix

$$\mathbf{W} = \begin{pmatrix} (\mathbf{Z}'_1 \mathbf{Z}_1)^{-1} \lambda & 0 \\ 0 & (\mathbf{Z}'_2 \mathbf{Z}_2)^{-1} (1 - \lambda) \end{pmatrix}$$

for  $\lambda \in (0, 1)$ .

Find an expression for  $\bar{\beta}$  which shows that it is a specific weighted average of  $\hat{\beta}$  and  $\tilde{\beta}$ .

**Exercise 13.22** Consider the just-identified model

$$\begin{aligned} y_i &= \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) &= 0 \end{aligned}$$

where  $\mathbf{x}_i = (\mathbf{x}'_{1i} \ \mathbf{x}'_{2i})'$  and  $\mathbf{z}_i$  are  $k \times 1$ . We want to test  $H_0 : \boldsymbol{\beta}_1 = 0$ . Three econometricians are called to advise on how to test  $H_0$ .

- Econometrician 1 proposes testing  $H_0$  by a Wald statistic.
- Econometrician 2 suggests testing  $H_0$  by the GMM Distance Statistic.
- Econometrician 3 suggests testing  $H_0$  using the test of overidentifying restrictions.

You are asked to settle this dispute. Explain the advantages and/or disadvantages of the different procedures, in this specific context.

**Exercise 13.23** Take the model

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) &= \mathbf{0} \\ \boldsymbol{\beta} &= \mathbf{Q} \boldsymbol{\theta} \end{aligned}$$

where  $\boldsymbol{\beta}$  is  $k \times 1$ ,  $\mathbf{Q}$  is  $k \times m$  with  $m < k$ , and  $\mathbf{Q}$  is known. Assume that the observations  $(y_i, \mathbf{x}_i)$  are i.i.d. across  $i = 1, \dots, n$ .

Under these assumptions, what is the efficient estimator of  $\boldsymbol{\theta}$ ?

**Exercise 13.24** Take the model

$$\begin{aligned} y_i &= \theta + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) &= 0 \end{aligned}$$

with  $(y_i, \mathbf{x}_i)$  a random sample.  $y_i$  is real-valued and  $\mathbf{x}_i$  is  $k \times 1$ ,  $k > 1$ .

- (a) Find the efficient GMM estimator of  $\theta$ .
- (b) Is this model over-identified or just-identified?
- (c) Find the GMM test statistic for over-identification.

**Exercise 13.25** Take the model

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(\mathbf{x}_i e_i) &= \mathbf{0} \end{aligned}$$

where  $\mathbf{x}_i$  contains an intercept so  $\mathbb{E}(e_i) = 0$ . An enterprising econometrician notices that this implies the  $n$  moment conditions

$$\mathbb{E}(e_i) = 0, i = 1, \dots, n.$$

Given an  $n \times n$  weight matrix  $\mathbf{W}$ , this implies a GMM criterion

$$J(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

- (a) Under i.i.d. sampling, show that the efficient weight matrix is  $\mathbf{W} = \sigma^{-2} \mathbf{I}_n$ , where  $\sigma^2 = \mathbb{E}(e_i^2)$ .
- (b) Using the weight matrix  $\mathbf{W} = \sigma^{-2} \mathbf{I}_n$ , find the GMM estimator  $\hat{\boldsymbol{\beta}}$  that minimizes  $J(\boldsymbol{\beta})$ .
- (c) Find a simple expression for the minimized criterion  $J(\hat{\boldsymbol{\beta}})$ .
- (d) Theorem 13.14 says that criterion such as  $J(\hat{\boldsymbol{\beta}})$  are asymptotically  $\chi_{\ell-k}^2$  where  $\ell$  is the number of moments. While the assumptions of Theorem 13.14 do not apply to this context, what is  $\ell$  here? That is, which  $\chi^2$  distribution is the asserted asymptotic distribution?
- (e) Does the answer in (d) make sense? Explain your reasoning.

**Exercise 13.26** Take the model

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(e_i | \mathbf{x}_i) &= 0 \\ \mathbb{E}(e_i^2 | \mathbf{x}_i) &= \sigma^2. \end{aligned}$$

An econometrician more enterprising than the one in previous question notices that this implies the  $nk$  moment conditions

$$\mathbb{E}(\mathbf{x}_i e_i) = 0, i = 1, \dots, n.$$

We can write the moments using matrix notation as

$$\mathbb{E}(\bar{\mathbf{X}}' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}))$$

where

$$\bar{\mathbf{X}} = \begin{pmatrix} \mathbf{x}'_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}'_2 & & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}'_n \end{pmatrix}.$$

Given an  $nk \times nk$  weight matrix  $\mathbf{W}$ , this implies a GMM criterion

$$J(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \bar{\mathbf{X}} \mathbf{W} \bar{\mathbf{X}}' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

- (a) Calculate  $\boldsymbol{\Omega} = \mathbb{E}(\bar{\mathbf{X}}' \mathbf{e} \mathbf{e}' \bar{\mathbf{X}})$ .
- (b) The econometrician decides to set  $\mathbf{W} = \boldsymbol{\Omega}^-$ , the Moore-Penrose generalized inverse of  $\boldsymbol{\Omega}$ . (See Section A.6.)

Note: A useful fact is that for a vector  $\mathbf{a}$ ,  $(\mathbf{a}\mathbf{a}')^- = \mathbf{a}\mathbf{a}'(\mathbf{a}'\mathbf{a})^{-2}$ .

- (c) Find the GMM estimator  $\hat{\boldsymbol{\beta}}$  that minimizes  $J(\boldsymbol{\beta})$ .

- (d) Find a simple expression for the minimized criterion  $J(\hat{\beta})$ .
- (e) Comment on whether the  $\chi^2$  approximation from Theorem 13.14 is appropriate for  $J(\hat{\beta})$ .

**Exercise 13.27** Continuation of Exercise 12.23, based on the empirical work reported in Acemoglu, Johnson and Robinson (2001).

- (a) Re-estimate the model estimated part (j) by efficient GMM. Use the 2SLS estimates as the first-step to get the weight matrix, and then calculate the GMM estimator from this weight matrix without further iteration. Report the estimates and standard errors.
- (b) Calculate and report the  $J$  statistic for overidentification.
- (c) Compare the GMM and 2SLS estimates. Discuss your findings

**Exercise 13.28** Continuation of Exercise 12.25, which involved estimation of a wage equation by 2SLS.

- (a) Re-estimate the model in part (a) by efficient GMM. Do the results change meaningfully?
- (b) Re-estimate the model in part (d) by efficient GMM. Do the results change meaningfully?
- (c) Report the  $J$  statistic for overidentification.

**Part IV**

**Dependent and Panel Data**

# Chapter 14

## Time Series

### 14.1 Introduction

A **time series**  $y_t \in \mathbb{R}^m$  is a process observed in sequence over time:  $t = 1, \dots, n$ . To denote the time period it is typical to use the subscript  $t$ . The time series is **univariate** if  $m = 1$  and **multivariate** if  $m > 1$ . This chapter is primarily focused on univariate time series models, though we describe the concepts for the multivariate case when the added generality does not add extra complications.

Most economic time series are recorded at discrete intervals such as annual, quarterly, monthly, weekly, or daily. The number of observations  $s$  per year is called the **frequency**.

Because of the sequential nature of time series, we expect that observations close in calendar time, e.g.  $y_t$  and its **lagged** value  $y_{t-1}$ , will be dependent. This type of dependence structure requires a different distributional theory than for cross-sectional and clustered observations, since we cannot divide the sample into independent groups. Many of the issues which distinguish time series from cross-section econometrics concern the modeling of these dependence relationships.

There are many excellent textbooks for time series analysis. The encyclopedic standard is Hamilton (1994). Others include Harvey (1990), Tong (1990), Brockwell and Davis (1991), Fan and Yao (2003), Lütkepohl (2005), Enders (2014), and Kilian and Lütkepohl (2017). For textbooks on the related subject of forecasting see Granger (1989), Granger and Newbold (1986), and Elliott and Timmermann (2016).

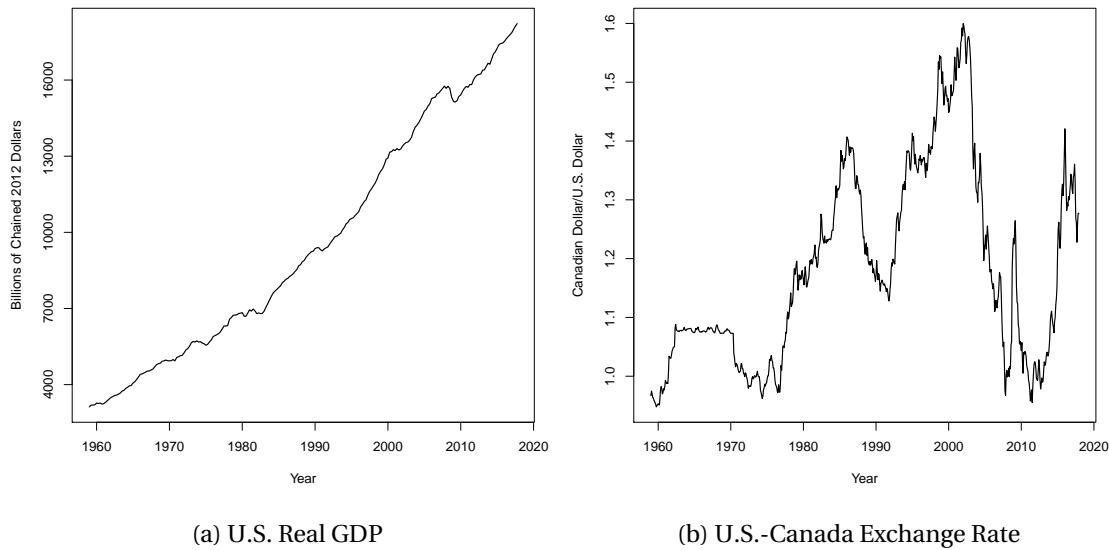


Figure 14.1: U.S. GDP and Exchange Rate

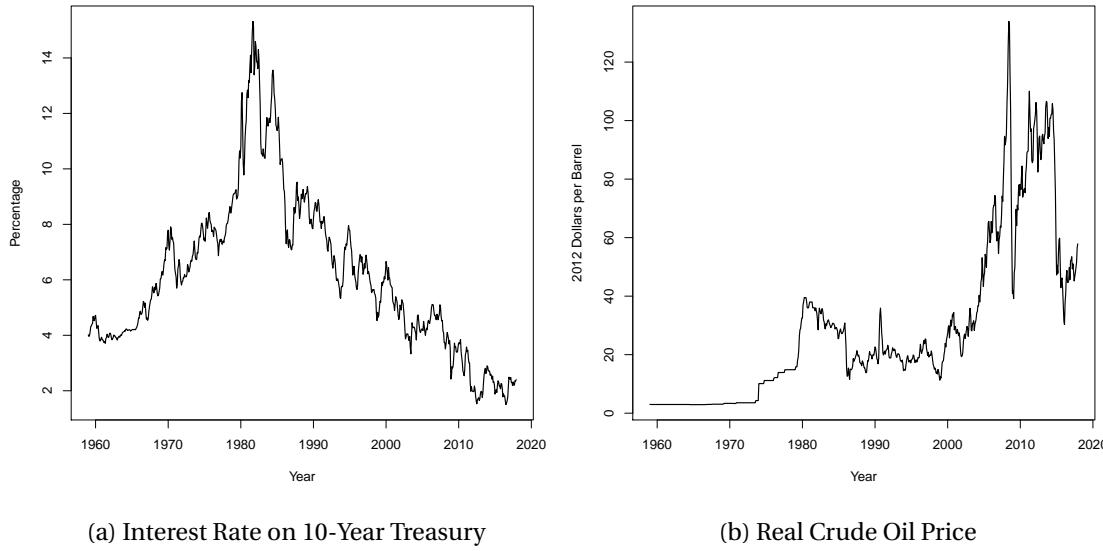


Figure 14.2: Interest Rate and Crude Oil Price

## 14.2 Examples

Many economic time series are macroeconomic variables. An excellent resource for U.S. macroeconomic data are the FRED-MD and FRED-QD databases, which contain a wide set of monthly and quarterly variables, assembled and maintained by the St. Louis Federal Reserve Bank. See McCracken and Ng (2015). The datasets FRED-MD and FRED-QD for 1959-2017 are posted on the course website. FRED-MD has 129 variables over 708 months. FRED-QD has 248 variables over 236 quarters.

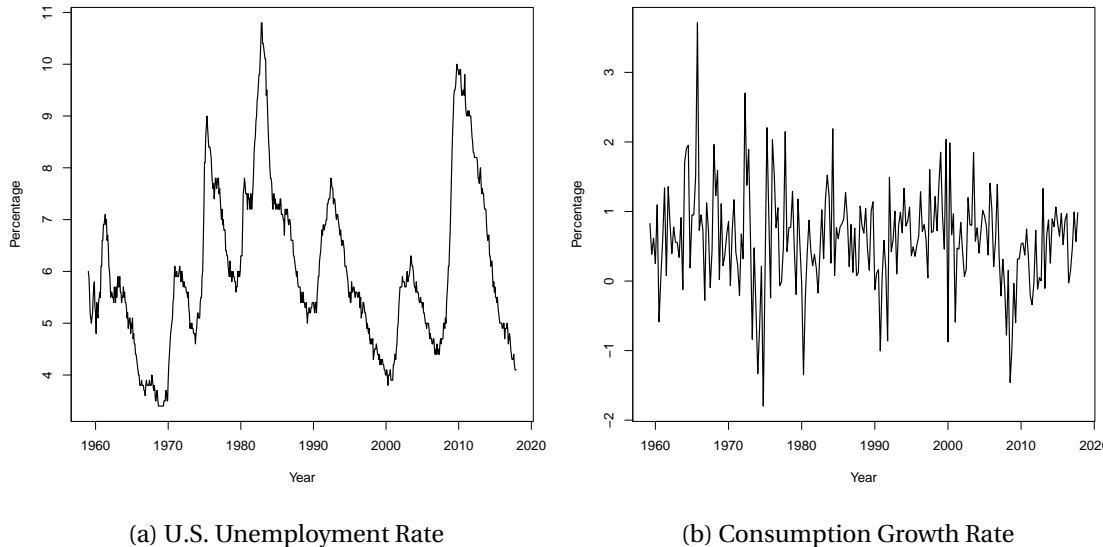


Figure 14.3: Unemployment Rate and Consumption Growth Rate

When working with time series data one of the first tasks is to plot the series against time. In Figures 14.1-14.4 we plot eight example time series from FRED-QD and FRED-MD. As is conventional in time series plots, the x-axis displays calendar dates (in this case years) and the y-axis displays the level of the series. The series plotted are: (1a) Real U.S. GDP (*gdpc1*); (1b) U.S.-Canada exchange rate (*excausx*); (2a)

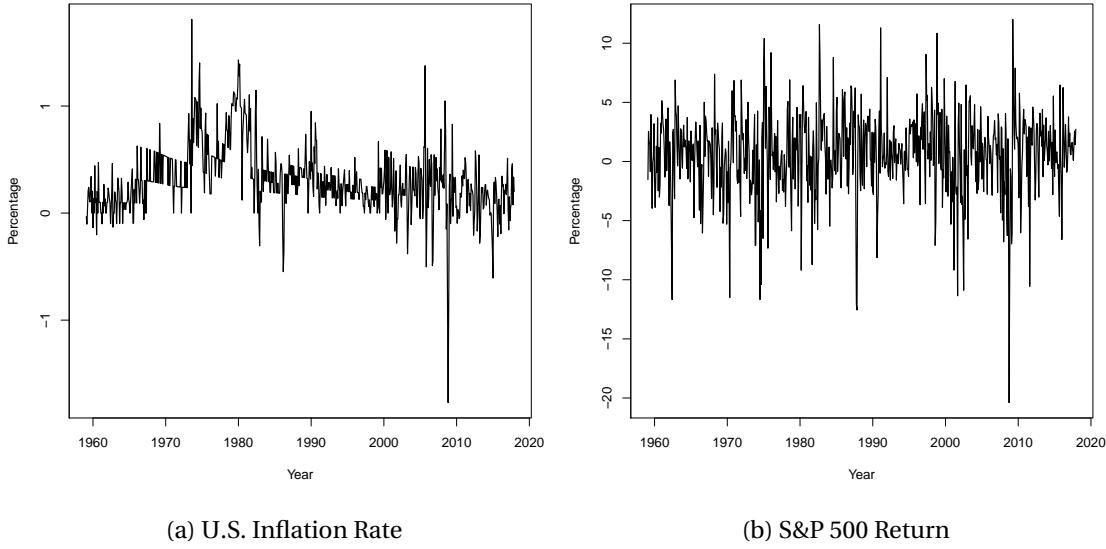


Figure 14.4: U.S. Inflation Rate and S&P 500 Return

Interest rate on U.S. 10-year Treasury (*gs10*); (2b) Real crude oil price (*oilpricex*); (3a) U.S. unemployment rate (*unrate*); (3b) U.S. real non-durables consumption growth rate (growth rate of *pcndx*); (4a) U.S. CPI inflation rate (growth rate of *cpiaucsl*); (4b) S&P 500 return (growth rate of *sp500*). (1a) and (3b) are quarterly series, the rest are monthly.

Many of the plots are smooth, meaning that the neighboring values (in calendar time) are very similar to one another and hence are correlated. Some of the plots are non-smooth, meaning that the neighboring values are not similar and hence less correlated. At least one plot (real GDP) displays a strong upward trend.

## 14.3 Differences and Growth Rates

It is common to transform many series by taking logarithms, differences and/or growth rates. Three of the series in Figures 14.3-14.4 (consumption growth, inflation [growth rate of CPI index], and S&P 500 return) are displayed as growth rates. This transformation may be done for a number of different reasons, but the most credible reason is that this is the suitable variable for the desired analysis.

Many aggregate series such as real GDP are transformed by taking natural logarithms. This flattens the apparent exponential growth, and makes fluctuations proportionate.

The first difference of a series  $y_t$  is

$$\Delta y_t = y_t - y_{t-1}.$$

The second difference is

$$\Delta^2 y_t = \Delta y_t - \Delta y_{t-1}.$$

Higher-order differences can be defined similarly but are not used in practice.

The annual, or year-on-year, change of a series  $y_t$  with frequency  $s$  is

$$\Delta_s y_t = y_t - y_{t-s}.$$

There are several methods to calculate growth rates. The one-period growth rate is the percentage change from period  $t - 1$  to period  $t$ :

$$q_t = 100 \left( \frac{\Delta y_t}{y_{t-1}} \right) = 100 \left( \frac{y_t}{y_{t-1}} - 1 \right). \quad (14.1)$$

The multiplication by 100 is not essential but scales  $q_t$  so that it is a percentage. This is the transformation used for the plots in Figures 14.3(b)-14.4(a)(b).

For non-annual data the one-period growth rate (14.1) may be unappealing for interpretation. Consequently, statistical agencies commonly report “annualized” growth rates, which is the annual growth which would occur if the one-period growth rate is compounded for a full year. For a series with frequency  $s$  the annualized growth rate is

$$a_t = 100 \left( \left( \frac{y_t}{y_{t-1}} \right)^s - 1 \right). \quad (14.2)$$

Notice that  $a_t$  is a nonlinear function of  $q_t$ .

Year-on-year growth rates are

$$Q_t = 100 \left( \frac{\Delta_s y_t}{y_{t-s}} \right) = 100 \left( \frac{y_t}{y_{t-s}} - 1 \right).$$

These do not need annualization.

Growth rates are closely related to logarithmic transformations. For small growth rates,  $q_t$ ,  $a_t$  and  $Q_t$  are approximately first differences in logarithms:

$$\begin{aligned} q_t &\simeq 100 \Delta \log y_t \\ a_t &\simeq 400 \Delta \log y_t \\ Q_t &\simeq 100 \Delta_s \log y_{tt}. \end{aligned}$$

For analysis using growth rates I recommend the one-period growth rates (14.1) or differenced logarithms rather than the annualized growth rates (14.2). While annualized growth rates are preferred for reporting, they are a highly nonlinear transformation which is unnatural for statistical analysis. Differenced logarithms are the most common choice, and are recommended for models which combine log-levels and growth rates for then the models are linear in all variables.

## 14.4 Stationarity

Recall that cross-sectional observations are conventionally treated as random draws from an underlying population. This is not an appropriate model for time series processes due to serial dependence. Instead, we treat the observed sample  $\{y_1, \dots, y_n\}$  as a realization of a dependent stochastic process. It is often useful to view  $\{y_1, \dots, y_n\}$  as a subset of an underlying doubly-infinite sequence  $\{\dots, y_{t-1}, y_t, y_{t+1}, \dots\}$ .

A random vector  $y_t$  can be characterized by its distribution, and a set such as  $(y_t, y_{t+1}, \dots, y_{t+\ell})$  can be characterized by its joint distribution. Important features of these distributions are their means, variances, and covariances. Since there is only one observed time series sample, in order to learn about these distributions there needs to be some sort of constancy. This may only hold after a suitable transformation such as growth rates (as discussed in the previous section).

The most commonly assumed form of constancy is **stationarity**. There are two definitions. The first is sufficient for construction of linear models.

**Definition 14.1**  $\{y_t\}$  is **covariance or weakly stationary** if the mean

$$\boldsymbol{\mu} = \mathbb{E}(y_t)$$

and variance matrix

$$\boldsymbol{\Sigma} = \text{var}(y_t) = \mathbb{E}((y_t - \boldsymbol{\mu})(y_t - \boldsymbol{\mu})')$$

are independent of  $t$ , and the **autocovariances**

$$\boldsymbol{\Gamma}(k) = \text{cov}(y_t, y_{t-k}) = \mathbb{E}((y_t - \boldsymbol{\mu})(y_{t-k} - \boldsymbol{\mu})')$$

are independent of  $t$  for all  $k$ .

In the univariate case we typically write the variance as  $\sigma^2$  and autocovariances as  $\gamma(k)$ .

The mean  $\boldsymbol{\mu}$  and variance  $\boldsymbol{\Sigma}$  are features of the marginal distribution of  $y_t$  (the distribution of  $y_t$  at a specific time period  $t$ ). Their constancy as stated in the above definition means that these features of the distribution are stable over time.

The autocovariances  $\boldsymbol{\Gamma}(k)$  are features of the bivariate distributions of  $(y_t, y_{t-k})$ . Their constancy as stated in the definition means that the correlation patterns between adjacent  $y_t$  are stable over time, and only depend on the number of time periods  $k$  separating the variables. By symmetry, we have  $\boldsymbol{\Gamma}(-k) = \boldsymbol{\Gamma}(k)'$ . In the univariate case this simplifies to  $\gamma(-k) = \gamma(k)$ .

The autocovariances summarize the linear dependence between  $y_t$  and its lags. A scale-free measure of linear dependence in the univariate case are the **autocorrelations**

$$\rho(k) = \text{corr}(y_t, y_{t-k}) = \frac{\text{cov}(y_t, y_{t-k})}{\sqrt{\text{var}(y_t)\text{var}(y_{t-1})}} = \frac{\gamma(k)}{\sigma^2} = \frac{\gamma(k)}{\gamma(0)}.$$

Notice by symmetry that  $\rho(-k) = \rho(k)$ .

The second definition of stationarity concerns the entire joint distribution.

**Definition 14.2**  $\{y_t\}$  is **strictly stationary** if the joint distribution of  $(y_t, \dots, y_{t+\ell})$  is independent of  $t$  for all  $\ell$ .

This is the natural generalization of the cross-section definition of identical distributions. Strict stationarity implies that the (marginal) distribution of  $y_t$  does not vary over time. It also implies that the bivariate distributions of  $(y_t, y_{t+1})$  and multivariate distributions of  $(y_t, \dots, y_{t+\ell})$  are stable over time. Under the assumption of a bounded variance a strictly stationary process is covariance stationary<sup>1</sup>.

For formal statistical theory we will generally require the stronger assumption of strict stationarity. Therefore, if we label a process as “stationary” you should interpret it as meaning “strictly stationary”.

The core meaning of both weak and strict stationarity is the same – that the distribution of  $y_t$  is stable over time. To understand the concept, it may be useful to review the plots in Figures 14.1–14.4. Are these stationary processes? If so, we would expect that the mean and variance would be stable over time. This seems unlikely to apply to the series in Figures 14.1 and 14.2, as in each case it is difficult to describe what is the “typical” value of the series. Stationarity may be appropriate for the series in Figures 14.3 and

<sup>1</sup>More generally, the two classes are non-nested since strictly stationary infinite variance processes are not covariance stationary.

14.4, as each oscillates with a fairly regular pattern. It is difficult, however, to know whether or not a given time series is stationary simply by examining a time series plot.

A straightforward but essential relationship is that an i.i.d. process is strictly stationary.

**Theorem 14.1** If  $y_t$  is i.i.d., then it strictly stationary.

Here are some examples of strictly stationary scalar processes. In each,  $e_t$  is i.i.d. and  $\mathbb{E}(e_t) = 0$ .

**Example 14.1**  $y_t = e_t + \theta e_{t-1}$ .

**Example 14.2**  $y_t = Z$  for some random variable  $Z$ .

**Example 14.3**  $y_t = (-1)^t Z$  for a random variable  $Z$  which is symmetrically distributed about 0.

**Example 14.4**  $y_t = Z \cos(\theta t)$  for a random variable  $Z$  symmetrically distributed about 0.

Here are some examples of processes which are not stationary.

**Example 14.5**  $y_t = t$ .

**Example 14.6**  $y_t = (-1)^t$ .

**Example 14.7**  $y_t = \cos(\theta t)$ .

**Example 14.8**  $y_t = \sqrt{t} e_t$ .

**Example 14.9**  $y_t = e_t + t^{-1/2} e_{t-1}$ .

**Example 14.10**  $y_t = y_{t-1} + e_t$  with  $y_0 = 0$ .

From the examples we can see that stationarity means that the distribution is constant over time. It does not mean, however, mean that the process has some sort of limited dependence, nor that there is an absence of periodic patterns. These restrictions are actually associated with the concepts of ergodicity and mixing, which we shall introduce in subsequent sections.

## 14.5 Transformations of Stationary Processes

One of the important properties of strict stationarity is that it is preserved by transformation. That is, transformations of strictly stationary processes are also strictly stationary. This includes transformations which include the full history of  $y_t$ .

**Theorem 14.2** If  $y_t$  is strictly stationary and  $x_t = \phi(y_t, y_{t-1}, y_{t-2}, \dots) \in \mathbb{R}^q$  is a random vector, then  $x_t$  is strictly stationary.

Theorem 14.2 is extremely useful both for the study of stochastic processes which are constructed from underlying errors, and for the study of sample statistics such as linear regression estimators which are functions of sample averages of squares and cross-products of the original data.

As an example, Theorem 14.2 applies to the infinite-order moving average transformation

$$x_t = \sum_{j=0}^{\infty} a_j y_{t-j} \quad (14.3)$$

where  $a_j$  are coefficients. We only need to verify that the series  $x_t$  converges almost surely. It turns out that a sufficient condition is that the coefficients are absolutely convergent.

**Theorem 14.3** If  $\sup_t \mathbb{E}|y_t| < \infty$  and  $\sum_{j=0}^{\infty} |a_j| < \infty$  then (14.3) converges almost surely. If, in addition,  $y_t$  is strictly stationary, then  $x_t$  is strictly stationary.

We give proofs of Theorems 14.2 and 14.3 in Section 14.46.

## 14.6 Convergent Series

Theorem 14.3 gives a condition under which the infinite series (14.3) is convergent. In this section we review some relevant concepts.

A series  $S_N = \sum_{j=0}^N a_j$  is **convergent** if it has a finite limit as  $N \rightarrow \infty$ , thus  $S_N \rightarrow S = \sum_{j=0}^{\infty} a_j$  with  $|S| < \infty$ . The series is **absolutely convergent** if  $\sum_{j=0}^{\infty} |a_j|$  has a finite limit, which holds if  $\sum_{j=0}^{\infty} |a_j| < \infty$ . Absolute convergence implies convergence.

There are several tests for convergence. Here are several.

1. The **comparison test** applies if  $0 \leq a_j \leq b_j$ . If  $\sum_{j=0}^{\infty} b_j$  converges then so does  $\sum_{j=0}^{\infty} a_j$ .
2. The **ratio test** applies if  $a_j \geq 0$ . If  $\lim_{N \rightarrow \infty} \frac{a_{N+1}}{a_N} < 1$  then the series is absolutely convergent.
3. The **integral test** applies if  $f(N) = a_N \geq 0$  and monotonically decreasing. If  $\int_1^{\infty} f(x) dx < \infty$  then  $\sum_{j=2}^{\infty} a_j \leq \int_1^{\infty} f(x) dx < \infty$  so is absolutely convergent.
4. The **Cauchy convergence** criterion states that  $S_N$  converges if and only if for all  $\varepsilon > 0$ , there is an  $N < \infty$  such that for all  $m \geq 1$ ,  $|S_{N+m} - S_N| \leq \varepsilon$ .

We now describe three convergent series which arise in our treatment of time series econometrics.

**Theorem 14.4**

1.  $\sum_{k=0}^{\infty} \beta^k = \frac{1}{1-\beta}$  is absolutely convergent if  $|\beta| < 1$ .
2.  $\sum_{k=1}^{\infty} k^q \beta^k$  is absolutely convergent if  $|\beta| < 1$ , for any  $q$ .
3.  $\sum_{k=1}^{\infty} k^{-r} \leq \frac{r}{r-1}$  is absolutely convergent if  $r > 1$ .

Parts 1 and 2 converge by the ratio test, and part 3 by the integral test.

To compute the limit for part 1,

$$A = \sum_{k=0}^{\infty} \beta^k = 1 + \sum_{k=1}^{\infty} \beta^k = 1 + \beta \sum_{k=0}^{\infty} \beta^k = 1 + \beta A.$$

Solving, we find  $A = 1/(1 - \beta)$ .

To close this section, we provide a few useful results.

**Theorem 14.5** (Silverman-Toeplitz). If  $a_\ell \rightarrow A$  as  $\ell \rightarrow \infty$ , and for weights  $w_{n\ell} \geq 0$  such that  $\sum_{\ell=1}^n w_{n\ell} \rightarrow 1$  and  $w_{n\ell} \rightarrow 0$  for each  $\ell$  as  $n \rightarrow \infty$ , then  $\sum_{\ell=1}^n w_{n\ell} a_\ell \rightarrow A$  as  $n \rightarrow \infty$ .

The proof is given in Section 14.46.

Setting  $w_{n\ell} = 1/n$  we obtain the following.

**Theorem 14.6** (Theorem of Cesàro means) If  $a_\ell \rightarrow A$  as  $\ell \rightarrow \infty$  then  $\frac{1}{n} \sum_{\ell=1}^n a_\ell \rightarrow A$  as  $n \rightarrow \infty$ .

The following useful result follows by taking the limit of the Riemann sum for the integral  $\int_0^1 x^r dx = 1/(1+r)$ .

**Theorem 14.7** As  $n \rightarrow \infty$ , for any  $r > 0$ ,  $n^{-1-r} \sum_{t=1}^n t^r \rightarrow 1/(1+r)$ .

## 14.7 Ergodicity

The assumption of stationarity is not sufficient for many purposes, as there are strictly stationary processes with no time series variation. As we described earlier, an example of a stationary process is  $y_t = Z$  for some random variable  $Z$ . This is random, but constant over all time. An implication is that the sample mean of  $y_t = Z$  will be inconsistent for the population mean.

We want a minimal sufficient assumption so that the law of large numbers will apply to the sample mean. It turns out that a sufficient condition is ergodicity. As it is a rather technical subject, we mention only a few highlights here. For a rigorous treatment see a standard textbook such as Walters (1982).

A time series  $y_t$  is **ergodic** if all invariant events are trivial, meaning that any event which is unaffected by time-shifts has probability either zero or one. This definition is rather abstract and difficult to grasp, but fortunately it is not needed by most economists.

A useful intuition is that if  $y_t$  is ergodic then its sample paths will pass through all parts of the sample space, never getting “stuck” in a subregion.

We will first describe the properties of ergodic series which will be relevant for our needs, and follow with the more rigorous technical definitions. For proof of the results, see Section 14.46.

First, many standard time series processes can be shown to be ergodic. A useful starting point is the observation that an i.i.d. sequence is ergodic.

**Theorem 14.8** If  $y_t$  is i.i.d., then it strictly stationary and ergodic.

Second, ergodicity, like stationarity, is preserved by transformation.

**Theorem 14.9** If  $y_t$  is strictly stationary and ergodic and  $x_t = \phi(y_t, y_{t-1}, y_{t-2}, \dots)$  is a random vector, then  $x_t$  is strictly stationary and ergodic.

As an example, the infinite-order moving average transformation (14.3) is ergodic if the input is ergodic and the coefficients are absolutely convergent.

**Theorem 14.10** If  $y_t$  is strictly stationary, ergodic,  $\mathbb{E}|y_t| < \infty$ , and  $\sum_{j=0}^{\infty} |a_j| < \infty$  then  $x_t = \sum_{j=0}^{\infty} a_j y_{t-j}$  is strictly stationary and ergodic.

We now present a useful property. It is that the Cesàro sum of the autocovariances limits to zero.

**Theorem 14.11** If  $y_t$  is strictly stationary, ergodic, and  $\mathbb{E}(y_t^2) < \infty$ , then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{\ell=1}^n \text{cov}(y_t, y_{t+\ell}) = 0. \quad (14.4)$$

The result (14.4) can be interpreted as that the covariances “on average” tend to zero. Some authors have mis-stated ergodicity as implying that the covariances tend to zero but this is not correct, as (14.4) allows, for example, the non-convergent sequence  $\text{cov}(y_t, y_{t+\ell}) = (-1)^\ell$ . The reason why (14.4) is particularly useful is because it is sufficient for the weak law of large numbers, as we discover later in Theorem 14.13.

We now give the formal definition of ergodicity for interested readers. As the concepts will not be used again, most readers can safely skip this discussion.

As we stated above, by definition the series  $y_t$  is ergodic if all invariant events are trivial. To understand this we introduce some technical definitions. First, we can write an event as  $A = \{\tilde{y}_t \in G\}$  where  $\tilde{y}_t = (\dots, y_{t-1}, y_t, y_{t+1}, \dots)$  is an infinite history and  $G \subset \mathbb{R}^{m\infty}$ . Second, the  $\ell^{\text{th}}$  **time-shift** of  $\tilde{y}_t$  is defined as  $\tilde{y}_{t+\ell} = (\dots, y_{t-1+\ell}, y_{t+\ell}, y_{t+1+\ell}, \dots)$ . Thus  $\tilde{y}_{t+\ell}$  replaces each observation in  $\tilde{y}_t$  by its  $\ell^{\text{th}}$  shifted value  $y_{t+\ell}$ . A time-shift of the event  $A = \{\tilde{y}_t \in G\}$  is  $A_\ell = \{\tilde{y}_{t+\ell} \in G\}$ . Third, an event  $A$  is called **invariant** if it is unaffected by a time-shift, so that  $A_\ell = A$ . Thus replacing any history  $\tilde{y}_t$  with its shifted history  $\tilde{y}_{t+\ell}$  doesn’t change the event. Invariant events are rather special. An example of an invariant event is  $A = \{\max_{-\infty < t < \infty} y_t \leq 0\}$ . Fourth, an event  $A$  is called **trivial** if either  $\mathbb{P}(A) = 0$  or  $\mathbb{P}(A) = 1$ . You can think of trivial events as essentially non-random. Recall, by definition,  $y_t$  is ergodic if all invariant events are trivial. This means that any event which is unaffected by a time shift is trivial – is essentially non-random. For example, again consider the invariant event  $A = \{\max_{-\infty < t < \infty} y_t \leq 0\}$ . If  $y_t = Z \sim N(0, 1)$  for all  $t$ , then  $\mathbb{P}(A) = \mathbb{P}(Z \leq 0) = 0.5$ . Since this does not equal 0 or 1 then  $y_t = Z$  is not ergodic. However, if  $y_t$  is i.i.d.  $N(0, 1)$  then  $\mathbb{P}\{\max_{-\infty < t < \infty} y_t \leq 0\} = 0$ . This is a trivial event. For  $y_t$  to be ergodic (it is in this case) all such invariant events must be trivial.

An important technical result is that ergodicity is equivalent to the following property.

**Theorem 14.12** A stationary series  $\mathbf{y}_t$  is ergodic if and only if for all events  $A$  and  $B$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{\ell=1}^n \mathbb{P}(A_\ell \cap B) = \mathbb{P}(A)\mathbb{P}(B). \quad (14.5)$$

This result is rather deep so we do not prove it here. See Walters (1982), Corollary 1.14.2, or Davidson (1994), Theorem 13.13. The limit in (14.5) is the Cesàro sum of  $\mathbb{P}(A_\ell \cap B)$ . Theorem 14.6 shows that a sufficient condition for (14.5) is that  $\mathbb{P}(A_\ell \cap B) \rightarrow \mathbb{P}(A)\mathbb{P}(B)$  which is known as **mixing**. Thus mixing implies ergodicity. Mixing, roughly, means that separated events are asymptotically independent. Ergodicity is weaker, only requiring that the events are asymptotically independent “on average”. We discuss mixing in Section 14.12.

## 14.8 Ergodic Theorem

The ergodic theorem is one of the most famous results in time series theory. There are actually several forms of the theorem, most of which concern almost sure convergence. For simplicity we state the theorem in terms of convergence in probability.

**Theorem 14.13 (Ergodic Theorem)** If  $\mathbf{y}_t$  is strictly stationary, ergodic, and  $\mathbb{E}\|\mathbf{y}_t\| < \infty$ , then as  $n \rightarrow \infty$ ,

$$\mathbb{E}\|\bar{\mathbf{y}} - \boldsymbol{\mu}\| \rightarrow 0 \quad (14.6)$$

and

$$\bar{\mathbf{y}} \xrightarrow{p} \boldsymbol{\mu} \quad (14.7)$$

where  $\boldsymbol{\mu} = \mathbb{E}(\mathbf{y}_t)$ .

The ergodic theorem shows that ergodicity is sufficient for consistent estimation. The moment condition  $\mathbb{E}\|\mathbf{y}_t\| < \infty$  is the same as in the WLLN for i.i.d. samples.

We now provide a proof of the ergodic theorem for the scalar case under the additional assumption that  $\text{var}(y_t) = \sigma^2 < \infty$ . A proof which relaxes this assumption is provided in Section 14.46.

By direct calculation

$$\text{var}(\bar{y}) = \frac{1}{n^2} \sum_{t=1}^n \sum_{j=1}^n \gamma(t-j)$$

where  $\gamma(\ell) = \text{cov}(x_t, x_{t+\ell})$ . The double sum is over all elements of an  $n \times n$  matrix whose  $tj^{th}$  element is  $\gamma(t-j)$ . The diagonal elements are  $\gamma(0) = \sigma^2$ , the first off-diagonal elements are  $\gamma(1)$ , the second off-diagonal elements are  $\gamma(2)$  and so on. This means that there are precisely  $n$  diagonal elements of  $\sigma^2$ ,  $2(n-1)$  equalling  $\gamma(1)$ , etc. Thus the above equals

$$\begin{aligned} \text{var}(\bar{y}) &= \frac{1}{n^2} (n\sigma^2 + 2(n-1)\gamma(1) + 2(n-2)\gamma(2) + \cdots + 2\gamma(n-1)) \\ &= \frac{\sigma^2}{n} + \frac{2}{n} \sum_{\ell=1}^n \left(1 - \frac{\ell}{n}\right) \gamma(\ell). \end{aligned} \quad (14.8)$$

This is a rather intriguing expression. It shows that the variance of the sample mean precisely equals  $\sigma^2/n$  (which is the variance of the sample mean under i.i.d. sampling) plus a weighted Cesàro mean of the autocovariances. The latter is zero under i.i.d. sampling, but is non-zero otherwise. Theorem 14.11

shows that the Cesàro mean of the autocovariances converges to zero. Let  $w_{n\ell} = 2(\ell/n^2)$ , which satisfy the conditions of Theorem 14.5 by Theorem 14.7.1. Then

$$\frac{2}{n} \sum_{\ell=1}^n \left(1 - \frac{\ell}{n}\right) \gamma(\ell) = \frac{2}{n^2} \sum_{\ell=1}^{n-1} \sum_{j=1}^{\ell} \gamma(j) = \sum_{\ell=1}^{n-1} w_{n\ell} \left( \frac{1}{\ell} \sum_{j=1}^{\ell} \gamma(j) \right) \rightarrow 0 \quad (14.9)$$

by Theorem 14.5.

Together, we have shown that (14.8) is  $o(1)$  under ergodicity. Hence  $\text{var}(\bar{y}) \rightarrow 0$ . Markov's inequality establishes that  $\bar{y} \xrightarrow{P} \mu$ .

## 14.9 Conditioning on Information Sets

In the past few sections we have introduced the concept of the infinite histories. We now consider conditional expectations given infinite histories.

First, some basics. Recall from probability theory that an **outcome** is an element of a sample space. An **event** is a set of outcomes. A probability law is a rule which assigns non-negative real numbers to events. When outcomes are infinite histories then events are collections of such histories, and a probability law is a rule which assigns numbers to collections of infinite histories.

Now we wish to define a conditional expectation given an infinite past history. Specifically, we wish to define

$$\mathbb{E}_{t-1}(y_t) = \mathbb{E}(y_t | y_{t-1}, y_{t-2}, \dots) \quad (14.10)$$

the expected value of  $y_t$  given the history  $\tilde{y}_{t-1} = (y_{t-1}, y_{t-2}, \dots)$  up to time  $t$ . Intuitively,  $\mathbb{E}_{t-1}(y_t)$  is the mean of the conditional distribution, the latter reflecting the information in the history. Mathematically this cannot be defined using (2.4) as the latter requires a joint density for  $(y_t, y_{t-1}, y_{t-2}, \dots)$  and the latter does not make much sense. Instead, we can appeal to Theorem 2.13, which states that the conditional expectation (14.10) exists if  $\mathbb{E}|y_t| < \infty$  and the probabilities  $\mathbb{P}(\tilde{y}_{t-1} \in A)$  are defined. The latter are the probabilities of events discussed in the previous paragraph. Thus the conditional expectations are well defined.

In this textbook we have avoided measure-theoretic terminology to keep the presentation accessible, and because it is my belief that measure theory is more distracting than helpful. However, it is standard in the time series literature to follow the measure-theoretic convention of writing (14.10) as the conditional expectation given by a  $\sigma$ -field. So at the risk of being overly-technical, we will follow this convention and write the expectation (14.10) as  $\mathbb{E}(y_t | \mathcal{F}_{t-1})$  where  $\mathcal{F}_{t-1} = \sigma(\tilde{y}_{t-1})$  is the  $\sigma$ -field generated by the history  $\tilde{y}_{t-1}$ . A  **$\sigma$ -field** (also known as a  $\sigma$ -algebra) is a collection of sets satisfying certain regularity conditions<sup>2</sup>. An important example is the Borel  $\sigma$ -field  $\mathcal{B}$ , which is the collection of open sets in  $\mathbb{R}$ . The  $\sigma$ -field generated by a random variable  $Y$  is the collection of measurable events involving  $Y$ . Similarly, the  $\sigma$ -field generated by an infinite history is the collection of measurable events involving this history. Intuitively,  $\mathcal{F}_{t-1}$  contains all the information available in the history  $\tilde{y}_{t-1}$ . Consequently, economists typically call  $\mathcal{F}_{t-1}$  an **information set** rather than a  $\sigma$ -field. As I said, in this textbook we endeavor to avoid measure theoretic complications so will follow the economists' label rather than the probabilists', but use the latter's notation as is conventional. To summarize, we will write  $\mathcal{F}_t = \sigma(y_t, y_{t-1}, \dots)$  to indicate the information set generated by an infinite history  $(y_t, y_{t-1}, \dots)$ , and will write (14.10) as  $\mathbb{E}(y_t | \mathcal{F}_{t-1})$ .

We now describe some properties about information sets  $\mathcal{F}_t$ .

First, they are nested:  $\mathcal{F}_{t-1} \subset \mathcal{F}_t$ . This means that information accumulates over time. Information is not lost.

Second, it is important to be precise about which variables are contained in the information set. Some economists are sloppy and refer to "the information set at time  $t$ " without specifying which variables are in the information set. It is better to be specific. For example, the information sets  $\mathcal{F}_{1t} = \sigma(y_t, y_{t-1}, \dots)$  and  $\mathcal{F}_{2t} = \sigma(y_t, x_t, y_{t-1}, x_{t-1}, \dots)$  are distinct, even though they are both dated at time  $t$ .

---

<sup>2</sup>A  $\sigma$ -field contains the universal set, is closed under complementation, and closed under countable unions.

Third, the conditional expectations (14.10) follow the law of iterated expectations and the conditioning theorem, thus

$$\begin{aligned}\mathbb{E}(\mathbb{E}(y_t | \mathcal{F}_{t-1}) | \mathcal{F}_{t-2}) &= \mathbb{E}(y_t | \mathcal{F}_{t-2}) \\ \mathbb{E}(\mathbb{E}(y_t | \mathcal{F}_{t-1})) &= \mathbb{E}(y_t),\end{aligned}$$

and

$$\mathbb{E}(y_{t-1}y_t | \mathcal{F}_{t-1}) = y_{t-1}\mathbb{E}(y_t | \mathcal{F}_{t-1}).$$

## 14.10 Martingale Difference Sequences

An important concept in economics is unforecastability, meaning that the conditional expectation is the unconditional expectation. This is similar to the properties of a regression error. An unforecastable process is called a **martingale difference sequence (MDS)**.

A MDS  $e_t$  is defined with respect to a specific sequence of information sets  $\mathcal{F}_t$ . Most commonly the latter are the **natural filtration**  $\mathcal{F}_t = \sigma(e_t, e_{t-1}, \dots)$  (the past history of  $e_t$ ), but it could be a larger information set. The only requirement is that  $e_t$  is adapted to  $\mathcal{F}_t$ , meaning that  $\mathbb{E}(e_t | \mathcal{F}_t) = e_t$ .

**Definition 14.3** The process  $(e_t, \mathcal{F}_t)$  is a **Martingale Difference Sequence (MDS)** if  $e_t$  is adapted to  $\mathcal{F}_t$ ,  $\mathbb{E}|e_t| < \infty$  and  $\mathbb{E}(e_t | \mathcal{F}_{t-1}) = 0$ .

In words, a MDS  $e_t$  is unforecastable in the mean. It is useful to notice that if we apply iterated expectations  $\mathbb{E}(e_t) = \mathbb{E}(\mathbb{E}(e_t | \mathcal{F}_{t-1})) = 0$ . Thus a MDS is mean zero.

The definition of a MDS requires the information sets  $\mathcal{F}_t$  to contain the information in  $e_t$ , but is broader in the sense that it can contain more information. When no explicit definition is given it is standard to assume that  $\mathcal{F}_t$  is the natural filtration. However, it is best to explicitly specify the information sets so there is no confusion.

The term “martingale difference sequence” refers to the fact that the summed process  $S_t = \sum_{j=1}^t e_j$  is a **martingale**, and  $e_t$  is its first-difference. A martingale  $S_t$  is defined as a process such that  $\mathbb{E}(S_t | \mathcal{F}_{t-1}) = S_{t-1}$ .

If  $e_t$  is i.i.d. and mean zero it is a MDS, but the reverse is not the case. To see this, first suppose that  $e_t$  is i.i.d. and mean zero. It is then independent of  $\mathcal{F}_{t-1} = \sigma(e_{t-1}, e_{t-2}, \dots)$ , so  $\mathbb{E}(e_t | \mathcal{F}_{t-1}) = \mathbb{E}(e_t) = 0$ . Thus an i.i.d. shock is a MDS as claimed.

To show that the reverse is not true let  $u_t$  be i.i.d.  $N(0, 1)$  and set

$$e_t = u_t u_{t-1}. \quad (14.11)$$

By the conditioning theorem,

$$\mathbb{E}(e_t | \mathcal{F}_{t-1}) = u_{t-1}\mathbb{E}(u_t | \mathcal{F}_{t-1}) = 0$$

so  $e_t$  is a MDS. The process (14.11) is not, however, i.i.d. One way to see this is to calculate the first autocovariance of  $e_t^2$ , which is

$$\begin{aligned}\text{cov}(e_t^2, e_{t-1}^2) &= \mathbb{E}(e_t^2 e_{t-1}^2) - \mathbb{E}(e_t^2)\mathbb{E}(e_{t-1}^2) \\ &= \mathbb{E}(u_t^2)\mathbb{E}(u_{t-1}^4)\mathbb{E}(u_{t-2}^2) - 1 \\ &= 2 \neq 0.\end{aligned}$$

Since the covariance is non-zero,  $e_t$  is not an independent sequence. Thus  $e_t$  is a MDS but not i.i.d.

An important property of a square integrable MDS is that it is serially uncorrelated. To see this, observe that by iterated expectations, the conditioning theorem, and the definition of a MDS, for  $k > 0$ ,

$$\begin{aligned}\text{cov}(e_t, e_{t-k}) &= \mathbb{E}(e_t e_{t-k}) \\ &= \mathbb{E}(\mathbb{E}(e_t e_{t-k} | \mathcal{F}_{t-1})) \\ &= \mathbb{E}(\mathbb{E}(e_t | \mathcal{F}_{t-1}) e_{t-k}) \\ &= \mathbb{E}(0 e_{t-k}) \\ &= 0.\end{aligned}$$

Thus the autocovariances and autocorrelations are zero.

A process that is serially uncorrelated, however, is not necessarily a MDS. Take the process

$$e_t = u_t + u_{t-1} u_{t-2}$$

where again  $u_t$  is i.i.d.  $N(0, 1)$ . The shock  $e_t$  is not a MDS since  $\mathbb{E}(e_t | \mathcal{F}_{t-1}) = u_{t-1} u_{t-2} \neq 0$ . However,

$$\begin{aligned}\text{cov}(e_t, e_{t-1}) &= \mathbb{E}(e_t e_{t-1}) \\ &= \mathbb{E}((u_t + u_{t-1} u_{t-2})(u_{t-1} + u_{t-2} u_{t-3})) \\ &= \mathbb{E}(u_t u_{t-1} + u_t u_{t-2} u_{t-3} + u_{t-1}^2 u_{t-2} + u_{t-1} u_{t-2}^2 u_{t-3}) \\ &= \mathbb{E}(u_t) \mathbb{E}(u_{t-1}) + \mathbb{E}(u_t) \mathbb{E}(u_{t-2}) \mathbb{E}(u_{t-3}) \\ &\quad + \mathbb{E}(u_{t-1}^2) \mathbb{E}(u_{t-2}) + \mathbb{E}(u_{t-1}) \mathbb{E}(u_{t-2}^2) \mathbb{E}(u_{t-3}) \\ &= 0.\end{aligned}$$

Similarly,  $\text{cov}(e_t, e_{t-k}) = 0$  for  $k \neq 0$ . Thus  $e_t$  is serially uncorrelated. We have proved the following.

**Theorem 14.14** If  $(e_t, \mathcal{F}_t)$  is a MDS and  $\mathbb{E}(e_t^2) < \infty$  then  $e_t$  is serially uncorrelated.

Another important special case is a homoskedastic martingale difference sequence.

**Definition 14.4** The MDS  $(e_t, \mathcal{F}_t)$  is a **Homoskedastic Martingale Difference Sequence (MDS)** if  $\mathbb{E}(e_t^2 | \mathcal{F}_{t-1}) = \sigma^2$ .

A homoskedastic MDS should more properly be called a conditionally homoskedastic MDS, because the property concerns the conditional distribution rather than the unconditional. That is, any strictly stationary MDS satisfies a constant variance  $\mathbb{E}(e_t^2)$  but only a homoskedastic MDS a constant conditional variance  $\mathbb{E}(e_t^2 | \mathcal{F}_{t-1})$ .

A homoskedastic MDS is analogous to a conditionally homoskedastic regression error. It is intermediate between a MDS and an i.i.d. sequence. Specifically, a (square integrable and mean zero) i.i.d. sequence is a homoskedastic MDS, and the latter is a MDS.

The reverse is not the case. First, a MDS is not necessarily conditionally homoskedastic. Consider the example  $e_t = u_t u_{t-1}$  given previously which we showed is a MDS. It is not conditionally homoskedastic, however, since

$$\mathbb{E}(e_t^2 | \mathcal{F}_{t-1}) = u_{t-1}^2 \mathbb{E}(u_t^2 | \mathcal{F}_{t-1}) = u_{t-1}^2$$

which is time-varying. Thus this MDS  $e_t$  is conditionally heteroskedastic. Second, a homoskedastic MDS is not necessarily i.i.d. Consider the following example. Set  $e_t = \sqrt{1 - 2/\eta_{t-1}} T_t$ , where  $T_t$  is distributed

with a student  $t$  distribution with degree of freedom parameter  $\eta_{t-1} = 2 + e_{t-1}^2$ . This is scaled so that  $\mathbb{E}(e_t | \mathcal{F}_{t-1}) = 0$  and  $\mathbb{E}(e_t^2 | \mathcal{F}_{t-1}) = 1$ , and is thus a homoskedastic MDS. The conditional distribution of  $e_t$  depends on  $e_{t-1}$  through the degree of freedom parameter. Hence  $e_t$  is not an independent sequence.

One way to think about the difference between MDS and i.i.d. shocks is in terms of forecastability. An i.i.d. process is fully unforecastable, in that no function of an i.i.d. process is forecastable. A MDS is unforecastable in the mean, but other moments may be forecastable.

## 14.11 CLT for Martingale Differences

We are interested in an asymptotic approximation for the distribution of standardized sample means such as

$$\mathbf{s}_n = \frac{1}{\sqrt{n}} \sum_{t=1}^n \mathbf{u}_t \quad (14.12)$$

where  $\mathbf{u}_t$  is mean zero with variance  $\mathbb{E}(\mathbf{u}_t \mathbf{u}'_t) = \Sigma < \infty$ . In this section we present a CLT for the case where  $\mathbf{u}_t$  is a martingale difference sequence.

**Theorem 14.15 (MDS CLT)** If  $\mathbf{u}_t$  is a strictly stationary and ergodic martingale difference sequence and  $\mathbb{E}(\mathbf{u}_t \mathbf{u}'_t) = \Sigma < \infty$ , then as  $n \rightarrow \infty$ ,

$$\mathbf{s}_n = \frac{1}{\sqrt{n}} \sum_{t=1}^n \mathbf{u}_t \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma).$$

The conditions for Theorem 14.15 are similar to the Lindeberg-Lévy CLT. The only difference is that the i.i.d. assumption has been replaced by the assumption of a strictly stationary and ergodic MDS.

It might be reasonable to conjecture that the CLT would hold under the broader assumption that  $\mathbf{u}_t$  is white noise. However, no such theory exists. At present, it is unknown if the MDS assumption can be weakened.

The proof of Theorem 14.15 is technically advanced so we do not present the full details, but instead refer readers to Theorem 3.2 of Hall and Heyde (1980) or Theorem 24.3 of Davidson (1994) (which are more general than Theorem 14.15, not requiring strict stationarity). To illustrate the role of the MDS assumption we give a sketch of the proof in Section 14.46.

## 14.12 Mixing

For many results, including a CLT for correlated (non-MDS) series, we need a stronger restriction on the dependence between observations than ergodicity.

Recalling the property (14.5) of ergodic sequences, we can measure the dependence between two events  $A$  and  $B$  by the discrepancy

$$\alpha(A, B) = |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|. \quad (14.13)$$

This equals 0 when  $A$  and  $B$  are independent, and is positive otherwise. In general,  $\alpha(A, B)$  can be used to measure the degree of dependence between the events  $A$  and  $B$ .

Now consider the two information sets ( $\sigma$ -fields)

$$\begin{aligned} \mathcal{F}_{-\infty}^t &= \sigma(\dots, \mathbf{y}_{t-1}, \mathbf{y}_t) \\ \mathcal{F}_t^\infty &= \sigma(\mathbf{y}_t, \mathbf{y}_{t+1}, \dots). \end{aligned}$$

The first is the history of the series up until period  $t$ , and the second is the history of the series starting in period  $t$  and going forward. We then separate the information sets by  $\ell$  periods, that is, take  $\mathcal{F}_{-\infty}^{t-\ell}$  and  $\mathcal{F}_t^\infty$ . We can measure the degree of dependence between the information sets by taking all events in each, and then taking the largest discrepancy (14.13). This is

$$\alpha(\ell) = \sup_{A \in \mathcal{F}_{-\infty}^{t-\ell}, B \in \mathcal{F}_t^\infty} \alpha(A, B).$$

The constants  $\alpha(\ell)$  are known as the **mixing coefficients**. We say that  $y_t$  is **strong mixing** if  $\alpha(\ell) \rightarrow 0$  as  $\ell \rightarrow \infty$ . This means that as the time separation increases between the information sets, the degree of dependence decreases, eventually reaching independence.

From the Theorem of Cesàro Means, strong mixing implies (14.5) which is equivalent to ergodicity. Thus a mixing process is ergodic.

An intuition concerning mixing can be colorfully illustrated by the following example due to Halmos (1956). A martini is a drink consisting of a large portion of gin and a small part of vermouth. Suppose that you pour a serving of gin into a martini glass, pour a small amount of vermouth on top, and then stir the drink with a swizzle stick. If your stirring process is mixing, with each turn of the stick the vermouth will become more evenly distributed throughout the gin, and asymptotically (as the number of stirs tends to infinity) the vermouth and gin distributions will become independent<sup>3</sup>. If so, we say this is a mixing process.

For applications, mixing is often useful when we can characterize the rate at which the coefficients  $\alpha(\ell)$  decline to zero. There are two types of conditions which are seen in asymptotic theory: rates and summation. Rate conditions take the form  $\alpha(\ell) = O(\ell^{-r})$  or  $\alpha(\ell) = o(\ell^{-r})$ . Summation conditions take the form  $\sum_{\ell=0}^{\infty} \alpha(\ell)^r < \infty$  or  $\sum_{\ell=0}^{\infty} \ell^s \alpha(\ell)^r < \infty$ .

There are alternative measures of dependence beyond (14.13) and many have been proposed. Strong mixing is one of the weakest (and thus embraces a wide set of time series processes) but is insufficiently strong for some applications. Another popular dependence measure is known as **absolute regularity** or  **$\beta$ -mixing**. The  $\beta$ -mixing coefficients are

$$\beta(\ell) = \sup_{A \in \mathcal{F}_t^\infty} \mathbb{E} \left| \mathbb{P}(A | \mathcal{F}_{-\infty}^{t-\ell}) - \mathbb{P}(A) \right|.$$

Absolute regularity is stronger than strong mixing in the sense that  $\beta(\ell) \rightarrow 0$  implies  $\alpha(\ell) \rightarrow 0$ , and rates conditions for the  $\beta$ -mixing coefficients imply the same rate conditions for the strong mixing coefficients.

One reason why mixing is useful for applications is that it is preserved by transformations.

**Theorem 14.16** If  $y_t$  has mixing coefficients  $\alpha_y(\ell)$  and  $x_t = \phi(y_t, y_{t-1}, y_{t-2}, \dots, y_{t-q})$  then  $x_t$  has mixing coefficients  $\alpha_x(\ell) \leq \alpha_y(\ell - q)$  (for  $\ell \geq q$ ). The coefficients  $\alpha_x(m)$  satisfy the same summation and rate conditions as  $\alpha_y(\ell)$ .

A limitation of the above result is that it is confined to a finite number of lags, unlike the transformation results for stationarity and ergodicity.

Mixing can be a useful tool because of the following inequalities.

---

<sup>3</sup>Of course, if you really make an asymptotic number of stirs, you will never finish stirring and you won't be able to enjoy the martini. Hence in practice it is advised to stop stirring before the number of stirs actually reaches infinity.

**Theorem 14.17** Suppose that  $x_{t-\ell}$  and  $z_t$  are random variables which are  $\mathcal{F}_{-\infty}^{t-\ell}$  and  $\mathcal{F}_t^\infty$  measurable, respectively.

1. If  $|x_t| \leq C_1$  and  $|z_t| \leq C_2$  then

$$|\text{cov}(x_{t-\ell}, z_t)| \leq 4C_1 C_2 \alpha(\ell).$$

2. If  $\mathbb{E}|x_t|^r < \infty$  and  $\mathbb{E}|z_t|^q < \infty$  for  $1/r + 1/q < 1$  then

$$|\text{cov}(x_{t-\ell}, z_t)| \leq 8 (\mathbb{E}|x_t|^r)^{1/r} (\mathbb{E}|z_t|^q)^{1/q} \alpha(\ell)^{1-1/r-1/q}.$$

3. If  $\mathbb{E}(y_t) = 0$  and  $\mathbb{E}|y_t|^r < \infty$  for  $r \geq 1$  then

$$\mathbb{E} \left| \mathbb{E} \left( y_t \mid \mathcal{F}_{-\infty}^{t-\ell} \right) \right| \leq 6 (\mathbb{E}|y_t|^r)^{1/r} \alpha(\ell)^{1-1/r}.$$

The proof is given in Section 14.46. The following follows fairly directly from the definition of mixing.

**Theorem 14.18** If  $y_t$  is i.i.d. then it is strong mixing and ergodic.

### 14.13 CLT for Correlated Observations

In this section we develop a CLT for the normalized mean  $S_n$  defined in (14.12) allowing the variables  $u_t$  to be serially correlated.

In (14.8) we found that in the scalar case

$$\text{var}(S_n) = \sigma^2 + 2 \sum_{\ell=1}^n \left( 1 - \frac{\ell}{n} \right) \gamma(\ell)$$

where  $\sigma^2 = \text{var}(u_t)$  and  $\gamma(\ell) = \text{cov}(u_t, u_{t-\ell})$ . Since  $\gamma(-\ell) = \gamma(\ell)$  this can be written as

$$\text{var}(S_n) = \sum_{\ell=-n}^n \left( 1 - \frac{|\ell|}{n} \right) \gamma(\ell). \quad (14.14)$$

In the vector case define the variance

$$\boldsymbol{\Sigma} = \mathbb{E}(\mathbf{u}_t \mathbf{u}'_t)$$

and the matrix covariance

$$\boldsymbol{\Gamma}(\ell) = \mathbb{E}(\mathbf{u}_t \mathbf{u}'_{t-\ell})$$

which satisfies  $\boldsymbol{\Gamma}(-\ell) = \boldsymbol{\Gamma}(\ell)'$ . We obtain by a calculation analogous to (14.14)

$$\begin{aligned} \text{var}(S_n) &= \boldsymbol{\Sigma} + \sum_{\ell=1}^n \left( 1 - \frac{\ell}{n} \right) (\boldsymbol{\Gamma}(\ell) + \boldsymbol{\Gamma}(\ell)') \\ &= \sum_{\ell=-n}^n \left( 1 - \frac{|\ell|}{n} \right) \boldsymbol{\Gamma}(\ell). \end{aligned} \quad (14.15)$$

A necessary condition for  $\mathbf{S}_n$  to converge to a normal distribution is that the variance (14.15) converges to a limit. Indeed,

$$\sum_{\ell=1}^n \left(1 - \frac{\ell}{n}\right) \Gamma(\ell) = \frac{1}{n} \sum_{\ell=1}^{n-1} \sum_{j=1}^{\ell} \Gamma(j) \longrightarrow \sum_{\ell=0}^{\infty} \Gamma(\ell) \quad (14.16)$$

where the convergence holds by Theorem 14.6 if the limit in (14.16) is convergent. A necessary condition for this to hold is that the covariances  $\Gamma(\ell)$  decline to zero as  $\ell \rightarrow \infty$ , which is stronger than ergodicity. A sufficient condition is that the covariances are absolutely summable, which can be verified using a mixing inequality. Using the triangle inequality (B.16) and Theorem 14.17.2, for  $r > 2$

$$\sum_{\ell=0}^{\infty} \|\Gamma(\ell)\| \leq 8 (\mathbb{E} \|\mathbf{u}_t\|^r)^{2/r} \sum_{\ell=0}^{\infty} \alpha(\ell)^{1-2/r}.$$

This implies that (14.15) converges if  $\mathbb{E} \|\mathbf{u}_t\|^r < \infty$  and  $\sum_{\ell=0}^{\infty} \alpha(\ell)^{1-2/r} < \infty$ . We conclude that under these assumptions

$$\text{var}(\mathbf{S}_n) \longrightarrow \sum_{\ell=-\infty}^{\infty} \Gamma(\ell) \stackrel{\text{def}}{=} \boldsymbol{\Omega}. \quad (14.17)$$

It turns out that these conditions are sufficient for the CLT.

**Theorem 14.19** If  $\mathbf{u}_t$  is strictly stationary with mixing coefficients  $\alpha(\ell)$ ,  $\mathbb{E}(\mathbf{u}_t) = \mathbf{0}$ , for some  $r > 2$ ,  $\mathbb{E} \|\mathbf{u}_t\|^r < \infty$  and  $\sum_{\ell=1}^{\infty} \alpha(\ell)^{1-2/r} < \infty$ , then (14.17) is convergent, and

$$\mathbf{S}_n = \frac{1}{\sqrt{n}} \sum_{t=1}^n \mathbf{u}_t \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}).$$

The proof is in Section 14.46.

The theorem requires  $r > 2$  finite moments which is stronger than the MDS CLT. The summability condition on the mixing coefficients in Theorem 14.19 is considerably stronger than ergodicity. There is a trade off involving the choice of  $r$ . A larger  $r$  means more moments are required finite, but a slower decay in the coefficients  $\alpha(\ell)$  is allowed. Smaller  $r$  is less restrictive regarding moments, but requires a faster decay rate in the mixing coefficients.

## 14.14 Linear Projection

In Chapter 2 we extensively studied the properties of linear projection models. In the context of stationary time series we can use similar tools. An important extension is to allow for projections onto infinite dimensional random vectors. For this analysis we assume that  $y_t$  is covariance stationary.

Recall that when  $(y, \mathbf{x})$  have a joint distribution with bounded variances, the linear projection of  $y$  onto  $\mathbf{x}$  (the best linear predictor) is the minimizer of

$$S(\boldsymbol{\beta}) = \mathbb{E}(y - \boldsymbol{\beta}' \mathbf{x})^2$$

and has the solution

$$\mathcal{P}(y | \mathbf{x}) = \mathbf{x}' (\mathbb{E}(\mathbf{x} \mathbf{x}'))^{-1} \mathbb{E}(\mathbf{x} y).$$

We are interested in the best linear predictor of the random variable  $y_t$  given the infinite past history  $\tilde{\mathbf{y}}_{t-1} = (\dots, y_{t-2}, y_{t-1})$ . Linear functions of  $\tilde{\mathbf{y}}_{t-1}$  take the form  $\alpha_0 + \sum_{j=1}^{\infty} \alpha_j y_{t-j}$ . The best linear predictor minimizes the mean squared prediction error

$$S(\alpha_0, \alpha_1, \dots) = \mathbb{E} \left( y_t - \alpha_0 - \sum_{j=1}^{\infty} \alpha_j y_{t-j} \right)^2.$$

The solution takes the form

$$\mathcal{P}_{t-1}(y_t) = \mathcal{P}(y_t | \tilde{\mathbf{y}}_{t-1}) = \alpha_0 + \sum_{j=1}^{\infty} \alpha_j y_{t-j}.$$

We call this the projection of  $y_t$  onto  $\tilde{\mathbf{y}}_{t-1}$ . This is the projection analog of the conditional expectation (14.10).

The projection error is

$$e_t = y_t - \mathcal{P}_{t-1}(y_t). \quad (14.18)$$

We can write the decomposition of  $y_t$  into projection and projection error as a regression equation

$$y_t = \alpha_0 + \sum_{j=1}^{\infty} \alpha_j y_{t-j} + e_t.$$

From the projection theorem for Hilbert spaces (see, e.g., Theorem 2.3.1 of Brockwell and Davis (1991)) the projection  $\mathcal{P}_{t-1}(y_t)$  and projection error  $e_t$  are unique. The projection error has finite variance

$$\sigma^2 = \mathbb{E}(e_t^2) \leq \mathbb{E}(y_t^2) < \infty.$$

Also, by Theorem 14.2, if  $y_t$  is strictly stationary then  $\mathcal{P}_{t-1}(y_t)$  and  $e_t$  are strictly stationary.

The projection error is mean zero and uncorrelated with the elements of  $\tilde{\mathbf{y}}_{t-1}$ . This implies that

$$\begin{aligned} \mathbb{E}(e_{t-\ell} e_t) &= \mathbb{E}\left(\left(y_{t-\ell} - \alpha_0 - \sum_{j=1}^{\infty} \alpha_j y_{t-\ell-j}\right) e_t\right) \\ &= \mathbb{E}(y_{t-\ell} e_t) - \alpha_0 \mathbb{E}(e_t) - \sum_{j=1}^{\infty} \alpha_j \mathbb{E}(y_{t-\ell-j} e_t) \\ &= 0. \end{aligned}$$

Thus the projection errors are serially uncorrelated.

We state these results formally.

**Theorem 14.20** If  $y_t$  is covariance stationary it has the projection equation

$$y_t = \alpha_0 + \sum_{j=1}^{\infty} \alpha_j y_{t-j} + e_t.$$

The projection error  $e_t$  satisfies

$$\begin{aligned} \mathbb{E}(e_t) &= 0 \\ \mathbb{E}(y_{t-j} e_t) &= 0 \quad j \geq 1 \\ \mathbb{E}(e_{t-j} e_t) &= 0 \quad j \geq 1 \end{aligned}$$

and

$$\sigma^2 = \mathbb{E}(e_t^2) \leq \mathbb{E}(y_t^2) < \infty. \quad (14.19)$$

If  $y_t$  is strictly stationary, then  $e_t$  is strictly stationary.

## 14.15 White Noise

The projection error  $e_t$  is mean zero, has a finite variance, and is serially uncorrelated. This describes what is known as a white noise process.

**Definition 14.5** The process  $e_t$  is **white noise** if  $\mathbb{E}(e_t) = 0$ ,  $\mathbb{E}(e_t^2) = \sigma^2 < \infty$ , and  $\text{cov}(e_t, e_{t-k}) = 0$  for  $k \neq 0$ .

A MDS is white noise (Theorem 14.14) but the reverse is not true as shown by the example  $e_t = u_t + u_{t-1}u_{t-2}$  given in Section 14.10, which is white noise but not a MDS.

Therefore, the following types of shocks are nested: i.i.d., MDS, and white noise, with i.i.d. being the most narrow class, and white noise the broadest.

## 14.16 The Wold Decomposition

In Section 14.14 we showed that we can express a stationary time series by a projection equation with white noise errors. An alternative is to express the time series as a linear function of the same white noise errors. This is a famous result known as the Wold decomposition.

**Theorem 14.21 (The Wold Decomposition)** If  $y_t$  is covariance stationary and  $\sigma^2 > 0$  where  $\sigma^2$  is the projection error variance from (14.19), then  $y_t$  has the linear representation

$$y_t = \mu_t + \sum_{j=0}^{\infty} b_j e_{t-j}, \quad (14.20)$$

where  $e_t$  are the white noise projection errors (14.18),  $b_0 = 1$ ,

$$\sum_{j=1}^{\infty} b_j^2 < \infty \quad (14.21)$$

and

$$\mu_t = \lim_{m \rightarrow \infty} \mathcal{P}_{t-m}(y_t). \quad (14.22)$$

The Wold decomposition shows that  $y_t$  can be written as a linear function of the white noise projection errors plus  $\mu_t$ . The infinite sum in (14.20) is also known as a **linear process**. The Wold decomposition is a foundational result for linear time series analysis. Since any covariance stationary process can be written in this format we can use linear parametric models (autoregressive and moving average) as approximations.

The series  $\mu_t$  is the projection of  $y_t$  on the history from the infinite past. It is the part of  $y_t$  which is perfectly predictable from its past values, and is called the **deterministic component**. In most cases  $\mu_t = \mu$ , the unconditional mean of  $y_t$ . However, it is possible for stationary processes to have more substantive deterministic components. An example is

$$\mu_t = \begin{cases} (-1)^t & \text{with probability } 1/2 \\ (-1)^{t+1} & \text{with probability } 1/2 \end{cases}.$$

This series is strictly stationary, has mean zero and variance one. However, it is perfectly predictable given the previous history, as it simply oscillates between  $-1$  and  $1$ .

In practical applied time series analysis, deterministic components are typically excluded by assumption. We call a stationary time series **non-deterministic**<sup>4</sup> if  $\mu_t = \mu$ , a constant. In this case the Wold decomposition has a simpler form.

**Theorem 14.22** If  $y_t$  is covariance stationary and non-deterministic then  $y_t$  has the linear representation

$$y_t = \mu + \sum_{j=0}^{\infty} b_j e_{t-j},$$

where  $b_j$  satisfy (14.21) and  $e_t$  are the white noise projection errors (14.18).

A limitation of the Wold decomposition is the restriction to linearity. Effectively, it says that there is a valid linear approximation within the class of linear models. It excludes alternative (nonlinear) models by assumption.

For a proof of Theorem 14.21 see Section 14.46.

## 14.17 Linear Models

In the previous sections we showed that any non-deterministic covariance stationary time series  $y_t$  has the projection representations

$$y_t = \alpha_0 + \sum_{j=1}^{\infty} \alpha_j y_{t-j} + e_t$$

and

$$y_t = \mu + \sum_{j=0}^{\infty} b_j e_{t-j}$$

where the errors  $e_t$  are white noise projection errors. These representations help us understand that linear models can be used as approximations for stationary time series.

For the next several sections we reverse the analysis. We will assume a specific linear model, and then study the properties of the resulting time series. In particular, we will be seeking conditions under which the process is stationary. This helps us understand the properties of linear models.

Throughout, we will be assuming that the error  $e_t$  is a strictly stationary and ergodic MDS with a finite variance. This allows as a special case the stronger assumption that  $e_t$  is i.i.d., but is less restrictive. In particular, it allows for conditional heteroskedasticity.

## 14.18 Moving Average Processes

The **first-order moving average process**, denoted MA(1), is

$$y_t = \mu + e_t + \theta e_{t-1}$$

where  $e_t$  is a strictly stationary and ergodic white noise process. The model is called a “moving average” because  $y_t$  is a weighted average of the shocks  $e_t$  and  $e_{t-1}$ .

---

<sup>4</sup>Most authors define purely non-deterministic as the case  $\mu_t = 0$ . We allow for a non-zero mean so to accommodate practical time series applications.

It is straightforward to calculate that a MA(1) has the following moments.

$$\begin{aligned}\mathbb{E}(y_t) &= \mu \\ \text{var}(y_t) &= (1 + \theta^2)\sigma^2 \\ \gamma(1) &= \theta\sigma^2 \\ \rho(1) &= \frac{\theta}{1 + \theta^2} \\ \gamma(k) &= \rho(k) = 0, \quad k \geq 2.\end{aligned}$$

Thus the MA(1) process has a non-zero first autocorrelation, with the remainder zero.

An MA(1) process with  $\theta \neq 0$  is serially correlated, with each pair of adjacent observations  $(y_{t-1}, y_t)$  correlated. If  $\theta > 0$  the pair are positively correlated, while if  $\theta < 0$  they are negatively correlated. The serial correlation, however, is limited in that observations separated by multiple periods are mutually independent.

The  $q^{th}$ -order moving average process, denoted **MA(q)**, is

$$y_t = \mu + \theta_0 e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q}$$

where  $\theta_0 = 1$ . It is straightforward to calculate that a MA(q) has the following moments.

$$\begin{aligned}\mathbb{E}(y_t) &= \mu \\ \text{var}(y_t) &= \left(\sum_{j=0}^q \theta_j^2\right)\sigma^2 \\ \gamma(k) &= \left(\sum_{j=0}^{q-k} \theta_{j+k}\theta_j\right)\sigma^2, \quad k \leq q \\ \rho(k) &= \frac{\sum_{j=0}^{q-k} \theta_{j+k}\theta_j}{\sum_{j=0}^q \theta_j^2} \\ \gamma(k) &= \rho(k) = 0, \quad k > q.\end{aligned}$$

In particular, a MA(q) has  $q$  non-zero autocorrelations, with the remainder zero.

A MA(q) process  $y_t$  is strictly stationary and ergodic.

A MA(q) process with moderately large  $q$  can have considerably more complicated dependence relations than an MA(1) process. One specific pattern which can be induced by a MA process is smoothing. Suppose that the coefficients  $\theta_j$  all equal 1. Then  $y_t$  is a smoothed version of the shocks  $e_t$ .

To illustrate, Figure 14.5(a) displays a plot of a simulated white noise (i.i.d.  $N(0, 1)$ ) process with  $n = 120$  observations. Figure 14.5(b) displays a plot of an MA(8) process constructed with the same innovations, with  $\theta_j = 1$ ,  $j = 1, \dots, 8$ . You can see that the white noise has no predictable behavior, while the MA(8) is very smooth.

## 14.19 Infinite-Order Moving Average Process

An **infinite-order moving average process**, denoted **MA( $\infty$ )**, also known as a **linear process**, is

$$y_t = \mu + \sum_{j=0}^{\infty} \theta_j e_{t-j}$$

where  $e_t$  is a strictly stationary and ergodic white noise process, and

$$\sum_{j=0}^{\infty} \theta_j^2 < \infty.$$

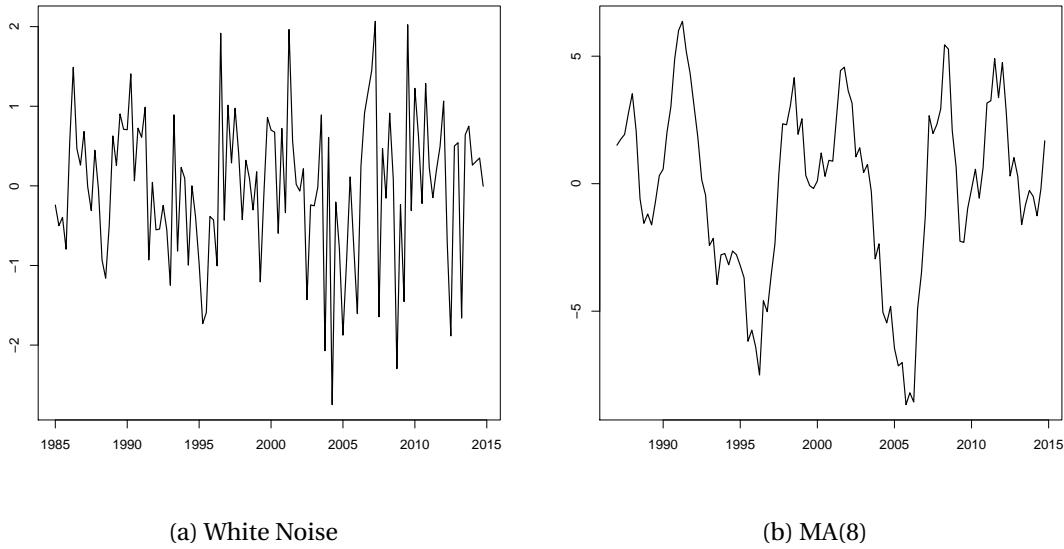


Figure 14.5: White Noise and MA(8)

A linear process has the following moments:

$$\begin{aligned}\mathbb{E}(y_t) &= \mu \\ \text{var}(y_t) &= \left( \sum_{j=0}^{\infty} \theta_j^2 \right) \sigma^2 \\ \gamma(k) &= \left( \sum_{j=0}^{\infty} \theta_{j+k} \theta_j \right) \sigma^2 \\ \rho(k) &= \frac{\sum_{j=0}^{\infty} \theta_{j+k} \theta_j}{\sum_{j=0}^q \theta_j^2}.\end{aligned}$$

**Theorem 14.23** The MA( $\infty$ ) process  $y_t$  converges almost surely, and is strictly stationary and ergodic.

For a proof see Section 14.46.

## 14.20 Lag Operator

An algebraic construct which is useful for the analysis of time series models is the lag operator.

**Definition 14.6** The **lag operator**  $L$  satisfies  $Ly_t = y_{t-1}$ .

Defining  $L^2 = LL$ , we see that  $L^2 y_t = Ly_{t-1} = y_{t-2}$ . In general,  $L^k y_t = y_{t-k}$ .

Using the lag operator, the MA(q) model can be written in the format

$$\begin{aligned} y_t &= \theta_0 e_t + \theta_1 L e_t + \cdots + \theta_q L^q e_t \\ &= (\theta_0 + \theta_1 L + \cdots + \theta_q L^q) e_t \\ &= \theta(L) e_t \end{aligned}$$

where

$$\theta(L) = \theta_0 + \theta_1 L + \cdots + \theta_q L^q$$

is a  $q^{th}$ -order polynomial in the lag operator  $L$ . The expression  $y_t = \theta(L)e_t$  is compact way to write the model.

## 14.21 First-Order Autoregressive Process

The **first-order autoregressive process**, denoted **AR(1)**, is

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + e_t \quad (14.23)$$

where  $e_t$  is a strictly stationary and ergodic white noise process. The AR(1) model is probably the single most important model in econometric time series analysis.

As a simple motivating example, let  $y_t$  be the employment level (number of jobs) in an economy. Suppose that a fixed fraction  $1 - \alpha_1$  of employees lose their jobs each period, and a random number  $u_t$  of new employees are hired each period. Setting  $\alpha_0 = \mathbb{E}(u_t)$  and  $e_t = u_t - \alpha_0$ , this implies the law of motion (14.23).

To illustrate the behavior of the AR(1) process, Figure 14.6 plots two simulated AR(1) processes. Each is generated using the white noise process  $e_t$  displayed in Figure 14.5(a). The plot in Figure 14.6(a) sets  $\alpha_1 = 0.5$  and the plot in Figure 14.6(b) sets  $\alpha_1 = 0.95$ . You can see how both are more smooth than the white noise process, and that the smoothing increases with  $\alpha$ .

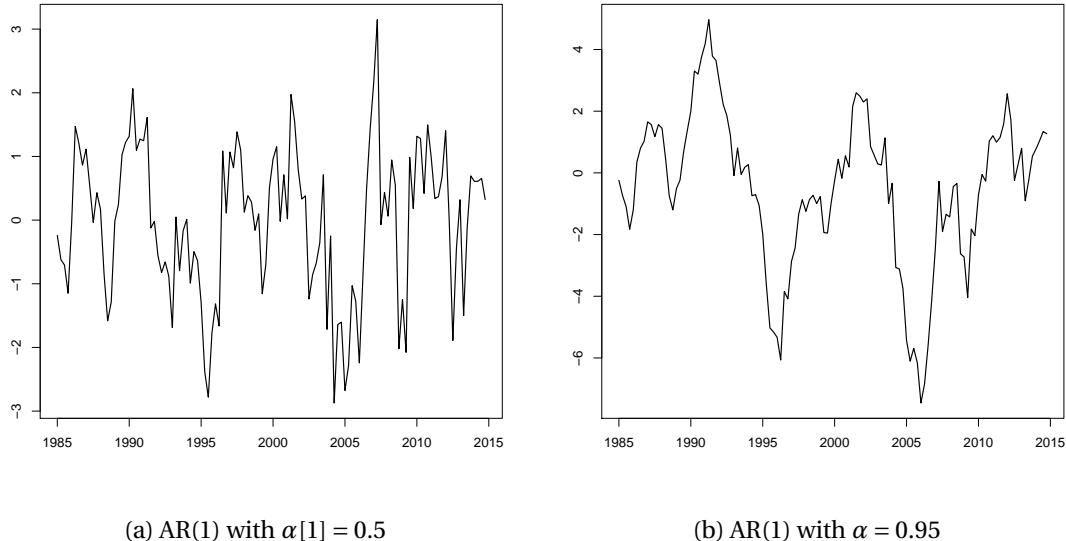


Figure 14.6: AR(1) Processes

Our first goal is to obtain conditions under which (14.23) is stationary. We can do so by showing that  $y_t$  can be written as a convergent linear process and then appealing to Theorem 14.9. To find a linear process representation for  $y_t$  we can use backward recursion. Notice that  $y_t$  in (14.23) depends on its

previous value  $y_{t-1}$ . If we take (14.23) and lag it one period we find  $y_{t-1} = \alpha_0 + \alpha_1 y_{t-2} + e_{t-1}$ . Substituting this into (14.23) we find

$$\begin{aligned} y_t &= \alpha_0 + \alpha_1 (\alpha_0 + \alpha_1 y_{t-2} + e_{t-1}) + e_t \\ &= \alpha_0 + \alpha_1 \alpha_0 + \alpha_1^2 y_{t-2} + \alpha_1 e_{t-1} + e_t. \end{aligned}$$

Similarly we can lag (14.28) twice to find  $y_{t-2} = \alpha_0 + \alpha_1 y_{t-3} + e_{t-2}$  and can be used to substitute out  $y_{t-2}$ . Continuing recursively  $t$  times, we find

$$\begin{aligned} y_t &= \alpha_0 (1 + \alpha_1 + \alpha_1^2 + \cdots + \alpha_1^{t-1}) + \alpha_1^t y_0 + \alpha_1^{t-1} e_1 + \alpha_1^{t-2} e_2 + \cdots + e_t \\ &= \alpha_0 \sum_{j=0}^{t-1} \alpha_1^j + \alpha_1^t y_0 + \sum_{j=0}^{t-1} \alpha_1^j e_{t-j}. \end{aligned}$$

Thus  $y_t$  equals an intercept plus the scaled initial condition  $\alpha_1^t y_0$  and the moving average  $\sum_{j=0}^{t-1} \alpha_1^j e_{t-j}$ .

Now suppose we continue this recursion into the infinite past. By Theorem 14.23 this converges if  $\sum_{j=0}^{\infty} |\alpha_1|^j < \infty$  which holds when  $|\alpha_1| < 1$  by Theorem 14.4.1. The intercept converges to  $\alpha_0 / (1 - \alpha_1)$ . We deduce the following:

**Theorem 14.24** If  $|\alpha_1| < 1$  then the AR(1) process (14.23) has the convergent representation

$$y_t = \mu + \sum_{j=0}^{\infty} \alpha_1^j e_{t-j} \quad (14.24)$$

where  $\mu = \alpha_0 / (1 - \alpha_1)$ . The AR(1) process  $y_t$  is strictly stationary and ergodic.

We can compute the moments of  $y_t$  from (14.24)

$$\begin{aligned} \mathbb{E}(y_t) &= \mu + \sum_{k=0}^{\infty} \alpha_1^k \mathbb{E}(e_{t-k}) = \mu \\ \text{var}(y_t) &= \sum_{k=0}^{\infty} \alpha_1^{2k} \text{var}(e_{t-k}) = \frac{\sigma^2}{1 - \alpha_1^2}. \end{aligned}$$

An alternative informal way to calculate the moments is as follows. Apply expectations to both sides of (14.23)

$$\mathbb{E}(y_t) = \alpha_0 + \alpha_1 \mathbb{E}(y_{t-1}) + \mathbb{E}(e_t) = \alpha_0 + \alpha_1 \mathbb{E}(y_{t-1}).$$

Stationarity implies  $\mathbb{E}(y_{t-1}) = \mathbb{E}(y_t)$ . Solving we find  $\mathbb{E}(y_t) = \alpha_0 / (1 - \alpha_1)$ . Similarly,

$$\text{var}(y_t) = \text{var}(\alpha y_{t-1} + e_t) = \alpha_1^2 \text{var}(y_{t-1}) + \text{var}(e_t) = \alpha_1^2 \text{var}(y_{t-1}) + \sigma^2.$$

Stationarity implies  $\text{var}(y_{t-1}) = \text{var}(y_t)$ . Solving we find  $\text{var}(y_t) = \sigma^2 / (1 - \alpha_1^2)$ . This method is useful for calculation of autocovariances and autocorrelations. For simplicity set  $\alpha_0 = 1$ . We find

$$\gamma(1) = \mathbb{E}(y_{t-1} y_t) = \mathbb{E}(y_{t-1} (\alpha_1 y_{t-1} + e_t)) = \alpha_1 \text{var}(y_t)$$

so

$$\rho(1) = \gamma(1) / \text{var}(y_t) = \alpha_1.$$

Furthermore,

$$\gamma(k) = \mathbb{E}(y_{t-k} y_t) = \mathbb{E}(y_{t-k} (\alpha_1 y_{t-1} + e_t)) = \alpha_1 \gamma(k-1).$$

By recursion we obtain

$$\begin{aligned}\gamma(k) &= \alpha_1^k \text{var}(y_t) \\ \rho(k) &= \alpha_1^k.\end{aligned}$$

Thus the AR(1) process with  $\alpha_1 \neq 0$  has non-zero autocorrelations of all orders which decay to zero geometrically as  $k$  increases. For  $\alpha_1 > 0$  the autocorrelations are all positive. For  $\alpha_1 < 0$  the autocorrelations alternate in sign.

We can also express the AR(1) process using the lag operator notation:

$$(1 - \alpha_1 L) y_t = \alpha_0 + e_t. \quad (14.25)$$

We can write this as

$$\alpha(L) y_t = \alpha_0 + e_t$$

where

$$\alpha(L) = 1 - \alpha_1 L.$$

We call  $\alpha(z) = 1 - \alpha_1 z$  the **autoregressive polynomial** of  $y_t$ .

This suggests an alternative way of obtaining the representation (14.24). We can invert the operator  $(1 - \alpha_1 L)$  to write  $y_t$  as a function of lagged  $e_t$ . That is, suppose that the inverse operator  $(1 - \alpha_1 L)^{-1}$  exists. Then we can use this operator on (14.25) to find

$$y_t = (1 - \alpha_1 L)^{-1} (1 - \alpha_1 L) y_t = (1 - \alpha_1 L)^{-1} (\alpha_0 + e_t). \quad (14.26)$$

What is the operator  $(1 - \alpha_1 L)^{-1}$ ? Recall from Theorem 14.4.1 that for  $|x| < 1$ ,

$$\sum_{j=0}^{\infty} x^j = \frac{1}{1-x} = (1-x)^{-1}.$$

Now evaluate this expression at  $x = \alpha_1 z$ . We find

$$(1 - \alpha_1 z)^{-1} = \sum_{j=0}^{\infty} \alpha_1^j z^j. \quad (14.27)$$

Setting  $z = L$  this is

$$(1 - \alpha_1 L)^{-1} = \sum_{j=0}^{\infty} \alpha_1^j L^j.$$

Substituted into (14.26) we obtain

$$\begin{aligned}y_t &= (1 - \alpha_1 L)^{-1} (\alpha_0 + e_t) \\ &= \left( \sum_{j=0}^{\infty} \alpha_1^j L^j \right) (\alpha_0 + e_t) \\ &= \sum_{j=0}^{\infty} \alpha_1^j L^j (\alpha_0 + e_t) \\ &= \sum_{j=0}^{\infty} \alpha_1^j (\alpha_0 + e_{t-j}) \\ &= \frac{\alpha_0}{1 - \alpha_1} + \sum_{j=0}^{\infty} \alpha_1^j e_{t-j}\end{aligned}$$

which is (14.24). This is valid for  $|\alpha_1| < 1$ .

This illustrates another important concept. We say that a polynomial  $\alpha(z)$  is **invertible** if it can be written as

$$\alpha(z)^{-1} = \sum_{j=0}^{\infty} a_j z^j$$

and is absolutely convergent. In particular, we have learned that the AR(1) autoregressive polynomial  $\alpha(z) = 1 - \alpha_1 z$  is invertible if  $|\alpha_1| < 1$ . This is the same condition as for stationarity of the AR(1) process. Invertibility turns out to be a very useful property.

## 14.22 Unit Root and Explosive AR(1) Processes

The AR(1) process (14.23) is stationary if  $|\alpha| < 1$ . What happens otherwise?

If  $\alpha_0 = 0$  and  $\alpha_1 = 1$  the model is known as a **random walk**.

$$y_t = y_{t-1} + e_t.$$

This is also called a unit root process, a martingale, or an integrated process. By back-substitution we find

$$y_t = y_0 + \sum_{j=1}^t e_j.$$

Thus the initial condition does not disappear for large  $t$ . Consequently the series is non-stationary. The autoregressive polynomial  $\alpha(z) = 1 - z$  is not invertible, meaning that  $y_t$  cannot be written as a convergent function of the infinite past history of  $e_t$ .

The stochastic behavior of a random walk is noticeably different from a stationary AR(1) process. It wanders up and down with equal likelihood, and is not mean-reverting. While it has no tendency to return to its previous values, the wandering nature of a random walk can give the illusion of mean reversion. The difference is that a random walk will take a very large number of time periods to “revert”.

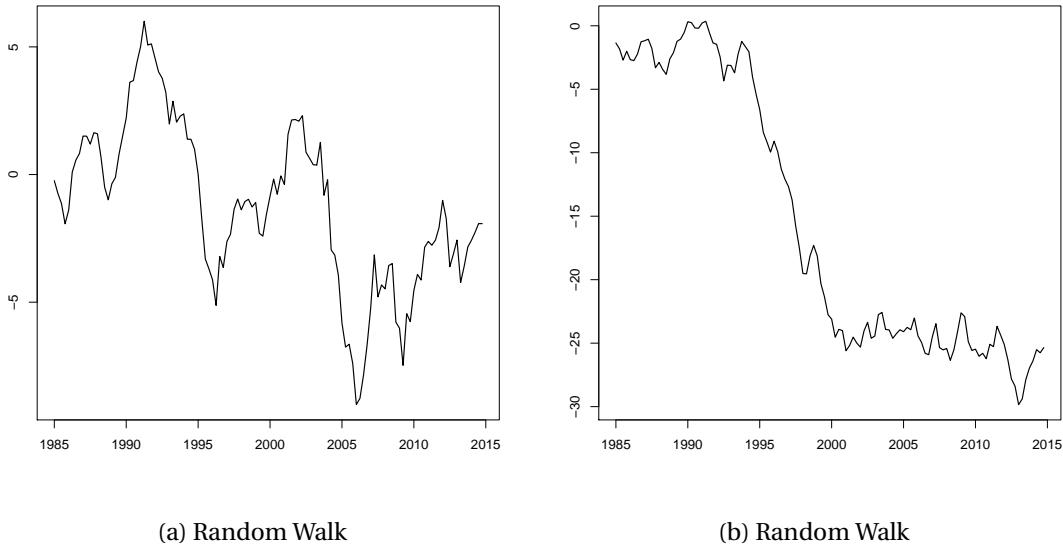


Figure 14.7: Random Walk Processes

To illustrate, Figure 14.7 plots two independent random walk processes. The plot in panel (a) uses the innovations from Figure 14.5(a). The plot in panel (b) uses an independent set of i.i.d.  $N(0, 1)$  errors. You can see that the plot in panel (a) appears similar to the MA(8) and AR(1) plots in the sense that the series is smooth with long swings, but the difference is that the series does not return to a long-term mean. It appears to have drifted down over time. The plot in panel (b) appears to have quite different behavior, falling dramatically over a 5-year period, and then appearing to stabilize. These are both common behaviors of random walk processes.

If  $\alpha_1 > 1$  the process is **explosive**. The model (14.23) with  $\alpha_1 > 1$  exhibits exponential growth, and high sensitivity to initial conditions. Explosive autoregressive processes do not seem to be good descriptions for most economic time series. While aggregate time series such as the GDP process displayed in Figure 14.1(a) exhibit a similar exponential growth pattern, the exponential growth can typically be removed by taking logarithms.

The case  $\alpha_1 < -1$  induces explosive oscillating growth and does not appear to be empirically relevant for economic applications.

## 14.23 Second-Order Autoregressive Process

The **second-order autoregressive process**, denoted AR(2), is

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + e_t \quad (14.28)$$

where  $e_t$  is a strictly stationary and ergodic white noise process. The dynamic patterns of an AR(2) process are more complicated than an AR(1) process.

As a motivating example consider the multiplier-accelerator model of Samuelson (1939). It might be a bit dated as a model, but it is simple so hopefully makes the point. Aggregate output (in an economy with no trade) is defined as  $Y_t = Consumption_t + Investment_t + Gov_t$ . Suppose that individuals make their consumption decisions on the previous period's income  $Consumption_t = bY_{t-1}$ , firms make their investment decisions on the change in consumption  $Investment_t = d\Delta C_t$ , and government spending is random,  $Gov_t = a + e_t$ . Then aggregate output follows

$$Y_t = a + b(1 + d)Y_{t-1} - bdY_{t-2} + e_t \quad (14.29)$$

which is an AR(2) process.

Using the lag operator we can write (14.28) as

$$y_t - \alpha_1 L y_t - \alpha_2 L^2 y_t = \alpha_0 + e_t,$$

or

$$\alpha(L)y_t = \alpha_0 + e_t$$

where

$$\alpha(L) = 1 - \alpha_1 L - \alpha_2 L^2.$$

We call  $\alpha(z)$  the autoregressive polynomial of  $y_t$ .

We would like to describe the conditions for the stationarity of  $y_t$ . For simplicity set  $\alpha_0 = 0$ . Factor the autoregressive polynomial as

$$\alpha(z) = 1 - \alpha_1 z - \alpha_2 z^2 = (1 - \beta_1 z)(1 - \beta_2 z)$$

which holds for

$$\beta_j = \frac{\alpha_1 \pm \sqrt{\alpha_1^2 + 4\alpha_2}}{2}. \quad (14.30)$$

These factors are real if  $\alpha_1^2 - 4\alpha_2 \geq 0$  but are complex conjugates otherwise. Equating the factors, we can see that  $\alpha_1 = \beta_1 + \beta_2$  and  $\alpha_2 = \beta_1 \beta_2$ .

The autoregressive polynomial  $\alpha(z)$  is invertible when the polynomials  $(1 - \beta_1 z)$  and  $(1 - \beta_2 z)$  are invertible. In the previous section we discovered that this occurs when  $|\beta_1| < 1$  and  $|\beta_2| < 1$ . Under these conditions the inverse equals

$$\alpha(z)^{-1} = (1 - \beta_2 z)^{-1}(1 - \beta_1 z)^{-1}.$$

Consequently

$$y_t = (1 - \beta_2 L)^{-1}(1 - \beta_1 L)^{-1}e_t.$$

If  $|\beta_1| < 1$ , by Theorem 14.24 the series  $u_t = (1 - \beta_1 L)^{-1}e_t$  is a convergent AR(1) process. Furthermore, when  $|\beta_2| < 1$  the series  $y_t = (1 - \beta_2 L)^{-1}u_t$  is also convergent, by Theorem 14.3. Thus sufficient conditions for stationarity are  $|\beta_1| < 1$  and  $|\beta_2| < 1$ .

With some algebra, we can show that  $|\beta_1| < 1$  and  $|\beta_2| < 1$  iff the following restrictions hold on the autoregressive coefficients:

$$\alpha_1 + \alpha_2 < 1 \quad (14.31)$$

$$\alpha_2 - \alpha_1 < 1 \quad (14.32)$$

$$\alpha_2 > -1 \quad (14.33)$$

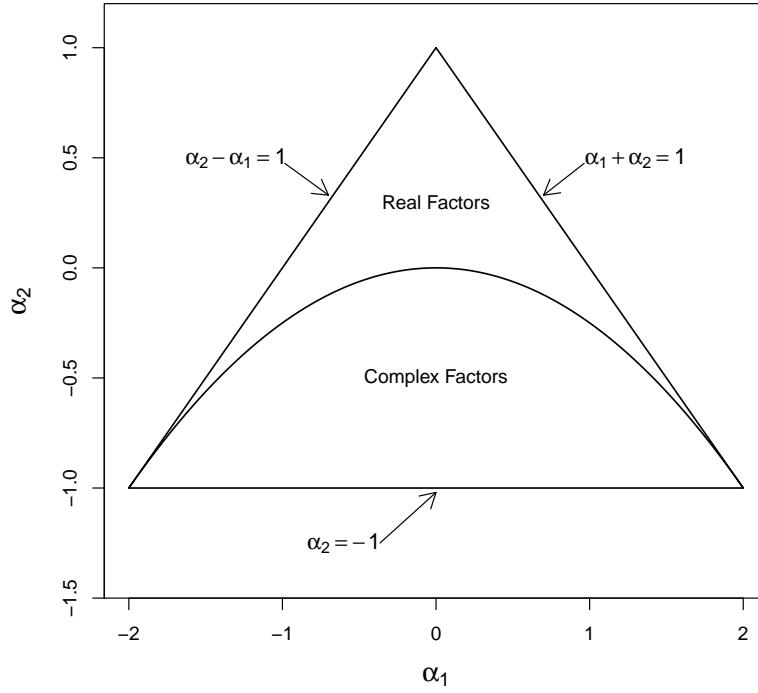


Figure 14.8: Stationarity Region for AR(2)

These restrictions describe a triangle in  $(\alpha_1, \alpha_2)$  space. This region is shown in Figure 14.8. Coefficients within this triangle correspond to a stationary AR(2) process.

Furthermore, the triangle is divided into two regions, the region above the parabola  $\alpha_1^2 - 4\alpha_2 = 0$  producing real factors  $\beta_j$ , and the region below the parabola producing complex factors  $\beta_j$ . These two regions are marked in Figure 14.8. This is potentially interesting because when the factors are complex the autocorrelations of  $y_t$  display damped oscillations. For this reason, the dynamic patterns of an AR(2) can be much more complicated than those of an AR(1).

Take, for example, the Samuelson multiplier-accelerator model (14.29). You can calculate that this model has complex factors (and thus oscillations) for certain values of  $b$  and  $d$ , including  $b \leq 0.8$  and  $d \geq 0.4$ .

**Theorem 14.25** If  $|\beta_j| < 1$  for  $\beta_j$  defined in (14.30), or equivalently if the inequalities (14.31)-(14.33) hold, then the AR(2) process (14.28) is absolutely convergent, strictly stationary, and ergodic.

The proof is presented in Section 14.46.

To illustrate, Figure 14.9 displays two simulated AR(2) processes. The plot in panel (a) sets  $\alpha_1 = \alpha_2 = 0.4$ . These coefficients produce real factors so the process displays behavior similar to that of the AR(1) processes. The plot in panel (b) sets  $\alpha_1 = 1.3$  and  $\alpha_2 = -0.8$ . These coefficients produce complex factors so the process displays oscillations.

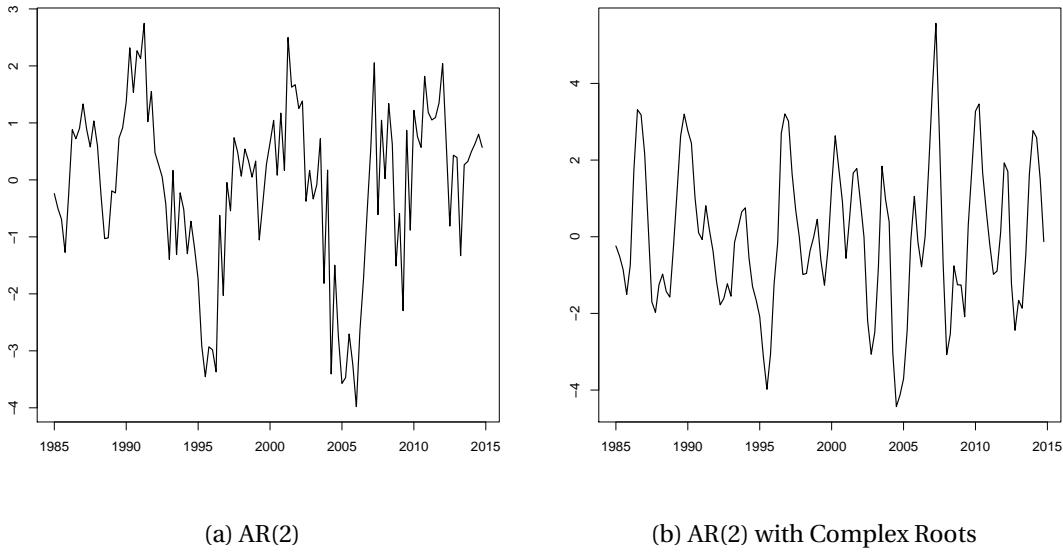


Figure 14.9: AR(2) Processes

## 14.24 AR(p) Processes

The  $p^{th}$ -order autoregressive process, denoted AR(p), is

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + e_t \quad (14.34)$$

where  $e_t$  is a strictly stationary and ergodic white noise process.

Using the lag operator,

$$y_t - \alpha_1 L y_t - \alpha_2 L^2 y_t - \cdots - \alpha_p L^p y_t = \alpha_0 + e_t,$$

or

$$\alpha(L)y_t = \alpha_0 + e_t$$

where

$$\alpha(L) = 1 - \alpha_1 L - \alpha_2 L^2 - \cdots - \alpha_p L^p. \quad (14.35)$$

We call  $\alpha(z)$  the autoregressive polynomial of  $y_t$ .

The **Fundamental Theorem of Algebra** states that any polynomial can be factored as

$$\alpha(z) = (1 - \beta_1 z)(1 - \beta_2 z) \cdots (1 - \beta_p z) \quad (14.36)$$

where the factors  $\beta_j$  can be real or in complex conjugate pairs. If  $|\beta_j| < 1$  then the polynomials  $(1 - \beta_j z)$  are invertible and thus so is  $\alpha(z)$ . The inverse is

$$\alpha(z)^{-1} = (1 - \beta_1 z)^{-1} (1 - \beta_2 z)^{-1} \cdots (1 - \beta_p z)^{-1}. \quad (14.37)$$

Consequently

$$\begin{aligned} y_t &= \alpha(L)^{-1} (\alpha_0 + e_t) \\ &= (1 - \beta_p L)^{-1} \cdots (1 - \beta_2 L)^{-1} (1 - \beta_1 L)^{-1} (\alpha_0 + e_t) \\ &= \mu + (1 - \beta_p L)^{-1} \cdots (1 - \beta_2 L)^{-1} (1 - \beta_1 L)^{-1} e_t \end{aligned} \quad (14.38)$$

where

$$\mu = \frac{\alpha_0}{1 - \alpha_1 - \cdots - \alpha_p}.$$

The series  $u_{1t} = (1 - \beta_1 L)^{-1} (\alpha_0 + e_t)$  is a strictly stationary and ergodic AR(1) process by Theorem 14.24. By induction, the series  $u_{jt} = (1 - \beta_j L)^{-1} u_{j-1,t}$  is strictly stationary and ergodic by Theorem 14.10. Thus  $y_t$  is strictly stationary and ergodic.

In general, we do not have explicit expressions for the factors  $\beta_j$  (though they can be calculated numerically from the coefficients). Instead, the following characterization may be insightful. Take the inverse factors  $\lambda_j = \beta_j^{-1}$ . Since  $1 - \beta_j \lambda_j = 0$ , then  $\alpha(\lambda_j) = 0$ . This means that  $\lambda_j$  is a **root** of the polynomial  $\alpha(z)$  (the point on the  $x$ -axis where the polynomial hits zero). The requirement  $|\beta_j| < 1$  is the same as  $|\lambda_j| > 1$ . We find that the following three conditions are equivalent.

1.  $|\beta_j| < 1$  for  $j = 1, \dots, p$ .
2. All roots  $\lambda_j$  of  $\alpha(z)$  satisfy  $|\lambda_j| > 1$ .
3.  $\alpha(z) \neq 0$  for all complex numbers  $z$  such that  $|z| \leq 1$ .

For complex numbers  $z$ , the equation  $|z| = 1$  defines the **unit circle** (the circle with radius of unity), the region  $|z| \leq 1$  is the interior of the unit circle, and the region  $|z| > 1$  is the exterior of the unit circle. We have established the following.

**Theorem 14.26** If all roots of  $\alpha(z)$  lie outside the unit circle, then the AR(p) process (14.34) is absolutely convergent, strictly stationary, and ergodic.

Thus to check if a specific autoregressive process satisfies the conditions for stationarity, we can (numerically) compute the roots  $\lambda_j$  of the autoregressive polynomial, calculate their modulus  $|\lambda_j|$  and check if  $|\lambda_j| > 1$ .

The equation (14.38) can be written as

$$y_t = \mu + b(L)e_t$$

where

$$b(z) = \alpha(z)^{-1} = \sum_{j=0}^{\infty} b_j z^j. \quad (14.39)$$

We have the following characterization of the moving average coefficients.

**Theorem 14.27** If all roots of the autoregressive polynomial  $\alpha(z)$  lie outside the unit circle then (14.39) holds with  $b_j = O(j^p \beta^j)$  and  $\sum_{j=0}^{\infty} |b_j| < \infty$ .

The proof is presented in Section 14.46.

## 14.25 Impulse Response Function

The coefficients of the moving average representation

$$\begin{aligned} y_t &= b(L)e_t \\ &= \sum_{j=0}^{\infty} b_j e_{t-j} \\ &= b_0 e_t + b_1 e_{t-1} + b_2 e_{t-2} + \dots \end{aligned}$$

are known among economists as the **impulse response function (IRF)**. (Often, scaled by the standard deviation of  $e_t$ . We discuss this scaling at the end of the section.) In linear models the impulse response function is defined as the change in  $y_{t+j}$  due to a shock at time  $t$ . This is

$$\frac{\partial}{\partial e_t} y_{t+j} = b_j.$$

This means that the coefficients  $b_j$  can be interpreted as the magnitude of the impact of a time  $t$  shock on the time  $t + j$  variable. Plots of  $b_j$  can then be used to assess the time-propagation of shocks. This is a standard method of analysis for multivariate time series.

It is desirable to have a convenient method to calculate the impulse responses  $b_j$  from the coefficients of an autoregressive model (14.34). There are two methods which we now describe.

The first uses a simple recursion. In the linear AR(p) model, we can see that the coefficient  $b_j$  is the simple derivative

$$b_j = \frac{\partial}{\partial e_t} y_{t+j} = \frac{\partial}{\partial e_0} y_j$$

We can therefore calculate  $b_j$  by generating a history and perturbing the shock  $e_0$ . Since this calculation is unaffected by all other shocks, we can simply set  $e_t = 0$  for  $t \neq 0$  and set  $e_0 = 1$ . This implies the recursion

$$\begin{aligned} b_0 &= 1 \\ b_1 &= \alpha_1 b_0 \\ b_2 &= \alpha_1 b_1 + \alpha_2 b_0 \\ &\vdots \\ b_j &= \alpha_1 b_{j-1} + \alpha_2 b_{j-2} + \cdots + \alpha_p b_{j-p}. \end{aligned}$$

Equivalently, this is achieved by the following simulation. Set  $y_t = 0$  for  $t \leq 0$ . Set  $e_0 = 1$  and  $e_t = 0$  for  $t > 1$ . Generate  $y_t$  for  $t \geq 0$  by  $y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + e_t$ . Then  $y_j = b_j$ .

A second method uses a vector representation of the AR(p) model. Let  $\tilde{\mathbf{y}}_t = (y_t, \dots, y_{t-p+1})'$  and  $\tilde{\mathbf{e}}_t = (e_t, 0, \dots, 0)'$ . Then

$$\begin{aligned} \tilde{\mathbf{y}}_t &= \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{p-1} & \alpha_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & & 0 & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \tilde{\mathbf{y}}_{t-1} + \tilde{\mathbf{e}}_t \\ &= \mathbf{A} \tilde{\mathbf{y}}_{t-1} + \tilde{\mathbf{e}}_t. \end{aligned} \tag{14.40}$$

By recursion

$$\tilde{\mathbf{y}}_t = \sum_{j=0}^{\infty} \mathbf{A}^j \tilde{\mathbf{e}}_{t-j}.$$

Here,  $\mathbf{A}^j = \mathbf{A} \cdots \mathbf{A}$  means the  $j^{th}$  matrix product of  $\mathbf{A}$  with itself. Setting  $S = (1, 0, \dots, 0)'$  we find

$$y_t = \sum_{j=0}^{\infty} S' \mathbf{A}^j S e_{t-j}.$$

By linearity

$$b_j = \frac{\partial}{\partial e_t} y_{t+j} = S' \mathbf{A}^j S. \tag{14.41}$$

Thus the coefficient  $b_j$  can be calculated by forming the matrix  $\mathbf{A}$ , its  $j$ -fold product  $\mathbf{A}^j$ , and then taking the upper-left element.

As mentioned at the beginning of the section, it is often desirable to define the IRF to be scaled so that it is the response to a one-deviation shock. Let  $\sigma^2 = \text{var}(e_t)$  and define  $\varepsilon_t = e_t/\sigma$  which has unit variance. Then the IRF at lag  $j$  is

$$\text{IRF}_j = \frac{\partial}{\partial \varepsilon_t} y_{t+j} = \sigma b_j.$$

## 14.26 ARMA and ARIMA Processes

The **autoregressive-moving-average process**, denoted **ARMA(p,q)**, is

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + \theta_0 e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q}$$

where  $e_t$  is a strictly stationary and ergodic white noise process. It can be written using lag operator notation as

$$\alpha(L)y_t = \alpha_0 + \theta(L)e_t.$$

**Theorem 14.28** The ARMA(p,q) process (14.34) is strictly stationary and ergodic if all roots of  $\alpha(z)$  lie outside the unit circle. In this case we can write

$$y_t = \mu + b(L)e_t$$

where  $b_j = O(j^p \beta^j)$  and  $\sum_{j=0}^{\infty} |b_j| < \infty$ .

The process  $y_t$  follows an **autoregressive-integrated moving-average process**, denoted **ARIMA(p,d,q)**, if  $\Delta^d y_t$  is ARMA(p,q). It can be written using lag operator notation as

$$\alpha(L)(1-L)^d y_t = \alpha_0 + \theta(L)e_t.$$

## 14.27 Mixing Properties of Linear Processes

There is a considerable probability literature investigating the mixing properties of time series processes. One challenge is that since autoregressive processes depend on the infinite past sequence of innovations  $e_t$  it is not immediately obvious if they satisfy the mixing conditions.

In fact, a simple AR(1) is not necessarily mixing. A counter-example was developed by Andrews (1984). He showed that if the error  $e_t$  has a two-point discrete distribution, then an AR(1)  $y_t$  is not strong mixing. The reason is that a discrete innovation combined with the autoregressive structure means that by observing  $y_t$  you can deduce with near certainty the past history of the shocks  $e_t$ . The example seems rather special, but shows the need to be careful with the theory. The intuition stemming from Andrews' finding is that for an autoregressive process to be mixing it is necessary for the errors  $e_t$  to not be discrete.

A useful characterization was provided by Pham and Tran (1985).

**Theorem 14.29** Suppose that  $y_t = \mu + \sum_{j=0}^{\infty} \theta_j e_{t-j}$  satisfies the following conditions:

1.  $e_t$  is i.i.d. with  $\mathbb{E}|e_t|^r < \infty$  for some  $r > 0$  and density  $f(x)$  which satisfies

$$\int |f(x-u) - f(x)| dx \leq C|u| \quad (14.42)$$

for some  $C < \infty$ .

2. All roots of  $\theta(z)$  lie outside the unit circle and  $\sum_{j=0}^{\infty} |\theta_j| < \infty$ .
3.  $\sum_{k=1}^{\infty} \left( \sum_{j=k}^{\infty} |\theta_j| \right)^{r/(1+r)} < \infty$ .

Then for some  $B < \infty$

$$\alpha(\ell) \leq 4\beta(\ell) \leq B \sum_{k=\ell}^{\infty} \left( \sum_{j=k}^{\infty} |\theta_j| \right)^{r/(1+r)}$$

and  $y_t$  is absolutely regular and strong mixing.

The condition (14.42) is rather unusual, but specifies that  $e_t$  has a smooth density. This rules out the counter-example discovered by Andrews (1984).

The summability condition on the coefficients in part 3 involves a trade-off with the number of moments  $r$ . If  $e_t$  has all moments finite (e.g. normal errors) then we can set  $r = \infty$  and this condition simplifies to  $\sum_{k=1}^{\infty} k|\theta_k| < \infty$ . For any  $r$ , the summability condition holds if  $\theta_j$  has geometric decay, as holds for an finite-order AR(p) process.

It is instructive to deduce how the decay in the coefficients  $\theta_j$  affects the rate for the mixing coefficients  $\alpha(\ell)$ . If  $|\theta_j| \leq O(j^{-\eta})$  then  $\sum_{j=k}^{\infty} |\theta_j| \leq O(k^{-(\eta-1)})$  so the rate is

$$\begin{aligned} \alpha(\ell) &\leq 4\beta(\ell) \leq O(\ell^{-s}) \\ s &= (\eta-1) \left( \frac{r}{1+r} \right) - 1. \end{aligned}$$

Mixing requires  $s > 0$ , which holds for sufficiently large  $\eta$ . For example, if  $r = 4$  it holds for  $\eta > 9/4$ .

The primary message from this section is that linear processes, including autoregressive and ARMA processes, are mixing, if the innovations satisfy suitable conditions. The mixing coefficients decay at rates related to the decay rates of the moving average coefficients.

## 14.28 Identification

The parameters of a model are identified if the parameters are uniquely determined by the probability distribution of the observations. In the case of linear time series analysis we typically focus on the second moments of the observations (means, variances, covariances). We therefore say that the coefficients of a stationary MA, AR, or ARMA model are **identified** if they are uniquely determined by the autocorrelation function. That is, given the autocorrelation function  $\rho(k)$ , are the coefficients unique?

It turns out that the answer is that MA and ARMA models are generally not identified. Identification is achieved by restricting the class of polynomial operators. In contrast, AR models are generally identified.

Let us start with the MA(1) model

$$y_t = e_t + \theta e_{t-1}.$$

It has first-order autocorrelation

$$\rho(1) = \frac{\theta}{1 + \theta^2}.$$

Set  $\omega = 1/\theta$ . Then

$$\frac{\omega}{1 + \omega^2} = \frac{1/\omega}{1 + (1/\omega)^2} = \frac{\theta}{1 + \theta^2} = \rho(1)$$

Thus the MA(1) model with coefficient  $\omega = 1/\theta$  produces the same autocorrelations as the MA(1) model with coefficient  $\theta$ . For example,  $\theta = 1/2$  and  $\omega = 2$  each yield  $\rho(1) = 2/5$ . There is no empirical way to distinguish between the models  $y_t = e_t + \theta e_{t-1}$  and  $y_t = e_t + \omega e_{t-1}$ . Thus the coefficient  $\theta$  is not identified.

The standard solution is to select the parameter which produce an invertible moving average polynomial. Since there is only one such choice this yields a unique solution. This may be sensible when there is reason to believe that shocks have their primary impact in the contemporaneous period, and secondary (lesser) impact in the second period.

Now consider the MA(2) model

$$y_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}.$$

The moving average polynomial can be factored as

$$\theta(z) = (1 - \beta_1 z)(1 - \beta_2 z)$$

so that  $\beta_1 \beta_2 = \theta_2$  and  $\beta_1 + \beta_2 = -\theta_1$ . The process has first- and second-order autocorrelations

$$\begin{aligned}\rho(1) &= \frac{\theta_1 + \theta_1 \theta_2}{1 + \theta_1^2 + \theta_2^2} = \frac{-\beta_1 - \beta_2 - \beta_1^2 \beta_2 - \beta_1 \beta_2^2}{1 + \beta_1^2 + \beta_2^2 + 2\beta_1 \beta_2 + \beta_1^2 \beta_2^2} \\ \rho(2) &= \frac{\theta_2}{1 + \theta_1^2 + \theta_2^2} = \frac{\beta_1 \beta_2}{1 + \beta_1^2 + \beta_2^2 + 2\beta_1 \beta_2 + \beta_1^2 \beta_2^2}.\end{aligned}$$

If we replace  $\beta_1$  with  $\omega_1 = 1/\beta_1$  we obtain

$$\begin{aligned}\rho(1) &= \frac{-1/\beta_1 - \beta_2 - \beta_2/\beta_1^2 - \beta_2^2/\beta_1}{1 + 1/\beta_1^2 + \beta_2^2 + 2\beta_2/\beta_1 + \beta_2^2/\beta_1^2} = \frac{-\beta_1 - \beta_2 \beta_1^2 - \beta_2 - \beta_2^2 \beta_1}{\beta_1^2 + 1 + \beta_2^2 \beta_1^2 + 2\beta_2 \beta_1 + \beta_2^2} \\ \rho(2) &= \frac{\beta_2/\beta_1}{1 + 1/\beta_1^2 + \beta_2^2 + 2\beta_2/\beta_1 + \beta_2^2/\beta_1^2} = \frac{\beta_1 \beta_2}{\beta_1^2 + 1 + \beta_1^2 \beta_2^2 + 2\beta_1 \beta_2 + \beta_2^2}\end{aligned}$$

which is unchanged. Similarly if we replace  $\beta_2$  with  $\omega_2 = 1/\beta_2$  we obtain unchanged first- and second-order autocorrelations. It follows that in the MA(2) model, the factors  $\beta_1$  and  $\beta_2$  are not identified. It follows that the coefficients  $\theta_1$  and  $\theta_2$  are not identified. Consequently there are four distinct MA(2) models which are identifiably indistinguishable.

This analysis extends to the MA( $q$ ) model. The factors of the MA polynomial can be replaced by their inverses, and consequently the coefficients are not identified.

The standard solution is to confine attention to MA( $q$ ) models with invertible roots. This technically solves the identification dilemma. One reason why this choice may be considered reasonable is because this corresponds to the Wold decomposition. This is because the Wold decomposition is defined in terms of the projection errors, which correspond to the invertible representation.

A deeper identification failure occurs in ARMA models. Consider an ARMA(1,1) model

$$y_t = \alpha y_{t-1} + e_t + \theta e_{t-1}.$$

Written in lag operator notation

$$(1 - \alpha L) y_t = (1 + \theta L) e_t.$$

The identification failure is that when  $\alpha = -\theta$  then the model simplifies to

$$y_t = e_t.$$

This means that the continuum of models with  $\alpha = -\theta$  are all identical and hence the coefficients are not identified.

This extends to higher order ARMA models. Take the ARMA(2,2) model written in factored lag operator notation

$$(1 - \alpha_1 L)(1 - \alpha_2 L)y_t = (1 + \theta_1 L)(1 + \theta_2 L)e_t.$$

Here we see that the models with  $\alpha_1 = -\theta_1$ ,  $\alpha_1 = -\theta_2$ ,  $\alpha_2 = -\theta_1$ , or  $\alpha_2 = -\theta_2$  all simplify to an ARMA(1,1) model. Thus all these models are identical and hence the coefficients are not identified.

The problem is called “cancelling roots” due to the fact that it arises when there are two identical lag polynomial factors in the AR and MA polynomials.

The standard solution in the ARMA literature is to *assume* that there are no cancelling roots. The trouble with this solution is that this is an assumption about the true process, which is unknown. Thus it is not really a solution to the identification problem. One recommendation is to be careful when using ARMA models, and be aware that highly parameterized models may not have unique coefficients.

Now consider the AR(p) model (14.34). It can be written as

$$y_t = \mathbf{x}'_t \boldsymbol{\alpha} + e_t \quad (14.43)$$

where  $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \dots, \alpha_p)'$  and  $\mathbf{x}_t = (1, y_{t-1}, \dots, y_{t-p})'$ . The MDS assumption implies that  $\mathbb{E}(e_t) = 0$ ,  $\mathbb{E}(y_{t-j}e_t) = 0$ , and hence

$$\mathbb{E}(\mathbf{x}_t e_t) = 0.$$

This means that from our standard analysis of projection models the coefficient  $\boldsymbol{\alpha}$  satisfies

$$\boldsymbol{\alpha} = (\mathbb{E}(\mathbf{x}_t \mathbf{x}'_t))^{-1} (\mathbb{E}(\mathbf{x}_t y_t)). \quad (14.44)$$

This equation is unique if  $\mathbf{Q} = \mathbb{E}(\mathbf{x}_t \mathbf{x}'_t)$  is positive definite. It turns out that this is generically true, so  $\boldsymbol{\alpha}$  is unique and identified.

**Theorem 14.30** In the AR(p) model (14.34), if  $0 < \sigma^2 < \infty$  then  $\mathbf{Q} > 0$  and  $\boldsymbol{\alpha}$  is unique and identified.

The assumption  $\sigma^2 > 0$  means that  $y_t$  is not purely deterministic.

We can extend this result to approximating AR(p) models. That is, consider the equation (14.43) without the assumption that  $y_t$  is necessarily a true AR(p) with a MDS error. Instead, suppose that  $y_t$  is a non-deterministic stationary process. (Recall, non-deterministic means that  $\sigma^2 > 0$  where  $\sigma^2$  is the projection error variance (14.19).) We then define the coefficient  $\boldsymbol{\alpha}$  as the best linear predictor, which is (14.44). The error  $e_t$  is then defined by the equation (14.43). This is a linear projection model.

As in the case of any linear projection, the error  $e_t$  satisfies  $\mathbb{E}(\mathbf{x}_t e_t) = 0$ . This means that  $\mathbb{E}(e_t) = 0$  and  $\mathbb{E}(y_{t-j}e_t) = 0$  for  $j = 1, \dots, p$ . However, the error  $e_t$  is not necessarily a MDS nor white noise.

The coefficient  $\boldsymbol{\alpha}$  is identified if  $\mathbf{Q} > 0$ . The proof of Theorem 14.30 (presented in Section 14.46) does not make use of the assumption that  $y_t$  is an AR(p) with a MDS error. Rather, it only uses the assumption that  $\sigma^2 > 0$ . This holds in the approximate AR(p) model as well under the assumption that  $y_t$  is non-deterministic. We conclude that any approximating AR(p) is identified.

**Theorem 14.31** If  $y_t$  is strictly stationary, not purely deterministic, and  $\mathbb{E}(y_t^2) < \infty$ , then for any  $p$ ,  $\mathbf{Q} = \mathbb{E}(\mathbf{x}_t \mathbf{x}'_t) > 0$  and thus the coefficient vector (14.44) is identified.

## 14.29 Estimation of Autoregressive Models

We consider estimation of an AR(p) model for stationary, ergodic, and non-deterministic  $y_t$ . The model is

$$y_t = \mathbf{x}'_t \boldsymbol{\alpha} + e_t \quad (14.45)$$

where  $\mathbf{x}_t = (1, y_{t-1}, \dots, y_{t-p})'$ . The coefficient  $\boldsymbol{\alpha}$  is defined by projection in (14.44). The error is defined by (14.45), and has variance  $\sigma^2 = \mathbb{E}(e_t^2)$ . This allows  $y_t$  to follow a true AR(p) process, but it is not necessary.

The least squares estimator of the AR(p) model is

$$\hat{\boldsymbol{\alpha}} = \left( \sum_{t=1}^n \mathbf{x}_t \mathbf{x}'_t \right)^{-1} \left( \sum_{t=1}^n \mathbf{x}_t y_t \right).$$

This notation presumes that there are  $n + p$  observations on  $y_t$ , from which the first  $p$  are used as initial conditions so that  $\mathbf{x}_1 = (1, y_0, y_{-1}, \dots, y_{-p+1})$  is defined. Effectively, this redefines the sample period. (An alternative notational choice is to define the estimator to have the sums range from observations  $p + 1$  to  $n$ .)

The least squares residuals are

$$\hat{e}_t = y_t - \mathbf{x}'_t \hat{\boldsymbol{\alpha}}.$$

The error variance can be estimated by

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{t=1}^n \hat{e}_t^2$$

or

$$s^2 = \frac{1}{n-p-1} \sum_{t=1}^n \hat{e}_t^2.$$

If  $y_t$  is strictly stationary and ergodic, then so are  $\mathbf{x}_t \mathbf{x}'_t$  and  $\mathbf{x}_t y_t$ . They have finite means if  $\mathbb{E}(y_t^2) < \infty$ . Under these assumptions the Ergodic Theorem implies that

$$\frac{1}{n} \sum_{t=1}^n \mathbf{x}_t y_t \xrightarrow{p} \mathbb{E}(\mathbf{x}_t y_t) \quad (14.46)$$

and

$$\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t \xrightarrow{p} \mathbb{E}(\mathbf{x}_t \mathbf{x}'_t) = \mathbf{Q}.$$

Theorem 14.31 shows that  $\mathbf{Q} > 0$ . Combined with the continuous mapping theorem, we see that

$$\hat{\boldsymbol{\alpha}} = \left( \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t \right)^{-1} \left( \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t y_t \right) \xrightarrow{p} (\mathbb{E}(\mathbf{x}_t \mathbf{x}'_t))^{-1} (\mathbb{E}(\mathbf{x}_t y_t)) = \boldsymbol{\alpha}.$$

It is straightforward to show that  $\hat{\sigma}^2$  is consistent as well.

**Theorem 14.32** If  $y_t$  is strictly stationary, ergodic, not purely deterministic, and  $\mathbb{E}(y_t^2) < \infty$ , then for any  $p$ ,  $\hat{\boldsymbol{\alpha}} \xrightarrow{p} \boldsymbol{\alpha}$  and  $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$  as  $n \rightarrow \infty$ .

This shows that under very mild conditions, the coefficients of an AR(p) model can be consistently estimated by least squares. Once again, this does not require that the series  $y_t$  is actually an AR(p) process. It holds for any stationary process with the coefficient defined by projection.

### 14.30 Asymptotic Distribution of Least Squares Estimator

The asymptotic distribution of the least squares estimator  $\hat{\alpha}$  depends on the stochastic assumptions. In this section we derive the asymptotic distribution under the assumption of correct specification.

Specifically, we assume that the error  $e_t$  is a MDS. An important implication of the MDS assumption is that since  $\mathbf{x}_t = (1, y_{t-1}, \dots, y_{t-p})'$  is part of the information set  $\mathcal{F}_{t-1}$ , by the conditioning theorem,

$$\mathbb{E}(\mathbf{x}_t e_t | \mathcal{F}_{t-1}) = \mathbf{x}_t \mathbb{E}(e_t | \mathcal{F}_{t-1}) = 0.$$

Thus  $\mathbf{x}_t e_t$  is a MDS. It has a finite variance if  $\mathbf{x}_t$  and  $e_t$  have finite fourth moments, which holds if  $y_t$  does. We can then apply the martingale difference CLT (Theorem 14.15) to see that

$$\frac{1}{\sqrt{n}} \sum_{t=1}^n \mathbf{x}_t e_t \xrightarrow{d} N(\mathbf{0}, \Sigma)$$

where

$$\Sigma = \mathbb{E}(\mathbf{x}_t \mathbf{x}_t' e_t^2).$$

**Theorem 14.33** If  $y_t$  follows the AR(p) model (14.34) with  $\mathbb{E}(e_t | \mathcal{F}_{t-1}) = 0$ ,  $\mathbb{E}(y_t^4) < \infty$ , and  $\sigma^2 > 0$ , then as  $n \rightarrow \infty$ ,

$$\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{d} N(\mathbf{0}, V)$$

where

$$V = Q^{-1} \Sigma Q^{-1}.$$

This is identical in form to the asymptotic distribution of least squares in cross-section regression. The implication is that asymptotic inference is the same. In particular, the asymptotic covariance matrix is estimated just as in the cross-section case.

### 14.31 Distribution Under Homoskedasticity

In cross-section regression we found that the variance matrix simplifies under the assumption of conditional homoskedasticity. The same occurs in the time series context. Assume that the error is a homoskedastic MDS:

$$\begin{aligned}\mathbb{E}(e_t | \mathcal{F}_{t-1}) &= 0 \\ \mathbb{E}(e_t^2 | \mathcal{F}_{t-1}) &= \sigma^2.\end{aligned}$$

In this case

$$\Sigma = \mathbb{E}(\mathbf{x}_t \mathbf{x}_t' \mathbb{E}(e_t^2 | \mathcal{F}_{t-1})) = Q \sigma^2$$

and the asymptotic distribution simplifies.

**Theorem 14.34** If  $y_t$  follows the AR(p) model (14.34) with  $\mathbb{E}(e_t | \mathcal{F}_{t-1}) = 0$ ,  $\mathbb{E}(y_t^4) < \infty$ , and  $\mathbb{E}(e_t^2 | \mathcal{F}_{t-1}) = \sigma^2 > 0$ , then as  $n \rightarrow \infty$ ,

$$\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{d} N(\mathbf{0}, V^0)$$

where

$$V^0 = \sigma^2 Q^{-1}.$$

These results show that under correct specification (a MDS error) the format of the asymptotic distribution of the least squares estimator exactly parallels the cross-section case. In general the covariance matrix takes a sandwich form, with components exactly equal to the cross-section case. Under conditional homoskedasticity the covariance matrix simplifies exactly as in the cross-section case.

A particularly useful insight which can be derived from Theorem 14.34 is to focus on the simple AR(1) with no intercept. In this case  $Q = \mathbb{E}(y_t^2) = \sigma^2/(1 - \alpha_1^2)$  so the asymptotic distribution simplifies to

$$\sqrt{n}(\hat{\alpha}_1 - \alpha_1) \xrightarrow{d} N(0, 1 - \alpha_1^2).$$

Thus the asymptotic variance depends only on  $\alpha_1$  and is decreasing with  $\alpha_1^2$ . An intuition is that larger  $\alpha_1^2$  means greater signal and hence greater estimation precision. This result also shows that the asymptotic distribution is non-similar: the variance is a function of the parameter of interest. This means that we can expect (from advanced statistical theory) asymptotic inference to be less accurate than indicated by nominal levels.

In the context of cross-section data we argued that the homoskedasticity assumption was dubious except for occasional theoretical insight. For practical applications, it is recommended to use heteroskedasticity-robust theory and methods when possible. The same argument applies to the time series case. While the distribution theory simplifies under conditional homoskedasticity, there is no reason to expect homoskedasticity to hold in practice. Therefore in applications it is better to use the heteroskedasticity-robust distributional theory when possible.

Unfortunately, many existing time series textbooks report the distribution theory from (14.34). This has influenced computer software packages, many of which also by default (or exclusively) use the homoskedastic distribution theory. This is unfortunate.

## 14.32 Asymptotic Distribution Under General Dependence

If the AR(p) model (14.34) holds with white noise errors, or if the AR(p) is an approximation with  $\boldsymbol{\alpha}$  defined as the best linear predictor, then the MDS central limit theory does not apply. Instead, if  $y_t$  is strong mixing we can use the central limit theory for mixing processes (Theorem 14.19).

**Theorem 14.35** If  $y_t$  is strictly stationary, ergodic,  $0 < \sigma^2 < \infty$ , and for some  $r > 4$ ,  $\mathbb{E}|y_t|^r < \infty$  and the mixing coefficients satisfy  $\sum_{\ell=1}^{\infty} \alpha(\ell)^{1-4/r} < \infty$ , then

$$\boldsymbol{\Omega} = \sum_{\ell=-\infty}^{\infty} \mathbb{E}(\mathbf{x}_{t-\ell} \mathbf{x}'_{t-\ell} e_t e_{t-\ell})$$

is convergent, and for the AR(p) least squares estimator  $\hat{\boldsymbol{\alpha}}$  and projection coefficients (14.44),

$$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{d} N(\mathbf{0}, \mathbf{V})$$

as  $n \rightarrow \infty$ , where

$$\mathbf{V} = \mathbf{Q}^{-1} \boldsymbol{\Omega} \mathbf{Q}^{-1}.$$

This result is substantially different from the cross-section case. It shows that model misspecification (misspecifying the order of the autoregression, or missing proper specification of the conditional mean) renders invalid the conventional “heteroskedasticity-robust” covariance matrix formula. Misspecified models do not have unforecastable (martingale difference) errors, so the regression scores  $\mathbf{x}_t e_t$  are potentially serially correlated.

### 14.33 Covariance Matrix Estimation

Under the assumption of correct specification, covariance matrix estimation is identical to the cross-section case. The asymptotic covariance matrix estimator under the assumption of homoskedasticity is

$$\begin{aligned}\hat{V}^0 &= \hat{\sigma}^2 \hat{Q}^{-1} \\ \hat{Q} &= \frac{1}{n} \sum_{t=1}^n \mathbf{x}_t \mathbf{x}'_t\end{aligned}$$

The estimator  $s^2$  may be used instead of  $\hat{\sigma}^2$ .

The heteroskedasticity-robust asymptotic covariance matrix estimator is

$$\hat{V} = \hat{Q}^{-1} \hat{\Sigma} \hat{Q}^{-1} \quad (14.47)$$

where

$$\hat{\Sigma} = \frac{1}{n} \sum_{t=1}^n \mathbf{x}_t \mathbf{x}'_t \hat{e}_t^2.$$

Degree-of-freedom adjustments may be made as in the cross-section case, though a theoretical justification has not been developed in the time series case.

Standard errors  $s(\hat{\alpha}_j)$  for individual coefficient estimates can be formed by taking the scaled diagonal elements of  $\hat{V}$ .

**Theorem 14.36** Under the assumptions of Theorem 14.35 as  $n \rightarrow \infty$ ,

$$\hat{V} \xrightarrow{p} V$$

and

$$\frac{\hat{\alpha}_j - \alpha_j}{s(\hat{\alpha}_j)} \xrightarrow{d} N(0, 1).$$

Theorem 14.36 shows that standard covariance matrix estimation is consistent and the resulting t-ratios are asymptotically normal. This means that for stationary autoregressions, inference can proceed using conventional regression methods.

### 14.34 Covariance Matrix Estimation Under General Dependence

Under the assumptions of Theorem 14.35, the conventional covariance matrix estimators are inconsistent as they do not capture the serial dependence in the regression scores  $\mathbf{x}_t e_t$ . To consistently estimate the covariance matrix, we need a different estimator. The appropriate class of estimators are called **Heteroskedasticity and Autocorrelation Consistent (HAC)** or **Heteroskedasticity and Autocorrelation Robust (HAR)** covariance matrix estimators.

To understand the methods, it is helpful to define the vector series  $\mathbf{u}_t = \mathbf{x}_t e_t$  and autocovariance matrices  $\Gamma(\ell) = \mathbb{E}(\mathbf{u}_{t-\ell} \mathbf{u}'_t)$  so that

$$\Omega = \sum_{\ell=-\infty}^{\infty} \Gamma(\ell).$$

Since this sum is convergent the autocovariance matrices converge to zero as  $\ell \rightarrow \infty$ . Therefore  $\Omega$  can be approximated by taking a finite sum of autocovariances, such as

$$\Omega_M = \sum_{\ell=-M}^M \Gamma(\ell).$$

The number  $M$  is sometimes called the **lag truncation** number. Other authors call it the **bandwidth**. An estimator of  $\Gamma(\ell)$  is

$$\widehat{\Gamma}(\ell) = \frac{1}{n} \sum_{1 \leq t-\ell \leq n} \widehat{\mathbf{u}}_{t-\ell} \widehat{\mathbf{u}}_t'$$

where  $\widehat{\mathbf{u}}_t = \mathbf{x}_t \widehat{e}_t$ . By the ergodic theorem we can show that for any  $\ell$ ,  $\widehat{\Gamma}(\ell) \xrightarrow{P} \Gamma(\ell)$ . Thus for any fixed  $M$ , the estimator

$$\widehat{\Omega}_M = \sum_{\ell=-M}^M \widehat{\Gamma}(\ell) \quad (14.48)$$

is consistent for  $\Omega_M$ .

If the serial correlation in  $\mathbf{x}_t e_t$  is known to be zero after  $M$  lags, then  $\Omega_M = \Omega$  and the estimator (14.48) is consistent for  $\Omega$ . This estimator was proposed by L. Hansen and Hodrick (1980) in the context of multiperiod forecasts, and by L. Hansen (1982) for the generalized method of moments weight matrix.

In the general case, we can select  $M$  to increase with sample size  $n$ . If the rate at which  $M$  increases is sufficiently slow, then  $\widehat{\Omega}_M$  will be consistent for  $\Omega$ , as first shown by White and Domowitz (1984).

Once we view the lag truncation number  $M$  as a choice made by the user, the estimator (14.48) has two potential deficiencies. One is that  $\widehat{\Omega}_M$  can change non-smoothly with  $M$ , which makes estimation results sensitive to the choice of  $M$ . The other is that  $\widehat{\Omega}_M$  may not be positive semi-definite and is therefore not a valid variance matrix estimator. We can see this in the simple case of scalar  $u_t$  and  $M = 1$ . In this case

$$\widehat{\Omega}_1 = \widehat{\gamma}(0) (1 + 2\widehat{\rho}(1))$$

which is negative when  $\widehat{\rho}(1) < -1/2$ . Thus if the data are strongly negatively autocorrelated the variance estimator can be negative. A negative variance estimator means that standard errors are ill-defined (a naive computation will produce a complex standard error which makes no sense<sup>5</sup>).

These two deficiencies can be resolved if we amend (14.48) by a weighted sum of autocovariances. In particular, Newey and West (1987b) suggested the estimator

$$\widehat{\Omega}_{NW} = \sum_{\ell=-M}^M \left(1 - \frac{|\ell|}{M+1}\right) \widehat{\Gamma}(\ell). \quad (14.49)$$

This is a weighted sum of the autocovariances. Other weight functions can be used; the one in (14.49) is known as the Bartlett kernel<sup>6</sup>. Newey and West (1987b) showed that this estimator has the algebraic property that  $\widehat{\Omega}_{NW} \geq 0$  (it is positive semi-definite), solving the negative variance estimator problem, and it is also a smooth function of  $M$  as well. Thus this estimator solves the two problems described above.

For  $\widehat{\Omega}_{NW}$  to be consistent for  $\Omega$ , the lag truncation number  $M$  must increase to infinity with  $n$ . Sufficient conditions were established by B. Hansen (1992).

**Theorem 14.37** Under the assumptions of Theorem 14.35, plus  $\sum_{\ell=1}^{\infty} \alpha(\ell)^{1/2-4/r} < \infty$ , if  $M \rightarrow \infty$  yet  $M^3/n = O(1)$ , then as  $n \rightarrow \infty$ ,

$$\widehat{\Omega}_{NW} \xrightarrow{P} \Omega.$$

The assumption  $M^3/n = O(1)$  technically means that  $M$  grows no faster than  $n^{1/3}$ , but this does not have a meaningful practical counterpart other than the implication that “ $M$  should be much smaller than  $n$ ”.

<sup>5</sup>A common computational mishap is a complex standard error. This occurs when a covariance matrix estimator has negative elements on the diagonal.

<sup>6</sup>See Andrews (1991b) for a description of popular options. In practice, the choice of weight function is much less important than the choice of lag truncation number  $M$ .

A important practical issue is how to select  $M$ . One way to think about it is that  $M$  impacts the precision of the estimator  $\hat{\Omega}_{NW}$  through its bias and variance. Since  $\hat{\Gamma}(\ell)$  is a sample average, its variance is  $O(1/n)$  so we expect the variance of  $\hat{\Omega}_M$  to be of order  $O(M/n)$ . The bias of  $\hat{\Omega}_{NW}$  for  $\Omega$  is harder to calculate, but depends on the rate at which the covariances  $\Gamma(\ell)$  decay to zero. Andrews (1991b) found that the  $M$  which minimizes the mean squared error of  $\hat{\Omega}_{NW}$  satisfies the rate  $M = Cn^{1/3}$  where the constant  $C$  depends on the autocovariances. Practical rules to estimate and implement this optimal lag truncation parameter have been proposed by Andrews (1991b) and Newey and West (1994). Stock and Watson (2014) show that a simplified version of Andrews' rule is  $M = 0.75n^{1/3}$ .

### 14.35 Testing the Hypothesis of No Serial Correlation

In some cases it may be of interest to test the hypothesis that the series  $y_t$  is serially uncorrelated against the alternative that it is serially correlated. There have been many proposed tests of this hypothesis. The most appropriate is based on the least squares regression of an AR(p) model. Take the model

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + e_t$$

with  $e_t$  a MDS. In this model, the series  $y_t$  is serially uncorrelated if the slope coefficieints are all zero. Thus the hypothesis of interest is

$$\begin{aligned}\mathbb{H}_0 : \alpha_1 &= \cdots = \alpha_p = 0 \\ \mathbb{H}_1 : \alpha_j &\neq 0 \text{ for some } j \geq 1.\end{aligned}$$

The test can be implemented by a Wald (or F) test. Estimate the AR(p) model by least squares. Form the Wald (or F) statistic using the variance estimator (14.47). (The Newey-West estimator should not be used as there is no serial correlation under the null hypothesis.) Accept the hypothesis if the test statistic is smaller than a conventional critical value (or if the p-value exceeds the significance level), and reject the hypothesis otherwise.

Implementation of this test requires a choice of autoregressive order  $p$ . This choice affects the power of the test. A sufficient number of lags should be included so to pick up potential serial correlation patterns, but not so many that the power of the test is diluted. A reasonable choice in many applications is to set  $p$  to equals  $s$ , the seasonal periodicity. Thus include four lags for quarterly data, or twelve lags for monthly data.

### 14.36 Testing for Omitted Serial Correlation

When using an AR(p) model it may be of interest to know if there is any remaining serial correlation. This can be expressed as a test for serial correlation in the error or equivalently as a test for a higher-order autogressive model.

Take the AR(p) model

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + u_t. \quad (14.50)$$

The null hypothesis is that  $u_t$  is serially uncorrelated, and the alternative hypothesis is that it is serially correlated. We can model the latter as a mean-zero autoregressive process

$$u_t = \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q} + e_t. \quad (14.51)$$

The hypothesis is

$$\begin{aligned}\mathbb{H}_0 : \theta_1 &= \cdots = \theta_q = 0 \\ \mathbb{H}_1 : \theta_j &\neq 0 \text{ for some } j \geq 1.\end{aligned}$$

A seemingly natural test for  $H_0$  uses a two-step method. First estimate (14.50) by least squares and obtain the residuals  $\hat{u}_t$ . Second, estimate (14.51) by least squares by regressing  $\hat{u}_t$  on its lagged values, and obtain the Wald (or F) test for  $H_0$ . This seems like a natural approach, but it is muddled by the fact that the distribution of the Wald statistic is distorted by the two-step procedure. The Wald statistic is not asymptotically chi-square distributed, so it is inappropriate to make a decision based on the conventional critical values and p-values. One approach to obtain the correct asymptotic distribution is to use the generalized method of moments, treating (14.50)-(14.51) as a two-equation just-identified system.

An easier solution is to re-write (14.50)-(14.51) as a higher-order autoregression so that we can use a standard test statistic. To illustrate how this works, for simplicity take the case  $q = 1$ . Take (14.50) and lag the equation once:

$$y_{t-1} = \alpha_0 + \alpha_1 y_{t-2} + \alpha_2 y_{t-3} + \cdots + \alpha_p y_{t-p-1} + u_{t-1}.$$

We then multiply this by  $\theta_1$  and subtract from (14.50), to find

$$\begin{aligned} y_t - \theta_1 y_{t-1} &= \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + u_t \\ &\quad - \theta_1 \alpha_0 - \theta_1 \alpha_1 y_{t-2} - \theta_1 \alpha_2 y_{t-3} - \cdots - \theta_1 \alpha_p y_{t-p-1} - \theta_1 u_{t-1} \end{aligned}$$

or

$$y_t = \alpha_0(1 - \theta_1) + (\alpha_1 + \theta_1) y_{t-1} + (\alpha_2 - \theta_1 \alpha_1) y_{t-2} + \cdots - \theta_1 \alpha_p y_{t-p-1} + e_t.$$

This is an AR(p+1). It simplifies to an AR(p) when  $\theta_1 = 0$ . Thus  $H_0$  is equivalent to the restriction that the coefficient on  $y_{t-p-1}$  is zero.

Thus testing the null hypothesis of an AR(p) (14.50) against the alternative that the error is an AR(1) is equivalent to testing an AR(p) against an AR(p+1). The latter test is implemented as a Wald (or F) test on the coefficient on  $y_{t-p-1}$ .

More generally, testing the null hypothesis of an AR(p) (14.50) against the alternative that the error is an AR(q) is equivalent to testing that  $y_t$  is an AR(p) against the alternative that  $y_t$  is an AR(p+q). The latter test is implemented as a Wald (or F) test on the coefficients on  $y_{t-p-1}, \dots, y_{t-p-1}$ . If the statistic is smaller than the critical values (or the p-value is larger than the significance level) then we reject the hypothesis that the AR(p) is correctly specified in favor of the alternative that there is omitted serial correlation. Otherwise we accept the hypothesis that the AR(p) model is correctly specified.

Another way of deriving the test is as follows. Write (14.50) and (14.51) using lag operator notation

$$\alpha(L)y_t = \alpha_0 + u_t$$

$$\theta(L)u_t = e_t.$$

Applying the operator  $\theta(L)$  to the first equation we obtain

$$\theta(L)\alpha(L)y_t = \alpha_0^* + e_t$$

where  $\alpha_0^* = \theta(1)\alpha_0$ . The product  $\theta(L)\alpha(L)$  is a polynomial of order  $p + q$ , so this is an AR(p+q) model for  $y_t$ .

While this discussion is all good fun, it is unclear if there is good reason to use the test described in this section. Economic theory does not typically produce hypotheses concerning the autoregressive order. Consequently there is rarely a case where there is scientific interest in testing, say, the hypothesis that a series is an AR(4), or any other specific autoregressive order. Instead, practitioners tend to use hypothesis tests for another purpose – model selection. That is, in practice users want to know “What autoregressive model should be used” in a specific application, and resort to hypothesis tests to aid in this decision. This is an inappropriate use of hypothesis tests, because tests are designed to provide answers to scientific questions, rather than being designed to select models with good approximation properties. Instead, model selection should be based on model selection tools. One is described in the following section.

## 14.37 Model Selection

What is the appropriate choice of autoregressive order  $p$  in practice? This is the problem of model selection.

A good choice is to minimize the Akaike information criterion (AIC)

$$\text{AIC}(p) = n \log \hat{\sigma}^2(p) + 2p$$

where  $\hat{\sigma}^2(p)$  is the estimated residual variance from an AR( $p$ ). The AIC is a penalized version of the Gaussian log-likelihood function for the estimated regression model. It is an estimate of the divergence between the fitted model and the true conditional density. By selecting the model with the smallest value of the AIC, you select the model with the smallest estimated divergence – the highest estimated fit between the estimated and true densities.

The AIC is also a monotonic transformation of an estimator of the one-step-ahead forecast mean squared error. Thus selecting the model with the smallest value of the AIC you are selecting the model with the smallest estimated forecast error.

One possible hiccup in computing the AIC criterion for multiple models is that the sample size available for estimation changes as  $p$  changes. (If you increase  $p$ , you need more initial conditions.) This renders AIC comparisons inappropriate. The same sample – the same number of observations – should be used for estimation of all models. The appropriate remedy is to fix a upper value  $\bar{p}$ , and then reserve the first  $\bar{p}$  as initial conditions. Then estimate the models AR(1), AR(2), ..., AR( $\bar{p}$ ) on this (unified) sample.

The AIC of an estimated regression model can be displayed in Stata by using the `estimates stats` command.

## 14.38 Illustrations

We illustrate autoregressive estimation with three empirical examples using U.S. quarterly time series from the FRED-QD data file.

The first example is real GDP growth rates (growth rate of *gdpc1*). We estimate autoregressive models of order 0 through 4 using the sample from 1980-2017<sup>7</sup>. This is a very commonly estimated model in applied macroeconomic practice, and is the empirical version of the Samuelson multiplier-accelerator model discussed in Section 14.23. The coefficient estimates, conventional (heteroskedasticity-robust) standard errors, Newey-West (with  $M = 5$ ) standard errors, and AIC, are displayed in Table 14.1. This sample has 152 observations. The model selected by the AIC criterion is the AR(2). The estimated model has positive and small values for the first two autoregressive coefficients. This means that quarterly output growth rates are positively correlated from quarter to quarter, but only mildly so, and most of the correlation is captured by the first lag. The coefficients of this model are in the real section of Figure 14.8, meaning that the dynamics of the estimated model do not display oscillations. The coefficients of the estimated AR(4) model are nearly identical to the AR(2) model. The conventional and Newey-West standard errors are somewhat different from one another for the AR(0) and AR(4) models, but are nearly identical to one another for the AR(1) and AR(2) models

Our second example is real non-durables consumption growth rates (growth rate of *pcndx*). This is motivated by an influential paper by Robert Hall (1978), who argued that the permanent income hypothesis implies that changes in consumption should be unpredictable (martingale differences). To test this model Hall (1978) estimated an AR(4) model. Our estimated regression using the full sample ( $n = 231$ ), with heteroskedasticity-robust standard errors, are reported in the following equation. Let  $c_t$  denote the consumption growth rate.

---

<sup>7</sup>This sub-sample was used for estimation as it has been argued that the growth rate of U.S. GDP slowed around this period. The goal was to estimate the model over a period of time when the series is plausibly stationary.

Table 14.1: U.S. GDP AR Models

	AR(0)	AR(1)	AR(2)	AR(3)	AR(4)
$\alpha_0$	0.65 (0.06) [0.09]	0.40 (0.08) [0.08]	0.34 (0.10) [0.09]	0.34 (0.10) [0.09]	0.34 (0.11) [0.09]
$\alpha_1$		0.39 (0.09) [0.10]	0.34 (0.10) [0.10]	0.33 (0.10) [0.10]	0.34 (0.10) [0.10]
$\alpha_2$			0.14 (0.11) [0.10]	0.13 (0.13) [0.10]	0.13 (0.14) [0.11]
$\alpha_3$				0.02 (0.11) [0.07]	0.03 (0.12) [0.09]
$\alpha_4$					-0.02 (0.12) [0.13]
AIC	329	306	305	307	309

1. Standard errors robust to heteroskedasticity in parenthesis.
2. Newey-West standard errors in square brackets, with  $M = 5$ .

$$\hat{c}_t = \begin{array}{llllllll} 0.15 & c_{t-1} + & 0.11 & c_{t-2} + & 0.13 & c_{t-3} + & 0.02 & c_{t-4} + & 0.35 \\ (0.07) & & (0.07) & & (0.07) & & (0.08) & & (0.09) \end{array} .$$

Hall's hypothesis is that all autoregressive coefficients should be zero. We test this joint hypothesis with an  $F$  statistic, and find  $F = 3.32$ , with a p-value of  $p = 0.012$ . This is significant at the 5% level, and close to the 1% level. The first three autoregressive coefficients appear to be positive, but small, indicating positive serial correlation. This evidence is (mildly) inconsistent with Hall's hypothesis. We report heteroskedasticity-robust standard errors, not Newey-West standard errors, since the purpose was to test the hypothesis of no serial correlation.

The third example is the first difference of CPI inflation (first difference of growth rate of *cpiaucsl*). This is motivated by Stock and Watson (2007) who examined forecasting models for inflation rates. We estimate autoregressive models of order 1 through 8 using the full sample ( $n = 226$ ); we report models 1 through 5 in Table 14.2. The model with the lowest AIC is the AR(4). All four estimated autoregressive coefficients are negative, most particularly the first two. The two sets of standard errors are quite similar for the AR(4) model. There are meaningful differences only for the lower order AR models.

## 14.39 Time Series Regression Models

Least squares regression methods can be used broadly with stationary time series. Interpretation and usefulness can depend, however, on constructive dynamic specifications. Furthermore, it is necessary to be aware of the serial correlation properties of the series involved, and to use the appropriate covariance matrix estimator when the dynamics have not been explicitly modeled.

Let  $(y_t, \mathbf{x}_t)$  be paired observations with  $y_t$  the dependent variable and  $\mathbf{x}_t$  a vector of regressors including an intercept. The regressors can contain lagged  $y_t$  so this framework includes the autoregressive

Table 14.2: U.S. Inflation AR Models

	AR(1)	AR(2)	AR(3)	AR(4)	AR(5)
$\alpha_0$	0.004 (0.034) [0.023]	0.003 (0.032) [0.028]	0.003 (0.032) [0.029]	0.003 (0.032) [0.031]	0.003 (0.032) [0.032]
$\alpha_1$	-0.26 (0.08) [0.05]	-0.36 (0.07) [0.07]	-0.36 (0.07) [0.07]	-0.36 (0.07) [0.07]	-0.37 (0.07) [0.07]
$\alpha_2$		-0.36 (0.07) [0.06]	-0.37 (0.06) [0.05]	-0.42 (0.06) [0.07]	-0.43 (0.06) [0.07]
$\alpha_3$			-0.00 (0.09) [0.09]	-0.06 (0.10) [0.12]	-0.08 (0.11) [0.13]
$\alpha_4$				-0.16 (0.08) [0.09]	-0.18 (0.08) [0.09]
$\alpha_5$					-0.04 (0.07) [0.06]
AIC	342	312	314	310	312

1. Standard errors robust to heteroskedasticity in parenthesis.
2. Newey-West standard errors in square brackets, with  $M = 5$ .

model as a special case. A linear regression model takes the form

$$y_t = \mathbf{x}'_t \boldsymbol{\beta} + e_t. \quad (14.52)$$

The coefficient vector is defined by projection and therefore equals

$$\boldsymbol{\beta} = (\mathbb{E}(\mathbf{x}_t \mathbf{x}'_t))^{-1} (\mathbb{E}(\mathbf{x}_t y_t)). \quad (14.53)$$

The error  $e_t$  is defined by (14.52) and thus its properties are determined by that relationship. Implicitly the model assumes that the variables have finite second moments and  $\mathbb{E}(\mathbf{x}_t \mathbf{x}'_t) > 0$ , otherwise the model is not uniquely defined and a regressor could be eliminated. By the property of projection the error is uncorrelated with the regressors

$$\mathbb{E}(\mathbf{x}_t e_t) = 0.$$

The least squares estimator of the coefficient is

$$\hat{\boldsymbol{\beta}} = \left( \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t \right)^{-1} \left( \sum_{t=1}^T \mathbf{x}_t y_t \right).$$

Under the assumption that the joint series  $(y_t, \mathbf{x}_t)$  is strictly stationary and ergodic, the estimator is consistent. Under the mixing and moment conditions of Theorem 14.35 the estimator is asymptotically normal with a general covariance matrix

However, under the stronger assumption that the error is a MDS the asymptotic covariance matrix simplifies. It is worthwhile investigating this condition further. The necessary condition is

$$\mathbb{E}(e_t | \mathcal{F}_{t-1}) = 0$$

where  $\mathcal{F}_{t-1}$  is an information set to which  $(e_{t-1}, \mathbf{x}_t)$  is adapted. This notation may appear somewhat odd, but recall in the autoregressive context that  $\mathbf{x}_t = (1, y_{t-1}, \dots, y_{t-p})$  contains variables dated time  $t-1$  and previously, thus  $\mathbf{x}_t$  in this context is a “time  $t-1$ ” variable. The reason why we need  $(e_{t-1}, \mathbf{x}_t)$  to be adapted to  $\mathcal{F}_{t-1}$  is for the MDS condition to hold, the regression function  $\mathbf{x}'_t \boldsymbol{\beta}$  in (14.52) must be the conditional mean of  $y_t$  given  $\mathcal{F}_{t-1}$ , and thus  $\mathbf{x}_t$  must be part of the information set  $\mathcal{F}_{t-1}$ . Under this assumption

$$\mathbb{E}(\mathbf{x}_t e_t | \mathcal{F}_{t-1}) = \mathbf{x}_t \mathbb{E}(e_t | \mathcal{F}_{t-1}) = 0$$

so  $(\mathbf{x}_t e_t, \mathcal{F}_t)$  is a MDS. This means we can apply the MDS CLT to obtain the asymptotic distribution.

We summarize this discussion with the following formal statement.

**Theorem 14.38** If  $(y_t, \mathbf{x}_t)$  is strictly stationary, ergodic, with finite second moments, and  $\mathbf{Q} = \mathbb{E}(\mathbf{x}_t \mathbf{x}'_t) > 0$ , then  $\boldsymbol{\beta}$  in (14.53) is uniquely defined and the least squares estimator is consistent,  $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ .

If in addition,  $\mathbb{E}(e_t | \mathcal{F}_{t-1}) = 0$ , where  $\mathcal{F}_{t-1}$  is an information set to which  $(e_{t-1}, \mathbf{x}_t)$  is adapted,  $\mathbb{E}|y_t|^4 < \infty$ , and  $\mathbb{E}\|\mathbf{x}_t\|^4 < \infty$ , then

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{Q}^{-1} \boldsymbol{\Omega} \mathbf{Q}^{-1}) \quad (14.54)$$

as  $n \rightarrow \infty$ , where  $\boldsymbol{\Omega} = \mathbb{E}(\mathbf{x}_t \mathbf{x}'_t e_t^2)$ .

Alternatively, if in addition, for some  $r > 4$ ,  $\mathbb{E}|y_t|^r < \infty$ ,  $\mathbb{E}\|\mathbf{x}_t\|^r < \infty$ , and the mixing coefficients for  $(y_t, \mathbf{x}_t)$  satisfy  $\sum_{\ell=1}^{\infty} \alpha(\ell)^{1-4/r} < \infty$ , then (14.54) holds with

$$\boldsymbol{\Omega} = \sum_{\ell=-\infty}^{\infty} \mathbb{E}(\mathbf{x}_{t-\ell} \mathbf{x}'_t e_t e_{t-\ell}).$$

## 14.40 Static, Distributed Lag, and Autoregressive Distributed Lag Models

In this section we describe standard linear time series regression models.

Let  $(y_t, \mathbf{z}_t)$  be paired observations with  $y_t$  the dependent variable and  $\mathbf{z}_t$  an observed regressor vector (which does not include lagged  $y_t$ ).

The simplest regression model is the static equation

$$y_t = \alpha + \mathbf{z}'_t \boldsymbol{\beta} + e_t.$$

This is (14.52) by setting  $\mathbf{x}_t = (1, \mathbf{z}'_t)'$ . Static models are motivated to describe how  $y_t$  and  $\mathbf{z}_t$  co-move. Their advantage is their simplicity. The disadvantage is that they are difficult to interpret. The coefficient is the best linear predictor (14.53) but almost certainly is dynamically misspecified since no lagged values are incorporated. The regression of  $y_t$  on contemporaneous  $\mathbf{z}_t$  is also difficult to interpret without a causal framework, since the two can be simultaneous. If this regression is estimated it is important that the standard errors be calculated using the Newey-West method to account for serial correlation in the error.

A model which allows the regressor to have impact over several periods is called a **distributed lag (DL)** model. It takes the form

$$y_t = \alpha + \mathbf{z}'_{t-1} \boldsymbol{\beta}_1 + \mathbf{z}'_{t-2} \boldsymbol{\beta}_2 + \cdots + \mathbf{z}'_{t-q} \boldsymbol{\beta}_q + e_t.$$

It is also possible to include the contemporaneous regressor  $\mathbf{z}_t$ . In this model the leading coefficient  $\boldsymbol{\beta}_1$  represents the initial impact of  $\mathbf{z}_t$  on  $y_t$ ,  $\boldsymbol{\beta}_2$  represents the impact in the second period, and so on. The cumulative impact is the sum of the coefficients  $\boldsymbol{\beta}_1 + \cdots + \boldsymbol{\beta}_q$ , which is called the **long-run multiplier**.

The distributed lag model falls in the class (14.52) by setting  $\mathbf{x}_t = (1, \mathbf{z}'_{t-1}, \mathbf{z}'_{t-2}, \dots, \mathbf{z}'_{t-q})'$ . While it allows for a lagged impact of  $\mathbf{z}_t$  on  $y_t$ , the model does not incorporate serial correlation, so the error  $e_t$  should be expected to be serially correlated. Thus the model is (typically) dynamically misspecified which can make interpretation difficult. It is also necessary to use Newey-West standard errors to account for the serial correlation.

A more complete model combines autoregressive and distributed lags. It takes the form

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \dots + \alpha_p y_{t-p} + \mathbf{z}'_{t-1} \boldsymbol{\beta}_1 + \dots + \mathbf{z}'_{t-q} \boldsymbol{\beta}_q + e_t.$$

This is called an **autoregressive distributed lag (AR-DL)** model. It nests both the autoregressive and distributed lag models, thereby combining serial correlation and dynamic impact. The AR-DL model falls in the class (14.52) by setting  $\mathbf{x}_t = (1, y_{t-1}, \dots, y_{t-p}, \mathbf{z}'_{t-1}, \dots, \mathbf{z}'_{t-q})'$ .

If the lag orders  $p$  and  $q$  are selected sufficiently large, the AR-DL model will have an error which is approximately white noise, in which case the model can be interpreted as dynamically well-specified, and conventional standard error methods can be used.

In an AR-DL specification, the long-run multiplier is

$$LR = \frac{\beta_1 + \dots + \beta_q}{1 - \alpha_1 - \dots - \alpha_p}$$

which is a nonlinear function of the coefficients.

## 14.41 Time Trends

Many economic time series have means which change over time. A useful way to think about this is the components model

$$y_t = T_t + u_t$$

where  $T_t$  is the trend component and  $u_t$  is the stochastic component. The latter can be modeled by a linear process or autoregression

$$\alpha(L)u_t = e_t.$$

The trend component is often modeled as a linear function in the time index

$$T_t = \beta_0 + \beta_1 t$$

or a quadratic function in time

$$T_t = \beta_0 + \beta_1 t + \beta_2 t^2.$$

These models are typically not thought of as being literally true, but rather as useful approximations.

When we write down time series models we write the index as  $t = 1, \dots, n$ . But in practical applications the time index corresponds to a date, e.g.  $t = 1960, 1961, \dots, 2017$ . Furthermore, if the data is at a higher frequency than annual then it is incremented in fractional units. This is not of fundamental importance; it merely changes the meaning of the intercept  $\beta_0$  and slope  $\beta_1$ . Consequently these should not be interpreted outside of how the time index is defined.

One traditional way of dealing with time trends is to “detrend” the data. This means using an estimation method to estimate the trend and subtract it off. The simplest method is least squares linear detrending. Given the linear model

$$y_t = \beta_0 + \beta_1 t + u_t \tag{14.55}$$

the coefficients are estimated by least squares. The detrended series is the residual  $\hat{u}_t$ . More intricate methods can be used but they have a similar flavor.

To understand the properties of the detrending method we can apply an asymptotic approximation. A time trend is not a stationary process so we should be thoughtful before applying standard theory. We will study asymptotics for non-stationary processes in more detail in Chapter 16, so our treatment here

will be brief. It turns out that most of our conventional procedures work just fine with time trends (and quadratics in time) as regressors. The rates of convergence change but this does not affect anything of practical importance.

Let us demonstrate that the least squares estimator of the coefficients in (14.55) is consistent. We can write the estimator as

$$\begin{pmatrix} \hat{\beta}_0 - \beta_0 \\ \hat{\beta}_1 - \beta_1 \end{pmatrix} = \begin{pmatrix} n & \sum_{t=1}^n t \\ \sum_{t=1}^n t & \sum_{t=1}^n t^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{t=1}^n u_t \\ \sum_{t=1}^n tu_t \end{pmatrix}.$$

We need to study the behavior of the sums in the design matrix. For this we appeal to Theorem 14.7, which showed that for any  $r > 0$

$$\frac{1}{n^{1+r}} \sum_{t=1}^n t \rightarrow \frac{1}{1+r}.$$

This implies that

$$\frac{1}{n^2} \sum_{t=1}^n t \rightarrow \frac{1}{2}$$

and

$$\frac{1}{n^3} \sum_{t=1}^n t^2 \rightarrow \frac{1}{3}.$$

This is new. The sums require normalizations other than  $n^{-1}$ !

To handle this in multiple regression it is convenient to define a scaling matrix which normalizes each elements in the regression by its convergence rate. Define the matrix  $D_n = \begin{bmatrix} 1 & \\ & n \end{bmatrix}$ . The first element is the the intercept and second for the time trend. Then

$$\begin{aligned} D_n \begin{pmatrix} \hat{\beta}_0 - \beta_0 \\ \hat{\beta}_1 - \beta_1 \end{pmatrix} &= D_n \begin{pmatrix} n & \sum_{t=1}^n t \\ \sum_{t=1}^n t & \sum_{t=1}^n t^2 \end{pmatrix}^{-1} D_n D_n^{-1} \begin{pmatrix} \sum_{t=1}^n u_t \\ \sum_{t=1}^n tu_t \end{pmatrix} \\ &= \left( D_n \begin{pmatrix} n & \sum_{t=1}^n t \\ \sum_{t=1}^n t & \sum_{t=1}^n t^2 \end{pmatrix} D_n \right)^{-1} \begin{pmatrix} \sum_{t=1}^n u_t \\ \frac{1}{n} \sum_{t=1}^n tu_t \end{pmatrix} \\ &= \begin{pmatrix} n & \frac{1}{n} \sum_{t=1}^n t \\ \frac{1}{n} \sum_{t=1}^n t & \frac{1}{n^2} \sum_{t=1}^n t^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{t=1}^n u_t \\ \frac{1}{n} \sum_{t=1}^n tu_t \end{pmatrix} \end{aligned}$$

Then multiplying by  $n^{1/2}$  we obtain

$$\begin{pmatrix} n^{1/2}(\hat{\beta}_0 - \beta_0) \\ n^{3/2}(\hat{\beta}_1 - \beta_1) \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{n^2} \sum_{t=1}^n t \\ \frac{1}{n^2} \sum_{t=1}^n t & \frac{1}{n^2} \sum_{t=1}^n t^2 \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{n^{1/2}} \sum_{t=1}^n u_t \\ \frac{1}{n^{3/2}} \sum_{t=1}^n tu_t \end{pmatrix}.$$

The denominator matrix satisfies

$$\begin{pmatrix} 1 & \frac{1}{n^2} \sum_{t=1}^n t \\ \frac{1}{n^2} \sum_{t=1}^n t & \frac{1}{n^2} \sum_{t=1}^n t^2 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{pmatrix}$$

which is invertible. Setting  $\mathbf{x}_{nt} = (t/n 1)$ , the numerator vector can be written as  $\frac{1}{n^{1/2}} \sum_{t=1}^n \mathbf{x}_{nt} u_t$ . It has variance

$$\begin{aligned} \left\| \text{var} \left( \frac{1}{n^{1/2}} \sum_{t=1}^n \mathbf{x}_{nt} u_t \right) \right\| &= \left\| \frac{1}{n} \sum_{t=1}^n \sum_{j=1}^n \mathbf{x}_t \mathbf{x}'_j \mathbb{E}(u_t u_j) \right\| \\ &\leq \sqrt{2} \sum_{\ell=-\infty}^{\infty} \|\mathbb{E}(u_t u_j)\| < \infty \end{aligned}$$

by Theorem 14.19 if  $u_t$  satisfies the mixing and moment conditions for the central limit theory. This means that the numerator vector is  $O_p(1)$ . (It is also asymptotically normal but we defer this demonstration for now.) We conclude that

$$\begin{pmatrix} n^{1/2}(\hat{\beta}_0 - \beta_0) \\ n^{3/2}(\hat{\beta}_1 - \beta_1) \end{pmatrix} = O_p(1).$$

This shows that both coefficients are consistent,  $\hat{\beta}_0$  converges at the standard  $n^{1/2}$  rate, and  $\hat{\beta}_1$  converges at the faster  $n^{3/2}$  rate. The consistency of the coefficient estimators means that the detrending method is consistent.

An alternative is simply to include a time trend in an estimated regression model. If we have an autoregression, a distributed lag, or an AR-DL model, we can add a time index to obtain a model of the form

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \mathbf{z}'_{t-1} \boldsymbol{\beta}_1 + \cdots + \mathbf{z}'_{t-q} \boldsymbol{\beta}_q + \gamma t + e_t.$$

Estimation by least squares is equivalent to estimation after linear detrending (by the FWL theorem). Inclusion of a linear (and possibly quadratic) time trend in a regression model is typically the easiest method to incorporate time trends.

## 14.42 Illustration

We illustrate the model described in the previous section using a classical macroeconomic model for inflation prediction based on the Phillips curve. A. W. Phillips (1958) famously observed that the unemployment rate and the wage inflation rate are negatively correlated over time. Equations relating the inflation rate, or the change in the inflation rate, to macroeconomic indicators such as the unemployment rate are typically described as “Phillips curves”. A simple Phillips curve takes the form

$$\Delta\pi_t = \alpha + \beta ur_t + e_t \quad (14.56)$$

where  $\pi_t$  is price inflation and  $ur_t$  is the unemployment rate. This specification relates the change in inflation in a given period to the level of the unemployment rate in the previous period.

The least squares estimate of (14.56) using U.S. quarterly series from FRED-QD is reported in the first column Table 14.3. Both heteroskedasticity-robust and Newey-West standard errors are reported. The Newey-West standard errors are the appropriate choice since the estimated equation is static – no modeling of the serial correlation. In this example, the measured impact of the unemployment rate on inflation appears minimal. The estimate is consistent with a small effect of the unemployment rate on the inflation rate, but it is not precisely estimated.

A distributed lag (DL) model takes the form

$$\Delta\pi_t = \alpha + \beta_1 ur_{t-1} + \beta_2 ur_{t-2} + \cdots + \beta_q ur_{t-q} + e_t. \quad (14.57)$$

The least squares estimate of (14.57) is reported in the second column of Table 14.3. The estimates are quite different from the static model. We see large negative impacts in the first and third periods, countered by a large positive impact in the second period. The model suggests that the unemployment rate has a strong impact on the inflation rate, but the long-run impact is mitigated. The long-run multiplier is reported at the bottom of the column. The point estimate of  $-0.022$  is quite small, and very similar to the static estimate. It implies that an increase in the unemployment rate by 5 percentage points (a typical recession) decreases the long-run annual inflation rate by about one-half of one percentage points.

An AR-DL takes the form

$$\Delta\pi_t = \alpha_0 + \alpha_1 \Delta\pi_{t-1} + \cdots + \alpha_p \Delta\pi_{t-p} + \beta_1 ur_{t-1} + \cdots + \beta_q ur_{t-q} + e_t. \quad (14.58)$$

The least squares estimate of (14.58) is reported in the third column of Table 14.3. The coefficient estimates appear similar to those from the distributed lag model. The point estimate of the long-run multiplier is also nearly identical, but with a somewhat smaller standard error.

## 14.43 Granger Causality

In the AR-DL model (14.58) the unemployment rate would have no predictive impact on the inflation rate under the coefficient restriction  $\beta_1 = \cdots = \beta_q = 0$ . This restriction is called “Granger non-causality”.

Table 14.3: Phillips Curve Regressions

	Static Model	DL Model	AR-DL Model
$ur_t$	-0.023 (0.025) [0.017]		
$ur_{t-1}$		-0.59 (0.20) [0.16]	-0.62 (0.16) [0.12]
$ur_{t-2}$		1.14 (0.29) [0.28]	0.88 (0.25) [0.21]
$ur_{t-3}$		-0.68 (0.22) [0.25]	-0.36 (0.25) [0.24]
$ur_{t-4}$		0.12 (0.11) [0.11]	0.05 (0.12) [0.12]
$\pi_{t-1}$			-0.43 (0.08) [0.08]
$\pi_{t-2}$			-0.47 (0.10) [0.09]
$\pi_{t-3}$			-0.14 (0.10) [0.11]
$\pi_{t-4}$			-0.19 (0.08) [0.09]
Multiplier	-0.023 [0.017]	-0.022 [0.012]	-0.021 [0.008]

1. Standard errors robust to heteroskedasticity in parenthesis.
2. Newey-West standard errors in square brackets, with  $M = 5$ .

This definition of causality was developed by Granger (1969) and Sims (1972). When the coefficients are non-zero we say that the unemployment rate “Granger causes” the inflation rate.

The reason why we call this “Granger Causality” rather than “causality” is because this is not a physical or structure definition of causality. An alternative label is “predictive causality”.

To be precise, assume that we have two series  $(y_t, z_t)$ . Consider the projection of  $y_t$  onto the lagged history of both series

$$\begin{aligned} y_t &= \mathcal{P}_{t-1}(y_t) + e_t \\ &= \alpha_0 + \sum_{j=1}^{\infty} \alpha_j y_{t-j} + \sum_{j=1}^{\infty} \beta_j z_{t-j} + e_t. \end{aligned}$$

We say that  $z_t$  does not Granger-cause  $y_t$  if  $\beta_j = 0$  for all  $j$ . If  $\beta_j \neq 0$  for some  $j$  then we say that  $z_t$  Granger-causes  $y_t$ .

It is important that the definition includes the projection on the past history of  $y_t$ . Granger causality means that  $z_t$  helps to predict  $y_t$  even after the past history of  $y_t$  has been accounted for.

The definition can alternatively be written in terms of conditional expectations rather than projections. We can say that  $z_t$  does not Granger-cause  $y_t$  if

$$\mathbb{E}(y_t | y_{t-1}, y_{t-2}; z_{t-1}, z_{t-2}) = \mathbb{E}(y_t | y_{t-1}, y_{t-2}).$$

Granger causality can be tested in AR-DL models using a standard Wald or F test. In the context of model (14.58) we report the F statistic for  $\beta_1 = \dots = \beta_q = 0$ . The test rejects the hypothesis (and thus finds evidence of Granger causality) if the statistic is larger than the critical value (if the p-value is small), and fails to reject the hypothesis (and thus finds no evidence of causality) if the statistic is smaller than the critical value.

For example, in the results presented in Table 14.3, the F statistic for the hypothesis  $\beta_1 = \dots = \beta_4 = 0$  using the Newey-West covariance matrix is  $F = 6.98$  with a p-value of 0.000. This is statistically significant at any conventional level so we can conclude that the unemployment rate has a predictively causal impact on inflation.

Granger causality should not be interpreted structurally outside the context of an economic model. For example consider the regression of GDP growth rates  $y_t$  on stock price growth rates  $r_t$ . We use the quarterly series from FRED-QD, using an AR-DL specification with two lags

$$\begin{aligned} y_t &= 0.22 \quad y_{t-1} + 0.14 \quad y_{t-2} + 0.03 \quad r_{t-1} + 0.01 \quad r_{t-2}. \\ &\quad (0.09) \quad (0.10) \quad (0.01) \quad (0.01) \end{aligned}$$

The coefficients on the lagged stock price growth rates are small in magnitude, but the first lag appears statistically significant. The F statistic for exclusion of  $(r_{t-1}, r_{t-2})$  is  $F = 9.3$  with a p-value of 0.0002, which is highly significant. We can therefore reject the hypothesis of no Granger causality, and deduce that stock price changes Granger-cause GDP growth. This should not be interpreted as suggesting that the stock market causes output fluctuations, as a more reasonable explanation from economic theory is that stock prices are forward-looking measures of expected future profits. When corporate profits are forecasted to rise, the value of corporate stock rises, bidding up stock prices. Thus stock prices move in advance of actual economic activity, but are not necessarily structurally causal.

### Clive W. J. Granger

Clive Granger (1934-2009) of England was one of the leading figures in time-series econometrics, and co-winner in 2003 of the Nobel Memorial Prize in Economic Sciences (along with Robert Engle). In addition to formalizing the definition of causality known as Granger causality, he invented the concept of cointegration, introduced spectral methods into econometrics, and formalized methods for the combination of forecasts.

## 14.44 Testing for Serial Correlation in Regression Models

Consider the problem of testing for omitted serial correlation in an AR-DL model such as

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \beta_1 z_{t-1} + \cdots + \beta_q z_{t-q} + u_t. \quad (14.59)$$

The null hypothesis is that  $u_t$  is serially uncorrelated, and the alternative hypothesis is that it is serially correlated. We can model the latter as a mean-zero autoregressive process

$$u_t = \theta_1 u_{t-1} + \cdots + \theta_r u_{t-r} + e_t. \quad (14.60)$$

The hypothesis is

$$\begin{aligned} H_0: \theta_1 &= \cdots = \theta_r = 0 \\ H_1: \theta_j &\neq 0 \text{ for some } j \geq 1. \end{aligned}$$

There are two ways to implement a test of  $H_0$  against  $H_1$ . The first is to estimate equations (14.59)-(14.60) sequentially by least squares and construct a test for  $H_0$  on the second equation. This test is complicated by the fact that the two-step nature of the second regression invalidates conventional asymptotic approximations. Therefore this approach is not recommended.

The second approach is to combine equations (14.59)-(14.60) into a single model and execute the test as a restriction within this model. One way to make this combination is by using lag operator notation. Write (14.59)-(14.60) as

$$\begin{aligned} \alpha(L)y_t &= \alpha_0 + \beta(L)z_{t-1} + u_t \\ \theta(L)u_t &= e_t \end{aligned}$$

Then applying the operator  $\theta(L)$  to the first equation we obtain

$$\theta(L)\alpha(L)y_t = \theta(L)\alpha_0 + \theta(L)\beta(L)z_{t-1} + \theta(L)u_t$$

or

$$\alpha^*(L)y_t = \alpha_0^* + \beta^*(L)z_{t-1} + e_t$$

where  $\alpha^*(L)$  is a  $p+r$  order polynomial and  $\beta^*(L)$  is a  $q+r$  order polynomial. The restriction  $H_0$  is that these are  $p$  and  $q$  order polynomials. Thus we can implement a test of  $H_0$  against  $H_1$  by estimating an AR-DL model with  $p+r$  and  $q+r$  lags, and testing the exclusion of the final  $r$  lags of  $y_t$  and  $z_t$ . This test has a conventional asymptotic distribution so is simple to implement.

The basic message is that testing for omitted serial correlation can be implemented in regression models by estimating and contrasting different dynamic specifications.

## 14.45 Bootstrap for Time Series

Recall that the bootstrap approximates the sampling distribution of estimators and test statistics by using the empirical distribution of the observations. The traditional non-parametric bootstrap is appropriate for case of independent observations. For dependent observations alternative methods should be used.

Bootstrapping for time series is considerably more complicated than the cross section case, and many methods have been proposed. Part of the challenge is that theoretical justifications are much more difficult to establish than in the independent observation case.

In this section we describe the most popular methods to implement bootstrap resampling for time series data.

### Recursive Bootstrap

1. Estimate a complete model such as an AR(p) with coefficients  $\hat{\alpha}$  and residuals  $\hat{e}_t$ .
2. Fix the initial condition  $(y_{-p+1}, y_{-p+2}, \dots, y_0)$ .
3. Simulate i.i.d. draws  $e_t^*$  from the empirical distribution of the residuals  $\{\hat{e}_1, \dots, \hat{e}_n\}$ .
4. Create the bootstrap series  $y_t^*$  by the recursive formula

$$y_t^* = \hat{\alpha}_0 + \hat{\alpha}_1 y_{t-1}^* + \hat{\alpha}_2 y_{t-2}^* + \dots + \hat{\alpha}_p y_{t-p}^* + e_t^*.$$

This construction creates bootstrap samples  $y_t^*$  with the stochastic properties of the estimated AR(p) model, including the auxiliary assumption that the errors are i.i.d. This method can work well if the true process is an AR(p). One flaw is that it imposes homoskedasticity on the errors  $e_t^*$ , which may be different than the properties of the actual  $e_t$ . Another limitation is that it is inappropriate for AR-DL models unless the conditioning variables are treated as strictly exogenous.

There are alternative versions of this basic method. First, instead of fixing the initial conditions at the sample values a random block can be drawn from the sample. The difference is that this produces an unconditional distribution rather than a conditional one. Second, instead of drawing the errors from the residuals a parametric (typically normal) distribution can be used. This can improve precision when sample sizes are very small but otherwise is not recommended.

### **Pairwise Bootstrap**

1. Write the sample as  $\{y_t, \mathbf{x}_t\}$  where  $\mathbf{x}_t = (y_{t-1}, \dots, y_{t-p})'$  contains the lagged values used in estimation.
2. Apply the traditional nonparametric bootstrap which samples pairs  $(y_t^*, \mathbf{x}_t^*)$  i.i.d. from  $\{y_t, \mathbf{x}_t\}$  with replacement to create the bootstrap sample.
3. Create the bootstrap estimates on this bootstrap sample, e.g. regress  $y_t^*$  on  $\mathbf{x}_t^*$ .

This construction is essentially the traditional nonparametric bootstrap, but applied to the paired sample  $\{y_t, \mathbf{x}_t\}$ . It does not mimic the time series correlations across observations. However, it does produce bootstrap statistics with the correct first-order asymptotic distribution (under MDS errors). This method may be useful when we are interested in the distribution of nonlinear functions of the coefficient estimates and therefore desire an improvement on the Delta Method approximation.

### **Fixed Design Residual Bootstrap**

1. Write the sample as  $\{y_t, \mathbf{x}_t, \hat{e}_t\}$  where  $\mathbf{x}_t = (y_{t-1}, \dots, y_{t-p})'$  contains the lagged values used in estimation and  $\hat{e}_t$  are the residuals.
2. Fix the regressors  $\mathbf{x}_t$  at their sample values.
3. Simulate i.i.d. draws  $e_t^*$  from the empirical distribution of the residuals  $\{\hat{e}_1, \dots, \hat{e}_n\}$ .
4. Set  $y_t^* = \mathbf{x}_t' \hat{\beta} + e_t^*$

This construction is similar to the pairwise bootstrap, but imposes an i.i.d. error. It is therefore only valid when the errors are i.i.d. (and thus excludes heteroskedasticity).

### **Fixed Design Wild Bootstrap**

1. Write the sample as  $\{y_t, \mathbf{x}_t, \hat{e}_t\}$  where  $\mathbf{x}_t = (y_{t-1}, \dots, y_{t-p})'$  contains the lagged values used in estimation and  $\hat{e}_t$  are the residuals.

2. Fix the regressors  $\mathbf{x}_t$  and residuals  $\hat{e}_t$  at their sample values.
3. Simulate i.i.d. auxiliary random  $\xi_t^*$  with mean zero and variance one. See Section 10.31 for a discussion of choices.
4. Set  $e_t^* = \xi_t^* \hat{e}_t$  and  $y_t^* = \mathbf{x}'_t \hat{\beta} + e_t^*$ .

This construction is similar to the pairwise and fixed design bootstrap methods, but uses the wild bootstrap method. This imposes the conditional mean assumption on the error but allows heteroskedasticity.

### Block Bootstrap

1. Write the sample as  $\{(y_t, \mathbf{x}_t)\}$  where  $\mathbf{x}_t = (y_{t-1}, \dots, y_{t-p})'$  contains the lagged values used in estimation.
2. Divide the sample of paired observations  $\{(y_t, \mathbf{x}_t)\}$  into  $n/m$  blocks of length  $m$ .
3. Resample complete blocks. For each simulated sample, draw  $n/m$  blocks.
4. Paste the blocks together to create the bootstrap time series  $\{y_t^*, \mathbf{x}_t^*\}$ .

This construction allows for arbitrary stationary serial correlation, heteroskedasticity, and for model misspecification. One challenge is that the block bootstrap is sensitive to the block length and the way that the data are partitioned into blocks. The method may also work less well in small samples. Notice that the block bootstrap with  $m = 1$  is equal to the pairwise bootstrap, and the latter is the traditional nonparametric bootstrap. Thus the block bootstrap is a natural generalization of the nonparametric bootstrap.

## 14.46 Technical Proofs\*

**Proof of Theorem 14.2.** Define  $\tilde{\mathbf{y}}_t = (\mathbf{y}_t, \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \dots) \in \mathbb{R}^{m \times \infty}$  as the history of  $\mathbf{y}_t$  up to time  $t$ . We can then write  $\mathbf{x}_t = \phi(\tilde{\mathbf{y}}_t)$ . Let  $B$  be the pre-image of  $\{\mathbf{x}_t \leq \mathbf{x}\}$  (the vectors  $\tilde{\mathbf{y}} \in \mathbb{R}^{m \times \infty}$  such that  $\phi(\tilde{\mathbf{y}}) \leq \mathbf{x}$ ). Then

$$\mathbb{P}(\mathbf{x}_t \leq \mathbf{x}) = \mathbb{P}(\phi(\tilde{\mathbf{y}}_t) \leq \mathbf{x}) = \mathbb{P}(\tilde{\mathbf{y}}_t \in B).$$

Since  $\mathbf{y}_t$  is strictly stationary the probability  $\mathbb{P}(\tilde{\mathbf{y}}_t \in B)$  is independent<sup>8</sup> of  $t$ . This means that the distribution of  $\mathbf{x}_t$  is independent of  $t$ . This argument can be extended to show that the distribution of  $(\mathbf{x}_t, \dots, \mathbf{x}_{t+\ell})$  is independent of  $t$ . This means that  $\mathbf{x}_t$  is strictly stationary as claimed. ■

**Proof of Theorem 14.3.** We need to verify that the series  $S_N = \sum_{j=0}^N a_j y_{t-j}$  converges almost surely as  $N \rightarrow \infty$ . By the Cauchy criterion for convergence (see Section 14.6), this holds if for all  $\varepsilon > 0$ , there is an  $N < \infty$  such that  $\max_{m \geq 1} |S_{N+m} - S_N| \leq \varepsilon$ . Using Markov's inequality (B.35) and the triangle inequality

---

<sup>8</sup>An astute reader may notice that the independence of  $\mathbb{P}(\tilde{\mathbf{y}}_t \in B)$  from  $t$  does not follow directly from the definition of strict stationarity. Indeed, a full derivation requires a measure-theoretic treatment. See Section 1.2.B of Petersen (1983) or Section 3.5 of Stout (1974).

(B.1)

$$\begin{aligned}
\mathbb{P}\left(\max_{m \geq 1} |S_{N+m} - S_N| > \varepsilon\right) &\leq \varepsilon^{-1} \mathbb{E}\left(\max_{m \geq 1} |S_{N+m} - S_N|\right) \\
&= \varepsilon^{-1} \mathbb{E}\left(\max_{m \geq 1} \left| \sum_{j=N+1}^{N+m} a_j y_{t-j} \right|\right) \\
&\leq \mathbb{E}\left(\sum_{j=N+1}^{\infty} |a_j| |y_{t-j}|\right) \\
&\leq \left(\sum_{j=N+1}^{\infty} |a_j|\right) \left(\sup_t \mathbb{E}|y_t|\right) \\
&\rightarrow 0
\end{aligned}$$

as  $N \rightarrow \infty$  since  $\sum_{j=0}^{\infty} |a_j| < \infty$ . Thus  $x_t = \sum_{j=0}^{\infty} a_j y_{t-j}$  converges almost surely.

If  $y_t$  is strictly stationary then  $x_t$  is as well by Theorem 14.2. ■

**Proof of Theorem 14.5.** Since  $\sum_{\ell=1}^n w_{n\ell} \rightarrow 1$  we can without loss of generality set  $A = 0$ . Fix  $\varepsilon > 0$ . Pick  $N$  such that  $|a_\ell| \leq \varepsilon$  for  $\ell \geq N$ . Pick  $n$  sufficiently large so that

$$\sum_{\ell=1}^N w_{n\ell} |a_\ell| \leq \varepsilon$$

which is feasible since  $w_{n\ell} \rightarrow 0$ . Then

$$\begin{aligned}
\left| \sum_{\ell=1}^n w_{n\ell} a_\ell \right| &\leq \sum_{\ell=1}^n w_{n\ell} |a_\ell| \\
&= \sum_{\ell=1}^N w_{n\ell} |a_\ell| + \sum_{\ell=N+1}^n w_{n\ell} |a_\ell| \\
&\leq 2\varepsilon.
\end{aligned}$$

Since  $\varepsilon$  is arbitrary this establishes that  $\frac{1}{n} \sum_{\ell=1}^n w_{n\ell} a_\ell \rightarrow 0$  as asserted. ■

**Proof of Theorem 14.8.** See Theorem 14.18. ■

**Proof of Theorem 14.9.** Strict stationarity follows from Theorem 14.2. Let  $\tilde{\mathbf{y}}_t$  and  $\tilde{\mathbf{x}}_t$  be the histories of  $\mathbf{y}_t$  and  $\mathbf{x}_t$ . We can write  $\mathbf{x}_t = \phi(\tilde{\mathbf{y}}_t)$ . Let  $A$  be an invariant event for  $\mathbf{x}_t$ . We want to show  $\mathbb{P}(A) = 0$  or 1. The event  $A$  is a collection of  $\tilde{\mathbf{x}}_t$  histories, and occurs if and only if an associated collection of  $\tilde{\mathbf{y}}_t$  histories occur. That is, for some sets  $G$  and  $H$ ,

$$A = \{\tilde{\mathbf{x}}_t \in G\} = \{\phi(\tilde{\mathbf{y}}_t) \in G\} = \{\tilde{\mathbf{y}}_t \in H\}.$$

The assumption that  $A$  is invariant means it is unaffected by the time shift, thus can be written as

$$A = \{\tilde{\mathbf{x}}_{t+\ell} \in G\} = \{\tilde{\mathbf{y}}_{t+\ell} \in H\}.$$

This means the event  $\{\tilde{\mathbf{y}}_{t+\ell} \in H\}$  is invariant. Since  $\mathbf{y}_t$  is ergodic, the event has probability 0 or 1. Hence  $\mathbb{P}(A) = 0$  or 1, as desired. ■

**Proof of Theorem 14.11.** Suppose  $y_t$  is discrete with support on  $(\tau_1, \dots, \tau_N)$  and without loss of generality

assume  $\mathbb{E}(y_t) = 0$ . Then by Theorem 14.12

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{\ell=1}^n \text{cov}(y_t, y_{t+\ell}) &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{\ell=1}^n \mathbb{E}(y_t y_{t+\ell}) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{\ell=1}^n \sum_{j=1}^N \sum_{k=1}^N \tau_j \tau_k \mathbb{P}(y_t = \tau_j, y_{t+\ell} = \tau_k) \\ &= \sum_{j=1}^N \sum_{k=1}^N \tau_j \tau_k \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{\ell=1}^n \mathbb{P}(y_t = \tau_j, y_{t+\ell} = \tau_k) \\ &= \sum_{j=1}^N \sum_{k=1}^N \tau_j \tau_k \mathbb{P}(y_t = \tau_j) \mathbb{P}(y_{t+\ell} = \tau_k) \\ &= \mathbb{E}(y_t) \mathbb{E}(y_{t+\ell}) \\ &= 0. \end{aligned}$$

which is (14.4). This can be extended to the case of continuous distributions using the monotone convergence theorem. See Corollary 13.14 of Davidson (1994). ■

**Proof of Theorem 14.13.** We show (14.6). (14.7) follows by Markov's inequality (B.35).

Without loss of generality we focus on the scalar case, and assume  $\mathbb{E}(y_t) = 0$ . Fix  $\varepsilon > 0$ . Pick  $B$  large enough such that

$$\mathbb{E}|y_t \mathbf{1}(|y_t| > B)| \leq \frac{\varepsilon}{4} \quad (14.61)$$

which is feasible since  $\mathbb{E}|y_t| < \infty$ . Define

$$\begin{aligned} w_t &= y_t \mathbf{1}(|y_t| \leq B) - \mathbb{E}(y_t \mathbf{1}(|y_t| \leq B)) \\ z_t &= y_t \mathbf{1}(|y_t| > B) - \mathbb{E}(y_t \mathbf{1}(|y_t| > B)). \end{aligned}$$

Notice that  $w_t$  is a bounded transformation of the ergodic series  $y_t$ . Thus by (14.4) and (14.9) there is an  $n$  sufficiently large so that

$$\frac{\text{var}(w_t)}{n} + \frac{2}{n} \sum_{m=1}^n \left(1 - \frac{m}{n}\right) \text{cov}(w_t, w_m) \leq \frac{\varepsilon^2}{4} \quad (14.62)$$

By the triangle inequality (B.1)

$$\mathbb{E}|\bar{y}| = \mathbb{E}|\bar{w} + \bar{z}| \leq \mathbb{E}|\bar{w}| + \mathbb{E}|\bar{z}|. \quad (14.63)$$

By another application of the triangle inequality and (14.61)

$$\mathbb{E}|\bar{z}| \leq \mathbb{E}|z_t| \leq 2\mathbb{E}|y_t \mathbf{1}(|y_t| > B)| \leq \frac{\varepsilon}{2}. \quad (14.64)$$

By Jensen's inequality (B.26), direct calculation, and (14.62)

$$\begin{aligned} (\mathbb{E}|\bar{w}|)^2 &\leq \mathbb{E}|\bar{w}|^2 \\ &= \frac{1}{n^2} \sum_{t=1}^n \sum_{j=1}^n \mathbb{E}(w_t w_j) \\ &= \frac{\text{var}(w_t)}{n} + \frac{2}{n} \sum_{m=1}^n \left(1 - \frac{m}{n}\right) \text{cov}(w_t, w_m) \\ &\leq \frac{\varepsilon^2}{4}. \end{aligned}$$

Thus

$$\mathbb{E}|\bar{w}| \leq \frac{\varepsilon}{2}. \quad (14.65)$$

Together, (14.63), (14.64) and (14.65) show that  $\mathbb{E}|\bar{y}| \leq \varepsilon$ . Since  $\varepsilon$  is arbitrary, this establishes (14.6) as claimed. ■

**Proof of Theorem 14.15 (sketch).** By the Cramér-Wold device (Theorem 6.10) it is sufficient to establish the result for scalar  $u_t$ . Let  $\sigma^2 = \mathbb{E}(u_t^2)$ . By a Taylor series expansion, for  $x$  small

$$\log(1+x) \simeq x - \frac{x^2}{2}.$$

Taking exponentials and rearranging, we obtain the approximation

$$\exp(x) \simeq (1+x) \exp\left(\frac{x^2}{2}\right). \quad (14.66)$$

Fix  $\lambda$ . Define

$$T_j = \prod_{i=1}^j \left(1 + \frac{\lambda}{\sqrt{n}} u_t\right)$$

$$V_n = \frac{1}{n} \sum_{t=1}^n u_t^2.$$

Since  $u_t$  is strictly stationary and ergodic,  $V_n \xrightarrow{p} \sigma^2$  by the Ergodic Theorem (Theorem 14.13). Since  $u_t$  is a MDS

$$\mathbb{E}(T_n) = 1. \quad (14.67)$$

To see this, define  $\mathcal{F}_t = \sigma(\dots, u_{t-1}, u_t)$ . Note  $T_j = T_{j-1} \left(1 + \frac{\lambda}{\sqrt{n}} u_j\right)$ . By iterated expectations

$$\begin{aligned} \mathbb{E}(T_n) &= \mathbb{E}(\mathbb{E}(T_n | \mathcal{F}_{n-1})) \\ &= \mathbb{E}\left(T_{n-1} \mathbb{E}\left(1 + \frac{\lambda}{\sqrt{n}} u_n \mid \mathcal{F}_{n-1}\right)\right) \\ &= \mathbb{E}(T_{n-1}) = \dots = \mathbb{E}(T_1) \\ &= 1. \end{aligned}$$

This is (14.67).

The moment generating function of  $S_n$  is

$$\begin{aligned} \mathbb{E}\left(\exp\left(\frac{\lambda}{\sqrt{n}} \sum_{t=1}^n u_t\right)\right) &= \mathbb{E}\left(\prod_{i=1}^n \exp\left(\frac{\lambda}{\sqrt{n}} u_t\right)\right) \\ &\simeq \mathbb{E}\left(\prod_{i=1}^n \left(1 + \frac{\lambda}{\sqrt{n}} u_t\right) \exp\left(\frac{\lambda^2}{2n} u_t^2\right)\right) \end{aligned} \quad (14.68)$$

$$\begin{aligned} &= \mathbb{E}\left(T_n \exp\left(\frac{\lambda^2 V_n}{2}\right)\right) \\ &\simeq \mathbb{E}\left(T_n \exp\left(\frac{\lambda^2 \sigma^2}{2}\right)\right) \end{aligned} \quad (14.69)$$

$$= \exp\left(\frac{\lambda^2 \sigma^2}{2}\right).$$

The approximation in (14.68) is (14.66). The approximation (14.69) is  $V_n \xrightarrow{p} \sigma^2$ . (A rigorous justification which allows this substitution in the expectation is quite technical.) The final equality is (14.67). This calculation shows that the moment generating function of  $S_n$  is approximately that of  $N(0, \sigma^2)$ , as claimed.

The assumption that  $u_t$  is a MDS is critical for (14.67).  $T_n$  is a nonlinear function of the errors  $u_t$  so a white noise assumption cannot be used instead. The MDS assumption is exactly the minimal condition needed to obtain (14.67). This is why the MDS assumption cannot be easily replaced by a milder assumption such as white noise. ■

**Proof of Theorem 14.17.1.** Without loss of generality suppose  $\mathbb{E}(x_t) = 0$  and  $\mathbb{E}(z_t) = 0$ . Set  $\eta = \text{sgn}(\mathbb{E}(z_t | \mathcal{F}_{-\infty}^{t-m}))$ . Then by iterated expectations,  $|x_t| \leq C_1$ ,  $|\mathbb{E}(z_t | \mathcal{F}_{-\infty}^{t-m})| = \eta \mathbb{E}(z_t | \mathcal{F}_{-\infty}^{t-m})$ , and again using iterated expectations

$$\begin{aligned} |\text{cov}(x_{t-m}, z_t)| &= |\mathbb{E}(\mathbb{E}(x_{t-m} z_t | \mathcal{F}_{-\infty}^{t-m}))| \\ &= |\mathbb{E}(x_{t-m} \mathbb{E}(z_t | \mathcal{F}_{-\infty}^{t-m}))| \\ &\leq C_1 \mathbb{E}(|\mathbb{E}(z_t | \mathcal{F}_{-\infty}^{t-m})|) \\ &= C_1 \mathbb{E}(\eta \mathbb{E}(z_t | \mathcal{F}_{-\infty}^{t-m})) \\ &= C_1 \mathbb{E}(\mathbb{E}(\eta z_t | \mathcal{F}_{-\infty}^{t-m})) \\ &= C_1 \mathbb{E}(\eta z_t) \\ &= C_1 \text{cov}(\eta, z_t). \end{aligned} \quad (14.70)$$

Setting  $\xi = \text{sgn}(\mathbb{E}(x_{t-m} | \mathcal{F}_t^\infty))$ , by a similar argument, (14.70) is bounded by  $C_1 C_2 \text{cov}(\eta, \xi)$ . Set  $A_1 = \mathbf{1}(\eta = 1)$ ,  $A_2 = \mathbf{1}(\eta = -1)$ ,  $B_1 = \mathbf{1}(\xi = 1)$ ,  $B_2 = \mathbf{1}(\xi = -1)$ . We calculate

$$\begin{aligned} |\text{cov}(\eta, \xi)| &= |\mathbb{P}(A_1 \cap B_1) + \mathbb{P}(A_2 \cap B_2) - \mathbb{P}(A_2 \cap B_1) - \mathbb{P}(A_1 \cap B_2) \\ &\quad - \mathbb{P}(A_1) \mathbb{P}(B_1) - \mathbb{P}(A_2) \mathbb{P}(B_2) + \mathbb{P}(A_2) \mathbb{P}(B_1) + \mathbb{P}(A_1) \mathbb{P}(B_2)| \\ &\leq 4\alpha(m). \end{aligned}$$

Together,  $|\text{cov}(x_{t-m}, z_t)| \leq 4C_1 C_2 \alpha(m)$  as claimed. ■

**Proof of Theorem 14.17.2.** Assume  $\mathbb{E}(x_t) = 0$  and  $\mathbb{E}(z_t) = 0$ . We first show that if  $|x_t| \leq C$  then

$$|\text{cov}(x_{t-\ell}, z_t)| \leq 6C(\mathbb{E}|z_t|^r)^{1/r} \alpha(\ell)^{1-1/r}. \quad (14.71)$$

Indeed, if  $\alpha(\ell) = 0$  the result is immediate so assume  $\alpha(\ell) > 0$ . Set  $D = \alpha(\ell)^{-1/r} (\mathbb{E}|z_t|^r)^{1/r}$ ,  $v_t = z_t \mathbf{1}(|z_t| \leq D)$  and  $w_t = z_t \mathbf{1}(|z_t| > D)$ . Using the triangle inequality (B.1) and then part 1, since  $|x_t| \leq C$  and  $|v_t| \leq D$ ,

$$\begin{aligned} |\text{cov}(x_{t-\ell}, z_t)| &\leq |\text{cov}(x_{t-\ell}, v_t)| + |\text{cov}(x_{t-\ell}, w_t)| \\ &\leq 4CD\alpha(\ell) + 2C\mathbb{E}|w_t|. \end{aligned}$$

Also,

$$\mathbb{E}|w_t| = \mathbb{E}|z_t \mathbf{1}(|z_t| > D)| = \mathbb{E}\left|\frac{|z_t|^r}{|z_t|^{r-1}} \mathbf{1}(|z_t| > D)\right| \leq \frac{\mathbb{E}|z_t|^r}{D^{r-1}} = \alpha(\ell)^{(r-1)/r} (\mathbb{E}|z_t|^r)^{1/r}$$

using the definition of  $D$ . Together we have

$$|\text{cov}(x_{t-\ell}, z_t)| \leq 6C(\mathbb{E}|x_t|^r)^{1/r} \alpha(\ell)^{1-1/r}.$$

which is (14.71) as claimed.

Now set  $C = \alpha(\ell)^{-1/r} (\mathbb{E}|x_t|^r)^{1/r}$ ,  $v_t = x_t \mathbf{1}(|x_t| \leq C)$  and  $w_t = x_t \mathbf{1}(|x_t| > C)$ . Then using the triangle inequality and (14.71)

$$|\text{cov}(x_{t-\ell}, z_t)| \leq |\text{cov}(v_{t-\ell}, z_t)| + |\text{cov}(w_{t-\ell}, z_t)|.$$

Since  $|v_t| \leq C$ , using (14.71) and the definition of  $C$

$$|\text{cov}(v_{t-\ell}, z_t)| \leq 6C(\mathbb{E}|z_t|^q)^{1/q} \alpha(\ell)^{1-1/q} = 6(\mathbb{E}|x_t|^r)^{1/r} (\mathbb{E}|z_t|^q)^{1/q} \alpha(\ell)^{1-1/q-1/r}.$$

Using Hölder's inequality (B.30) and the definition of  $C$

$$\begin{aligned} |\text{cov}(w_{t-\ell}, z_t)| &\leq 2 \left( \mathbb{E}|w_t|^{q/(q-1)} \right)^{(q-1)/q} (\mathbb{E}|z_t|^q)^{1/q} \\ &= 2 \left( \mathbb{E}(|x_t|^{q/(q-1)} \mathbf{1}(|x_t| > C)) \right)^{(q-1)/q} (\mathbb{E}|z_t|^q)^{1/q} \\ &= 2 \left( \mathbb{E} \left( \frac{|x_t|^r}{|x_t|^{r-q/(q-1)}} \mathbf{1}(|x_t| > C) \right) \right)^{(q-1)/q} (\mathbb{E}|z_t|^q)^{1/q} \\ &\leq \frac{2}{C^{r(q-1)/q-1}} (\mathbb{E}|x_t|^r)^{(q-1)/q} (\mathbb{E}|z_t|^q)^{1/q} \\ &= 2 (\mathbb{E}|x_t|^r)^{1/r} (\mathbb{E}|z_t|^q)^{1/q} \alpha(\ell)^{1-1/q-1/r}. \end{aligned}$$

Together we have

$$|\text{cov}(x_{t-\ell}, z_t)| \leq 8 (\mathbb{E}|x_t|^r)^{1/r} (\mathbb{E}|z_t|^q)^{1/q} \alpha(\ell)^{1-1/r-1/q}$$

as claimed. ■

**Proof of Theorem 14.17.3.** Set  $\eta = \text{sgn}(\mathbb{E}(y_t | \mathcal{F}_{-\infty}^{t-\ell}))$  which satisfies  $|\eta| \leq 1$ . Then since  $\eta$  is  $\mathcal{F}_{-\infty}^{t-\ell}$ -measurable, iterated expectations, using (14.71) with  $C = 1$ , the conditional Jensen's inequality (B.27), and iterated expectations,

$$\begin{aligned} \mathbb{E}|\mathbb{E}(y_t | \mathcal{F}_{-\infty}^{t-\ell})| &= \mathbb{E}(\eta \mathbb{E}(y_t | \mathcal{F}_{-\infty}^{t-\ell})) \\ &= \mathbb{E}(\mathbb{E}(\eta y_t | \mathcal{F}_{-\infty}^{t-\ell})) \\ &= \mathbb{E}(\eta y_t) \\ &\leq 6 (\mathbb{E}|\mathbb{E}(y_t | \mathcal{F}_{-\infty}^{t-\ell})|^r)^{1/r} \alpha(\ell)^{1-1/r} \\ &\leq 6 (\mathbb{E}(\mathbb{E}(|y_t|^r | \mathcal{F}_{-\infty}^{t-\ell})))^{1/r} \alpha(\ell)^{1-1/r} \\ &= 6 (\mathbb{E}|y_t|^r)^{1/r} \alpha(\ell)^{1-1/r} \end{aligned}$$

as claimed. ■

**Proof of Theorem 14.19.** By the Cramér-Wold device (Theorem 6.10) it is sufficient to prove the result for the scalar case. Our proof method is based on a MDS approximation. The trick is to establish the relationship

$$u_t = e_t + z_t - z_{t+1} \quad (14.72)$$

where  $e_t$  is a strictly stationary and ergodic MDS with  $\mathbb{E}(e_t^2) = \Omega$  and  $\mathbb{E}|z_t| < \infty$ . Defining  $S_n^e = \frac{1}{\sqrt{n}} \sum_{t=1}^n e_t$ , we have

$$S_n = \frac{1}{\sqrt{n}} \sum_{t=1}^n (e_t + z_t - z_{t+1}) = S_n^e + \frac{z_1}{\sqrt{n}} - \frac{z_{n+1}}{\sqrt{n}}. \quad (14.73)$$

The first component on the right side is asymptotically  $N(0, \Omega)$  by the MDS CLT (Theorem 14.15), and the second and third terms are  $o_p(1)$  by Markov's inequality (B.35).

The desired relationship (14.72) holds as follows. Set  $\mathcal{F}_t = \sigma(..., u_{t-1}, u_t)$ ,

$$e_t = \sum_{\ell=0}^{\infty} (\mathbb{E}(u_{t+\ell} | \mathcal{F}_t) - \mathbb{E}(u_{t+\ell} | \mathcal{F}_{t-1})) \quad (14.74)$$

and

$$z_t = \sum_{\ell=0}^{\infty} \mathbb{E}(u_{t+\ell} | \mathcal{F}_{t-1}).$$

You can verify that these definitions satisfy (14.72), given  $\mathbb{E}(u_t | \mathcal{F}_t) = u_t$ . The variable  $z_t$  has a finite mean since by the triangle inequality (B.1), Theorem 14.17.3, and the assumptions

$$\begin{aligned}\mathbb{E}|z_t| &= \mathbb{E}\left|\sum_{\ell=0}^{\infty} \mathbb{E}(u_{t+\ell} | \mathcal{F}_{t-1})\right| \\ &\leq 6(\mathbb{E}|u_t|^r)^{1/r} \sum_{\ell=0}^{\infty} \alpha(\ell)^{1-1/r} \\ &< \infty.\end{aligned}$$

Since  $\sum_{\ell=0}^{\infty} \alpha(\ell)^{1-2/r} < \infty$  implies  $\sum_{\ell=0}^{\infty} \alpha(\ell)^{1-1/r} < \infty$ .

The series  $e_t$  in (14.74) has a finite mean by the same calculation as for  $z_t$ . It is a MDS since by iterated expectations

$$\begin{aligned}\mathbb{E}(e_t | \mathcal{F}_{t-1}) &= \mathbb{E}\left(\sum_{\ell=0}^{\infty} (\mathbb{E}(u_{t+\ell} | \mathcal{F}_t) - \mathbb{E}(u_{t+\ell} | \mathcal{F}_{t-1})) | \mathcal{F}_{t-1}\right) \\ &= \sum_{\ell=0}^{\infty} (\mathbb{E}(\mathbb{E}(u_{t+\ell} | \mathcal{F}_t) | \mathcal{F}_{t-1}) - \mathbb{E}(\mathbb{E}(u_{t+\ell} | \mathcal{F}_{t-1}) | \mathcal{F}_{t-1})) \\ &= \sum_{\ell=0}^{\infty} (\mathbb{E}(u_{t+\ell} | \mathcal{F}_{t-1}) - \mathbb{E}(u_{t+\ell} | \mathcal{F}_{t-1})) \\ &= 0.\end{aligned}$$

It is strictly stationary and ergodic by Theorem 14.2 since it is a function of the history  $(..., u_{t-1}, u_t)$ .

The proof is completed by showing that  $e_t$  has a finite variance which equals  $\Omega$ . The trickiest step is to show that  $\text{var}(e_t) < \infty$ . Since

$$\mathbb{E}|S_n| \leq \sqrt{\text{var}(S_n)} \longrightarrow \sqrt{\Omega}$$

(as shown in (14.17)) it follows that  $\mathbb{E}|S_n| \leq 2\sqrt{\Omega}$  for  $n$  sufficiently large. Using (14.73) and  $\mathbb{E}|z_t| < \infty$ , for  $n$  sufficiently large,

$$\mathbb{E}|S_n^e| \leq \mathbb{E}|S_n| + \frac{\mathbb{E}|z_1|}{\sqrt{n}} + \frac{\mathbb{E}|z_{n+1}|}{\sqrt{n}} \leq 3\sqrt{\Omega}. \quad (14.75)$$

Now define  $e_{Bt} = e_t \mathbf{1}(|e_t| \leq B) - \mathbb{E}(e_t \mathbf{1}(|e_t| \leq B) | \mathcal{F}_{t-1})$  which is a bounded MDS. By Theorem 14.15,  $\frac{1}{\sqrt{n}} \sum_{t=1}^n e_{Bt} \xrightarrow{d} N(0, \sigma_B^2)$  where  $\sigma_B^2 = \mathbb{E}(e_{Bt}^2)$ . Since the sequence is also uniformly integrable, this implies

$$\mathbb{E}\left|\frac{1}{\sqrt{n}} \sum_{t=1}^n e_{Bt}\right| \longrightarrow \mathbb{E}|N(0, \sigma_B^2)| = \sqrt{\frac{2}{\pi}} \sigma_B \quad (14.76)$$

using  $\mathbb{E}|N(0, 1)| = 2/\pi$ . We want to show that  $\text{var}(e_t) < \infty$ . Suppose not. Then  $\sigma_B \rightarrow \infty$  as  $B \rightarrow \infty$ , so there will be some  $B$  sufficiently large such that the right-side of (14.76) exceeds the right-side of (14.75). This is a contradiction. We deduce that  $\text{var}(e_t) < \infty$ .

Examining (14.73), we see that since  $\text{var}(S_n) \longrightarrow \Omega < \infty$  and  $\text{var}(S_n^e) = \text{var}(e_t) < \infty$  then  $\text{var}(z_1 - z_{n+1}) / n < \infty$ . Since  $z_t$  is stationary, we deduce that  $\text{var}(z_1 - z_{n+1}) < \infty$ . Equation (14.73) implies  $\text{var}(e_t) = \text{var}(S_n^e) = \text{var}(S_n) + o(1) \longrightarrow \Omega$ . We deduce that  $\text{var}(e_t) = \Omega$  as claimed. ■

**Proof of Theorem 14.21.** Consider the projection of  $y_t$  onto  $(..., e_{t-1}, e_t)$ . Since the projection errors  $e_t$  are uncorrelated, the coefficients of this projection are the bivariate projection coefficients  $b_j = \mathbb{E}(y_t e_{t-j}) / \mathbb{E}(e_{t-j}^2)$ . The leading coefficient is

$$\begin{aligned}b_0 &= \frac{\mathbb{E}(y_t e_t)}{\sigma^2} \\ &= \frac{\sum_{j=1}^{\infty} \alpha_j \mathbb{E}(y_{t-j} e_t) + \mathbb{E}(e_t^2)}{\sigma^2} \\ &= 1\end{aligned}$$

using Theorem 14.20. By Bessel's Inequality (Brockwell and Davis, 1991, Corollary 2.4.1),

$$\sum_{j=1}^{\infty} b_j^2 = \sigma^{-4} \sum_{j=1}^{\infty} (\mathbb{E}(y_t e_t))^2 \leq \sigma^{-4} (\mathbb{E}(y_t^2))^2 < \infty$$

since  $\mathbb{E}(y_t^2) < \infty$  by the assumption of covariance stationarity.

The error from the projection of  $y_t$  onto  $(..., e_{t-1}, e_t)$  is  $\mu_t = y_t - \sum_{j=0}^{\infty} b_j e_{t-j}$ . The fact that this can be written as (14.22) is technical. For the complete argument see Theorem 5.7.1 of Brockwell and Davis (1991). ■

**Proof of Theorem 14.23.** We need to verify that the series  $S_N = \sum_{j=0}^N \theta_j e_{t-j}$  converges almost surely as  $N \rightarrow \infty$ . By the Cauchy criterion for convergence (see Section 14.6), this holds if for all  $\varepsilon > 0$ , there is an  $N < \infty$  such that  $\max_{m \geq 1} |S_{N+m} - S_N| \leq \varepsilon$ . By Kolmogorov's inequality (B.52), since  $e_t$  is a MDS<sup>9</sup> with  $\mathbb{E}(e_t^2) = \sigma^2 < \infty$

$$\begin{aligned} \mathbb{P}\left(\max_{m \geq 1} |S_{N+m} - S_N| > \varepsilon\right) &= \mathbb{P}\left(\max_{m \geq 1} \left| \sum_{j=N+1}^{N+m} \theta_j e_{t-j} \right| > \varepsilon\right) \\ &\leq \varepsilon^{-2} \sum_{j=N+1}^{\infty} \mathbb{E}(\theta_j e_{t-j})^2 \\ &= \sigma^2 \varepsilon^{-2} \sum_{j=N+1}^{\infty} \theta_j^2 \rightarrow 0 \end{aligned}$$

as  $N \rightarrow \infty$  since  $\sum_{j=0}^{\infty} \theta_j^2 < \infty$ . Thus  $y_t = \sum_{j=0}^{\infty} \theta_j e_{t-j}$  converges almost surely.

Since  $e_t$  is strictly stationary and ergodic then  $y_t$  is as well by Theorem 14.9. ■

**Proof of Theorem 14.25.** In the text we showed that  $|\beta_j| < 1$  is sufficient for  $y_t$  to be strictly stationary and ergodic. We now verify that this is equivalent to (14.31)-(14.33). The roots are  $\beta_j = (\alpha_1 \pm \sqrt{\alpha_1^2 + 4\alpha_2})/2$ . Consider separately the cases of real roots and complex roots.

Suppose that the roots are real, which occurs when  $\alpha_1^2 + 4\alpha_2 \geq 0$ . Then  $|\beta_j| < 1$  iff  $|\alpha_1| < 2$  and

$$\frac{\alpha_1 + \sqrt{\alpha_1^2 + 4\alpha_2}}{2} < 1 \quad \text{and} \quad -1 < \frac{\alpha_1 - \sqrt{\alpha_1^2 + 4\alpha_2}}{2}.$$

Equivalently, this holds iff

$$\alpha_1^2 + 4\alpha_2 < (2 - \alpha_1)^2 = 4 - 4\alpha_1 + \alpha_1^2 \quad \text{and} \quad \alpha_1^2 + 4\alpha_2 < (2 + \alpha_1)^2 = 4 + 4\alpha_1 + \alpha_1^2$$

or equivalently iff

$$\alpha_2 < 1 - \alpha_1 \quad \text{and} \quad \alpha_2 < 1 + \alpha_1$$

which are (14.31) and (14.32).  $\alpha_1^2 + 4\alpha_2 \geq 0$  and  $|\alpha_1| < 2$  imply  $\alpha_2 \geq -\alpha_1^2/4 \geq -1$ , which is (14.33).

Now suppose the roots are complex, which occurs when  $\alpha_1^2 + 4\alpha_2 < 0$ . The squared modulus of the roots  $\beta_j = (\alpha_1 \pm \sqrt{\alpha_1^2 + 4\alpha_2})/2$  are

$$|\beta_j|^2 = \left(\frac{\alpha_1}{2}\right)^2 - \left(\frac{\sqrt{\alpha_1^2 + 4\alpha_2}}{2}\right)^2 = -\alpha_2.$$

Thus the requirement  $|\beta_j| < 1$  is satisfied iff  $\alpha_2 > -1$ , which is (14.33).  $\alpha_1^2 + 4\alpha_2 < 0$  and  $\alpha_2 > -1$  imply  $\alpha_1^2 < -4\alpha_2 < 4$ , so  $|\alpha_1| < 2$ .  $\alpha_1^2 + 4\alpha_2 < 0$  and  $|\alpha_1| < 2$  imply  $\alpha_1 + \alpha_2 < \alpha_1 - \alpha_1^2/4 < 1$  and  $\alpha_2 - \alpha_1 < -\alpha_1^2/4 - \alpha_1 < 1$  which are (14.31) and (14.32). ■

<sup>9</sup>We state Kolmogorov's inequality (B.52) under the assumption of independent errors, but it holds more broadly under MDS errors.

**Proof of Theorem 14.27.** The assumption that the roots of  $\alpha(z)$  lie outside the unit circle implies that the factors  $\beta_\ell$  satisfy  $|\beta_\ell| < 1$ . Using the factorization (14.37), and (14.27) under  $|\beta_\ell| < 1$ , we find

$$\begin{aligned}\alpha(z)^{-1} &= \prod_{\ell=1}^p (1 - \beta_\ell z)^{-1} \\ &= \prod_{\ell=1}^p \left( \sum_{j=0}^{\infty} \beta_\ell^j z^j \right) \\ &= \sum_{j=0}^{\infty} \left( \sum_{i_1+\dots+i_p=j} \beta_1^{i_1} \cdots \beta_p^{i_p} \right) z^j \\ &= \sum_{j=0}^{\infty} b_j z^j\end{aligned}$$

with

$$b_j = \sum_{i_1+\dots+i_p=j} \beta_1^{i_1} \cdots \beta_p^{i_p}.$$

Set  $\beta = \max_\ell |\beta_\ell| < 1$ . Using the triangle inequality and the stars and bars theorem (from combinatorics theory)

$$\begin{aligned}|b_j| &\leq \sum_{i_1+\dots+i_p=j} |\beta_1|^{i_1} \cdots |\beta_p|^{i_p} \\ &\leq \sum_{i_1+\dots+i_p=j} \beta^j \\ &\leq \binom{p+j-1}{j} \beta^j \\ &= \frac{(p+j-1)!}{(p-1)! j!} \beta^j \\ &\leq (j+1)^p \beta^j \\ &= O(j^p \beta^j).\end{aligned}$$

From Theorem 14.4.3,  $\sum_{j=0}^{\infty} |b_j| \leq \sum_{j=0}^{\infty} (j+1)^p \beta^j < \infty$  is convergent since  $\beta < 1$ . ■

**Proof of Theorem 14.30.** If  $\mathbf{Q}$  is singular then there is some  $\boldsymbol{\gamma}$  such that  $\boldsymbol{\gamma}' \mathbf{Q} \boldsymbol{\gamma} = 0$ . We can normalize  $\boldsymbol{\gamma}$  to have a unit coefficient on  $y_{t-1}$  (or the first non-zero coefficient other than the intercept). We then have that  $\mathbb{E}(y_{t-1} - (1, y_{t-2}, \dots, y_{t-p}))' \boldsymbol{\phi}^2 = 0$  for some  $\boldsymbol{\phi}$ , or equivalently  $\mathbb{E}(y_t - (1, y_{t-1}, \dots, y_{t-p+1}))' \boldsymbol{\phi}^2 = 0$ . Setting  $\boldsymbol{\beta} = (\boldsymbol{\phi}', 0)'$  this implies  $\mathbb{E}(y_t - \boldsymbol{\beta}' \mathbf{x}_t)^2 = 0$ . Since  $\boldsymbol{\alpha}$  is the best linear predictor we must have  $\boldsymbol{\beta} = \boldsymbol{\alpha}$ . This implies  $\sigma^2 = \mathbb{E}(y_t - \boldsymbol{\alpha}' \mathbf{x}_t)^2 = 0$ . This contradicts the assumption  $\sigma^2 > 0$ . We conclude that  $\mathbf{Q}$  is not singular. ■

## 14.47 Exercises

**Exercise 14.1** For a scalar time series  $y_t$  define the sample autocovariance and autocorrelation

$$\hat{\gamma}(k) = n^{-1} \sum_{t=k+1}^n (y_t - \bar{y})(y_{t-k} - \bar{y})$$

$$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)} = \frac{\sum_{t=k+1}^n (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^n (y_t - \bar{y})^2}.$$

Assume the series is strictly stationary, ergodic, and strictly stationary and  $\mathbb{E}(y_t^2) < \infty$ .

Show that  $\hat{\gamma}(k) \xrightarrow{p} \gamma(k)$  and  $\hat{\rho}(k) \xrightarrow{p} \rho(k)$  as  $n \rightarrow \infty$ . (Use the Ergodic Theorem.)

**Exercise 14.2** Show that if  $(e_t, \mathcal{F}_t)$  is a MDS and  $x_t$  is  $\mathcal{F}_t$ -measurable that  $u_t = x_{t-1}e_t$  is a MDS.

**Exercise 14.3** Let  $\sigma_t^2 = \mathbb{E}(e_t^2 | \mathcal{F}_{t-1})$ . Show that  $u_t = e_t^2 - \sigma_t^2$  is a MDS.

**Exercise 14.4** Continuing the previous exercise, show that if  $\mathbb{E}(e_t^4) < \infty$  then

$$n^{-1/2} \sum_{t=1}^n (e_t^2 - \sigma_t^2) \xrightarrow{d} N(0, v^2).$$

Express  $v^2$  in terms of the moments of  $e_t$ .

**Exercise 14.5** A stochastic volatility model is

$$y_t = \sigma_t e_t$$

$$\log \sigma_t^2 = \omega + \beta \log \sigma_{t-1}^2 + u_t$$

where  $e_t$  and  $u_t$  are independent i.i.d.  $N(0, 1)$  shocks.

(a) Write down an information set for which  $y_t$  is a MDS.

(b) Show that if  $|\beta| < 1$  then  $y_t$  is strictly stationary and ergodic.

**Exercise 14.6** Verify the formula  $\rho(1) = \theta / (1 + \theta^2)$  for a MA(1) process.

**Exercise 14.7** Verify the formula  $\rho(k) = (\sum_{j=0}^{\infty} \theta_{j+k} \theta_j) / (\sum_{j=0}^q \theta_j^2)$  for a MA( $\infty$ ) process.

**Exercise 14.8** Suppose  $y_t = y_{t-1} + e_t$  with  $e_t$  i.i.d.  $(0, 1)$  and  $y_0 = 0$ . Find  $\text{var}(y_t)$ . Is  $y_t$  stationary?

**Exercise 14.9** Take the AR(1) model with no intercept  $y_t = \alpha_1 y_{t-1} + e_t$ .

(a) Find the impulse response function  $b_j = \frac{\partial}{\partial e_t} y_{t+j}$ .

(b) Let  $\hat{\alpha}_1$  be the least squares estimator of  $\alpha_1$ . Find the estimator of  $b_j$ .

(c) Let  $s(\hat{\alpha}_1)$  be a standard error for  $\hat{\alpha}_1$ . Use the delta method to find a 95% asymptotic confidence interval for  $b_j$ .

**Exercise 14.10** Take the AR(2) model  $y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + e_t$ .

(a) Find expressions for the impulse responses  $b_1, b_2, b_3$  and  $b_4$ .

(b) Let  $(\hat{\alpha}_1, \hat{\alpha}_2)$  be the least squares estimator. Find the estimator of  $b_2$ .

- (c) Let  $\hat{V}$  be the estimated covariance matrix for the coefficients. Use the delta method to find a 95% asymptotic confidence interval for  $b_2$ .

**Exercise 14.11** Show that the models

$$\alpha(L)y_t = \alpha_0 + e_t$$

and

$$\begin{aligned}\alpha(L)y_t &= \mu + u_t \\ \alpha(L)u_t &= e_t\end{aligned}$$

are identical. Find an expression for  $\mu$  in terms of  $\alpha_0$  and  $\alpha(L)$ .

**Exercise 14.12** Take the model

$$\begin{aligned}\alpha(L)y_t &= u_t \\ \beta(L)u_t &= e_t\end{aligned}$$

where  $\alpha(L)$  and  $\beta(L)$  are  $p$  and  $q$  order lag polynomials, respectively. Show that these equations imply that

$$\gamma(L)y_t = e_t$$

for some lag polynomial  $\gamma(L)$ . What is the order of  $\gamma(L)$ ?

**Exercise 14.13** Suppose that

$$\begin{aligned}y_t &= u_t + e_t \\ u_t &= v_t + \theta v_{t-1}\end{aligned}$$

where  $u_t$  and  $e_t$  are mutually independent i.i.d. mean zero processes. Show that  $y_t$  is a MA(1) process

$$y_t = \eta_t + \psi \eta_{t-1}$$

for an i.i.d. error  $\eta_t$ . Find an expression for  $\psi$ .

**Exercise 14.14** Suppose that

$$\begin{aligned}y_t &= x_t + e_t \\ x_t &= \alpha x_{t-1} + u_t\end{aligned}$$

where the errors  $e_t$  and  $u_t$  are mutually independent i.i.d. processes. Show that  $y_t$  is an ARMA(1,1) process.

**Exercise 14.15** A Gaussian AR model is an autoregression with i.i.d.  $N(0, \sigma^2)$  errors. Consider the Gaussian AR(1) model

$$\begin{aligned}y_t &= \alpha_0 + \alpha_1 y_{t-1} + e_t \\ e_t &\sim N(0, \sigma^2)\end{aligned}$$

with  $|\alpha_1| < 1$ . Show that the marginal distribution of  $y_t$  is also normal:

$$y_t \sim N\left(\frac{\alpha_0}{1 - \alpha_1}, \frac{\sigma^2}{1 - \alpha_1^2}\right).$$

Hint: Use the MA representation of  $y_t$ .

**Exercise 14.16** Assume that  $y_t$  is a Gaussian AR(1) as in the previous exercise. Calculate the moments

$$\begin{aligned}\mu &= \mathbb{E}(y_t) \\ \sigma_y^2 &= \mathbb{E}(y_t - \mu)^2 \\ \kappa &= \mathbb{E}(y_t - \mu)^4\end{aligned}$$

A colleague suggests estimating the parameters  $(\alpha_0, \alpha_1, \sigma^2)$  of the Gaussian AR(1) model by GMM applied to the corresponding sample moments. He points out that there are three moments and three parameters, so it should be identified. Can you find a flaw in his approach?

Hint: This is subtle.

**Exercise 14.17** Take the nonlinear process

$$y_t = y_{t-1}^\alpha u_t^{1-\alpha}$$

where  $u_t$  is i.i.d. with strictly positive support.

- (a) Find the condition under which  $y_t$  is strictly stationary and ergodic.
- (b) Find an explicit expression for  $y_t$  as a function of  $(u_t, u_{t-1}, \dots)$ .

**Exercise 14.18** Take the quarterly series *pnpfix* (nonresidential real private fixed investment) from FRED-QD.

- (a) Transform the series into quarterly growth rates.
- (b) Estimate an AR(4) model. Report using heteroskedastic-consistent standard errors.
- (c) Repeat using the Newey-West standard errors, using  $M = 5$ .
- (d) Comment on the magnitude and interpretation of the coefficients.
- (e) Calculate (numerically) the impulse responses for  $j = 1, \dots, 10$ .

**Exercise 14.19** Take the quarterly series *oilpricex* (real price of crude oil) from FRED-QD.

- (a) Transform the series by taking first differences.
- (b) Estimate an AR(4) model. Report using heteroskedastic-consistent standard errors.
- (c) Test the hypothesis that the real oil prices is a random walk by testing that the four AR coefficients jointly equal zero.
- (d) Interpret the coefficient estimates and test result.

**Exercise 14.20** Take the monthly series *unrate* (unemployment rate) from FRED-MD.

- (a) Estimate AR(1) through AR(8) models, using the sample starting in 1960m1 so that all models use the same observations.
- (b) Compute the AIC for each AR model and report.
- (c) Which AR model has the lowest AIC?
- (d) Report the coefficient estimates and standard errors for the selected model.

**Exercise 14.21** Take the quarterly series *unrate* (unemployment rate) and *claimsx* (initial claims) from FRED-QD. “Initial claims” are the number of individuals who file for unemployment insurance.

- (a) Estimate a distributed lag regression of the unemployment rate on initial claims. Use lags 1 through 4. Which standard error method is appropriate?
- (b) Estimate an autoregressive distributed lag regression of the unemployment rate on initial claims. Use lags 1 through 4 for both variables.
- (c) Test the hypothesis that initial claims does not Granger cause the unemployment rate.
- (d) Interpret your results.

**Exercise 14.22** Take the quarterly series *gdpc1* (real GDP) and *houst* (housing starts) from FRED-QD. “Housing starts” are the number of new houses which on which construction is started.

- (a) Transform the real GDP series into its one quarter growth rate.
- (b) Estimate a distributed lag regression of GDP growth on housing starts. Use lags 1 through 4. Which standard error method is appropriate?
- (c) Estimate an autoregressive distributed lag regression of GDP growth on housing starts. Use lags 1 through 2 for GDP growth and 1 through 4 for housing starts.
- (d) Test the hypothesis that housing starts does not Granger cause GDP growth.
- (e) Interpret your results.

# Chapter 15

## Multivariate Time Series

### 15.1 Introduction

A multivariate time series  $\mathbf{y}_t = (y_{1t}, \dots, y_{mt})'$  is an  $m \times 1$  vector process observed in sequence over time,  $t = 1, \dots, n$ . Multivariate time series models primarily focus on the joint modeling of the vector series  $\mathbf{y}_t$ . The most common multivariate time series models used by economists are vector autoregressions (VARs). VARs were introduced to econometrics by Sims (1980).

Some excellent textbooks and review articles on multivariate time series include Hamilton (1994), Watson (1994), Canova (1995), Lütkepohl (2005), Ramey (2016), Stock and Watson (2006), and Kilian and Lütkepohl (2017).

### 15.2 Multiple Equation Time Series Models

To motivate vector autoregressions let us start by reviewing the autoregressive distributed lag model of Section 14.40 for the case of two series  $\mathbf{y}_t = (y_{1t}, y_{2t})'$  with a single lag. An AR-DL model for  $y_{1t}$  takes the form

$$y_{1t} = \alpha_0 + \alpha_1 y_{1t-1} + \beta_1 y_{2t-1} + e_{1t}.$$

Similarly, an AR-DL model for  $y_{2t}$  takes the form

$$y_{2t} = \gamma_0 + \gamma_1 y_{2t-1} + \delta_1 y_{1t-1} + e_{2t}.$$

These two equations specify that each variable is a linear function of its own lag and the lag of the other variable. In so doing, we find that the variables on the right hand side of each equation are identical and equal  $\mathbf{y}_{t-1}$ .

We can simplify the equations by combining the regressors, stacking the two equations together, and writing the vector error as  $\mathbf{e}_t = (e_{1t}, e_{2t})'$  to find

$$\mathbf{y}_t = \mathbf{a}_0 + \mathbf{A}_1 \mathbf{y}_{t-1} + \mathbf{e}_t$$

where  $\mathbf{a}_0$  is  $2 \times 1$  and  $\mathbf{A}_1$  is  $2 \times 2$ . This is a bivariate vector autoregressive model for  $\mathbf{y}_t$ . It specifies that the multivariate process  $\mathbf{y}_t$  is a linear function of its own lag  $\mathbf{y}_{t-1}$  plus the  $\mathbf{e}_t$ . It is the combination of two equations, each of which is an autoregressive distributed lag model. Thus a multivariate autoregression is simply a set of autoregressive distributed lag models.

The above derivation assumed a single lag. If the equations include  $p$  lags of each variable, we obtain the  $p^{th}$  order **vector autoregressive (VAR)** model

$$\mathbf{y}_t = \mathbf{a}_0 + \mathbf{A}_1 \mathbf{y}_{t-1} + \mathbf{A}_2 \mathbf{y}_{t-2} + \cdots + \mathbf{A}_p \mathbf{y}_{t-p} + \mathbf{e}_t. \quad (15.1)$$

This is a bivariate vector autoregressive model for  $\mathbf{y}_t$ .

Furthermore, there is nothing special about the two variable case. The notation in (15.1) allows  $\mathbf{y}_t$  to be a vector of dimension  $m$ , in which case the matrices  $\mathbf{A}_\ell$  are  $m \times m$  and the error  $\mathbf{e}_t$  is  $m \times 1$ . We will denote the elements of  $\mathbf{A}_\ell$  using the notation

$$\mathbf{A}_\ell = \begin{bmatrix} a_{11,\ell} & a_{12,\ell} & \cdots & a_{1m,\ell} \\ a_{21,\ell} & a_{22,\ell} & \cdots & a_{2m,\ell} \\ \vdots & \vdots & & \vdots \\ a_{m1,\ell} & a_{m2,\ell} & \cdots & a_{mm,\ell} \end{bmatrix}.$$

The error  $\mathbf{e}_t = (e_{1t}, \dots, e_{mt})'$  is the component of  $\mathbf{y}_t = (y_{1t}, \dots, y_{mt})'$  which is unforecastable at time  $t-1$ . However, the components of  $\mathbf{e}_t$  are contemporaneously correlated. Therefore the contemporaneous covariance matrix

$$\Sigma = \mathbb{E}(\mathbf{e}_t \mathbf{e}'_t)$$

is non-diagonal.

The VAR model falls in the class of multivariate regression models studied in Chapter 11.

In the following several sections we take a step back and provide a rigorous foundation for vector autoregressions for stationary time series.

### 15.3 Linear Projection

In Section 14.14 we derived the linear projection of the univariate series  $y_t$  on its infinite past history. We now extend this to the multivariate case. Define the multivariate infinite past history  $\tilde{\mathbf{y}}_{t-1} = (\dots, \mathbf{y}_{t-2}, \mathbf{y}_{t-1})$ . The best linear predictor of each component of  $\mathbf{y}_t$  is linear in the lags  $\mathbf{y}_{t-\ell}$ . Stacking together we obtain the linear projection of the vector  $\mathbf{y}_t$  on its past history

$$\mathcal{P}_{t-1}(\mathbf{y}_t) = \mathcal{P}(\mathbf{y}_t | \tilde{\mathbf{y}}_{t-1}) = \mathbf{a}_0 + \sum_{\ell=1}^{\infty} \mathbf{A}_\ell \mathbf{y}_{t-\ell}.$$

The projection error is the difference

$$\mathbf{e}_t = \mathbf{y}_t - \mathcal{P}_{t-1}(\mathbf{y}_t) \quad (15.2)$$

giving rise to the regression equation

$$\mathbf{y}_t = \mathbf{a}_0 + \sum_{\ell=1}^{\infty} \mathbf{A}_\ell \mathbf{y}_{t-\ell} + \mathbf{e}_t. \quad (15.3)$$

We will typically call the projection errors  $\mathbf{e}_t$  the “innovations”.

The innovations  $\mathbf{e}_t$  are mean zero, uncorrelated with lagged  $\mathbf{y}_{t-1}$ , and are serially uncorrelated. We state this formally.

**Theorem 15.1** If  $\mathbf{y}_t$  is covariance stationary it has the projection equation

$$\mathbf{y}_t = \mathbf{a}_0 + \sum_{\ell=1}^{\infty} \mathbf{A}_\ell \mathbf{y}_{t-\ell} + \mathbf{e}_t.$$

The innovations  $\mathbf{e}_t$  satisfy

$$\begin{aligned} \mathbb{E}(\mathbf{e}_t) &= 0 \\ \mathbb{E}(\mathbf{y}_{t-\ell} \mathbf{e}'_t) &= 0 \quad \ell \geq 1 \\ \mathbb{E}(\mathbf{e}_{t-\ell} \mathbf{e}'_t) &= 0 \quad \ell \geq 1 \end{aligned}$$

and

$$\Sigma = \mathbb{E}(\mathbf{e}_t \mathbf{e}'_t) < \infty.$$

If  $\mathbf{y}_t$  is strictly stationary then  $\mathbf{e}_t$  is strictly stationary.

We can write the model using the lag operator notation as

$$\mathbf{A}(\mathbf{L}) \mathbf{y}_t = \boldsymbol{\alpha}_0 + \mathbf{e}_t$$

where

$$\mathbf{A}(z) = \mathbf{I}_m - \sum_{\ell=1}^{\infty} \mathbf{A}_{\ell} z^{\ell}.$$

The multivariate innovations  $\mathbf{e}_t$  are mean zero and serially uncorrelated. This describes what is known as a multivariate white noise process.

**Definition 15.1** The vector process  $\mathbf{e}_t$  is **multivariate white noise** if  $\mathbb{E}(\mathbf{e}_t) = 0$ ,  $\mathbb{E}(\mathbf{e}_t \mathbf{e}'_t) = \Sigma < \infty$ , and  $\mathbb{E}(\mathbf{e}_t \mathbf{e}'_{t-\ell}) = 0$  for  $\ell \neq 0$ .

## 15.4 Multivariate Wold Decomposition

By projecting  $\mathbf{y}_t$  onto the past history of the white noise innovations  $\mathbf{e}_t$  we obtain a multivariate version of the Wold decomposition.

**Theorem 15.2** If  $\mathbf{y}_t$  is covariance stationary and non-deterministic then it has the linear representation

$$\mathbf{y}_t = \boldsymbol{\mu} + \sum_{\ell=0}^{\infty} \boldsymbol{\Theta}_{\ell} \mathbf{e}_{t-\ell} \quad (15.4)$$

where  $\mathbf{e}_t$  are the white noise projection errors and  $\boldsymbol{\Theta}_0 = \mathbf{I}_m$ . The coefficient matrices  $\boldsymbol{\Theta}_{\ell}$  are  $m \times m$ .

We can write the moving average representation using the lag operator notation as

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Theta}(\mathbf{L}) \mathbf{e}_t$$

where

$$\boldsymbol{\Theta}(z) = \sum_{\ell=0}^{\infty} \boldsymbol{\Theta}_{\ell} z^{\ell}.$$

If invertible, the lag polynomials satisfy the relationships  $\boldsymbol{\Theta}(z) = \mathbf{A}(z)^{-1}$  and  $\mathbf{A}(z) = \boldsymbol{\Theta}(z)^{-1}$ .

For some purposes (such as impulse response calculations) we need to calculate the moving average coefficient matrices  $\boldsymbol{\Theta}_{\ell}$  from the projection coefficient matrices  $\mathbf{A}_{\ell}$ . While there is not a closed-form solution there is a simple recursion by which the coefficients may be calculated.

**Theorem 15.3** For  $j \geq 1$

$$\boldsymbol{\Theta}_j = \sum_{\ell=1}^j \mathbf{A}_{\ell} \boldsymbol{\Theta}_{j-\ell}$$

To see this, suppose for simplicity  $\boldsymbol{a}_0 = 0$  and that the innovations satisfy  $e_t = 0$  for  $t \neq 0$ . Then  $\mathbf{y}_t = 0$  for  $t < 0$ . Using the regression equation (15.3) for  $t \geq 0$  we solve for each  $\mathbf{y}_t$ . For  $t = 0$

$$\mathbf{y}_0 = \mathbf{e}_0 = \boldsymbol{\Theta}_0 \mathbf{e}_0$$

where

$$\boldsymbol{\Theta}_0 = \mathbf{I}_m.$$

For  $t = 1$

$$\mathbf{y}_1 = \mathbf{A}_1 \mathbf{y}_0 = \mathbf{A}_1 \boldsymbol{\Theta}_0 \mathbf{e}_0 = \boldsymbol{\Theta}_1 \mathbf{e}_0$$

where

$$\boldsymbol{\Theta}_1 = \mathbf{A}_1 \boldsymbol{\Theta}_0.$$

For  $t = 2$

$$\mathbf{y}_2 = \mathbf{A}_1 \mathbf{y}_1 + \mathbf{A}_2 \mathbf{y}_0 = \mathbf{A}_1 \boldsymbol{\Theta}_1 \mathbf{e}_0 + \mathbf{A}_2 \boldsymbol{\Theta}_0 \mathbf{e}_0 = \boldsymbol{\Theta}_2 \mathbf{e}_0$$

where

$$\boldsymbol{\Theta}_2 = \mathbf{A}_1 \boldsymbol{\Theta}_1 + \mathbf{A}_2 \boldsymbol{\Theta}_0.$$

For  $t = 3$

$$\mathbf{y}_3 = \mathbf{A}_1 \mathbf{y}_2 + \mathbf{A}_2 \mathbf{y}_1 + \mathbf{A}_3 \mathbf{y}_0 = \mathbf{A}_1 \boldsymbol{\Theta}_2 \mathbf{e}_0 + \mathbf{A}_2 \boldsymbol{\Theta}_1 \mathbf{e}_0 + \mathbf{A}_3 \boldsymbol{\Theta}_0 \mathbf{e}_0 = \boldsymbol{\Theta}_3 \mathbf{e}_0$$

where

$$\boldsymbol{\Theta}_3 = \mathbf{A}_1 \boldsymbol{\Theta}_2 + \mathbf{A}_2 \boldsymbol{\Theta}_1 + \mathbf{A}_3 \boldsymbol{\Theta}_0.$$

The coefficients satisfy the stated recursion as claimed.

## 15.5 Impulse Response

One of the most important concepts in applied multivariate time series is the **impulse response function (IRF)**, which is defined as the change in  $\mathbf{y}_t$  due to a change in an innovation or shock. In this section we define the baseline IRF – the unnormalized non-orthogonalized impulse response function – which is the change in  $\mathbf{y}_t$  due to a change in an innovation  $\mathbf{e}_t$ . Specifically, we define the impulse response of variable  $i$  with respect to innovation  $j$  as the change in the time  $t$  projection of the  $i^{th}$  variable  $y_{it+h}$  due to the  $j^{th}$  innovation  $e_{jt}$

$$\text{IRF}_{ij}(h) = \frac{\partial}{\partial e_{jt}} \mathcal{P}_t(y_{it+h}).$$

There are  $m^2$  such responses for each horizon  $h$ . We can write them as an  $m \times m$  matrix

$$\text{IRF}(h) = \frac{\partial}{\partial \mathbf{e}'_t} \mathcal{P}_t(\mathbf{y}_{t+h}).$$

Recall the multivariate Wold representation

$$\mathbf{y}_t = \boldsymbol{\mu} + \sum_{\ell=0}^{\infty} \boldsymbol{\Theta}_\ell \mathbf{e}_{t-\ell}.$$

We can calculate that the projection onto the history at time  $t$  is

$$\mathcal{P}_t(\mathbf{y}_{t+h}) = \boldsymbol{\mu} + \sum_{\ell=h}^{\infty} \boldsymbol{\Theta}_\ell \mathbf{e}_{t+h-\ell} = \boldsymbol{\mu} + \sum_{\ell=0}^{\infty} \boldsymbol{\Theta}_{h+\ell} \mathbf{e}_{t-\ell}.$$

We deduce that the impulse response matrix is

$$\text{IRF}(h) = \boldsymbol{\Theta}_h$$

the  $h^{th}$  moving average coefficient matrix. The individual impulse response is

$$\text{IRF}_{ij}(h) = \Theta_{h,ij}$$

the  $i,j^{th}$  element of  $\Theta_h$ .

Here we have defined the impulse response in terms of the linear projection operator. An alternative is to define the impulse response in terms of the conditional expectation operator. The two coincide when the innovations  $\epsilon_t$  are a martingale difference sequence (and thus when the true process is linear) but otherwise will not coincide.

Typically we view impulse responses as a function of the horizon  $h$ , and plot them as a function of  $h$  for each pair  $(i,j)$ . The impulse response function  $\text{IRF}_{ij}(h)$  is interpreted as how the  $i^{th}$  variable responds over time to the  $j^{th}$  innovation.

In a linear vector autoregression, the impulse response function is symmetric in negative and positive innovations. That is, the impact on  $y_{it+h}$  of a positive innovation  $e_{jt} = 1$  is  $\text{IRF}_{ij}(h)$  and the impact of a negative innovation  $e_{jt} = -1$  is  $-\text{IRF}_{ij}(h)$ . Furthermore, the magnitude of the impact is linear in the magnitude of the innovation. Thus the impact of the innovation  $e_{jt} = 2$  is  $2\text{IRF}_{ij}(h)$  and the impact of the innovation  $e_{jt} = -2$  is  $-2\text{IRF}_{ij}(h)$ . This means that the shape of the impulse response function is unaffected by the magnitude of the innovation. (These are consequences of the linearity of the vector autoregressive model, not necessarily features of the true world.)

The impulse response functions can be scaled as desired. One standard choice is to scale so that the innovations correspond to one unit of the impulse variable. Thus if the impulse variable is measured in dollars, the impulse response can be scaled to correspond to a change in \$1 or some multiple such as a million dollars. If the impulse variable is measured in percentage points (e.g. an interest rate) then the impulse response can be scaled to correspond to a change of one full percentage point (e.g. from 3% to 4%) or to correspond to a change of one basis point (e.g. from 3.05% to 3.06%). Another standard choice is to scale the impulse responses to correspond to a “one standard deviation” innovation. This occurs when the innovations have been scaled to have unit variances. In this latter case impulse response functions can be interpreted as responses due to a “typical” sized (one standard deviation) innovation.

Closely related is the **cumulative impulse response function (CIRF)**, defined as

$$\text{CIRF}(h) = \sum_{\ell=1}^h \frac{\partial}{\partial \epsilon'_t} \mathcal{P}_t(y_{t+\ell}) = \sum_{\ell=1}^h \Theta_\ell.$$

The cumulative impulse response is the accumulated (summed) responses on  $y_t$  from time  $t$  to  $t+h$ . The limit of the cumulative impulse response as  $h \rightarrow \infty$  is the **long-run impulse response**

$$C = \lim_{h \rightarrow \infty} \text{CIRF}(h) = \sum_{\ell=1}^{\infty} \Theta_\ell = \Theta(1) = A(1)^{-1}.$$

This is the full (summed) effect of the innovation, over all time.

It is useful to observe that when a VAR is estimated on differenced observations  $\Delta y_t$  then cumulative impulse response is

$$\text{CIRF}(h) = \frac{\partial}{\partial \epsilon'_t} \mathcal{P}_t \left( \sum_{\ell=1}^h \Delta y_{t+\ell} \right) = \frac{\partial}{\partial \epsilon'_t} \mathcal{P}_t(y_{t+h})$$

which is the impulse response function for the variable  $y_t$  in levels. More generally, when a VAR is estimated with some variables in levels and some in differences, then the cumulative impulse response function for the second group will coincide with the impulse responses for the same variables measured in levels.

It is typical to report cumulative impulse response functions (rather than impulse response functions) for variables which enter a VAR in differences. In fact, many authors will label a cumulative impulse response as “the impulse response”.

## 15.6 VAR(1) Model

The **first-order vector autoregressive process**, denoted **VAR(1)**, is

$$\mathbf{y}_t = \boldsymbol{\alpha}_0 + \mathbf{A}_1 \mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t$$

where  $\boldsymbol{\epsilon}_t$  is a strictly stationary and ergodic white noise process.

We are interested in conditions under which  $\mathbf{y}_t$  is a stationary process. Let  $\lambda_{\max}(\mathbf{A})$  be the largest absolute eigenvalue of  $\mathbf{A}$ .

**Theorem 15.4** If  $\lambda_{\max}(\mathbf{A}_1) < 1$  then the VAR(1) process  $\mathbf{y}_t$  is strictly stationary and ergodic.

The VAR(1) generalizes the AR(1) model to multivariate systems. The dynamics of a VAR(1) can be considerably more involved than those of an AR(1).

The proof of Theorem 15.4 follows from the following technical result by applying back-substitution to write  $\mathbf{y}_t = \sum_{\ell=0}^{\infty} \mathbf{A}_1^\ell (\boldsymbol{\alpha}_0 + \boldsymbol{\epsilon}_{t-\ell})$ .

**Theorem 15.5** Suppose  $\lambda_{\max}(\mathbf{A}) < 1$ ,  $\mathbf{u}_t$  is strictly stationary and ergodic, and  $\mathbb{E} \|\mathbf{u}_t\| < \infty$ . Then  $\mathbf{x}_t = \sum_{\ell=0}^{\infty} \mathbf{A}^\ell \mathbf{u}_{t-\ell}$  is convergent with probability one, and is strictly stationary and ergodic.

The proof is given in Section 15.31.

## 15.7 VAR(p) Model

The  $p^{th}$ -order vector autoregressive process, denoted **VAR(p)**, is

$$\mathbf{y}_t = \boldsymbol{\alpha}_0 + \mathbf{A}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{A}_p \mathbf{y}_{t-p} + \boldsymbol{\epsilon}_t$$

where  $\boldsymbol{\epsilon}_t$  is a strictly stationary and ergodic white noise process.

We can write the model using the lag operator notation as

$$\mathbf{A}(\mathbf{L}) \mathbf{y}_t = \boldsymbol{\alpha}_0 + \boldsymbol{\epsilon}_t$$

where

$$\mathbf{A}(z) = \mathbf{I}_m - \mathbf{A}_1 z - \cdots - \mathbf{A}_p z^p.$$

The condition for stationarity of the system can be expressed as a restriction on the roots of the determinantal equation.

**Theorem 15.6** If all roots  $\lambda$  of  $\det(\mathbf{A}(z)) = 0$  satisfy  $|\lambda| > 1$  then the VAR(p) process  $\mathbf{y}_t$  is strictly stationary and ergodic.

The proof is given in Section 15.31.

## 15.8 Regression Notation

Defining the  $(mp + 1) \times 1$  vector

$$\mathbf{x}_t = \begin{pmatrix} 1 \\ \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \vdots \\ \mathbf{y}_{t-p} \end{pmatrix}$$

and the  $m \times (mp + 1)$  matrix

$$\mathbf{A}' = (\mathbf{a}_0 \quad \mathbf{a}_1 \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_p).$$

Then the VAR system of equations can be written as

$$\mathbf{y}_t = \mathbf{A}' \mathbf{x}_t + \mathbf{e}_t. \quad (15.5)$$

This is a multivariate regression model. The error has covariance matrix

$$\Sigma = \mathbb{E}(\mathbf{e}_t \mathbf{e}_t'). \quad (15.6)$$

We can also write the coefficient matrix as

$$\mathbf{A} = (\mathbf{a}_1 \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_m)$$

where  $\mathbf{a}_j$  is the vector of coefficients for the  $j^{th}$  equation. Thus

$$y_{jt} = \mathbf{a}'_j \mathbf{x}_t + e_{jt}.$$

In general, if  $\mathbf{y}_t$  is strictly stationary we can define the coefficient matrix  $\mathbf{A}$  by linear projection.

$$\mathbf{A} = (\mathbb{E}(\mathbf{x}_t \mathbf{x}'_t))^{-1} \mathbb{E}(\mathbf{x}_t \mathbf{y}'_t).$$

This holds whether or not  $\mathbf{y}_t$  is actually a VAR(p) process. By the properties of projection errors

$$\mathbb{E}(\mathbf{x}_t \mathbf{e}'_t) = 0. \quad (15.7)$$

The projection coefficient matrix  $\mathbf{A}$  is identified if  $\mathbb{E}(\mathbf{x}_t \mathbf{x}'_t)$  is invertible.

**Theorem 15.7** If  $\mathbf{y}_t$  is strictly stationary and  $0 < \Sigma < \infty$  for  $\Sigma$  defined in (15.6), then  $\mathbf{Q} = \mathbb{E}(\mathbf{x}_t \mathbf{x}'_t) > 0$  and the coefficient vector (14.44) is identified.

The proof is given in Section 15.31.

## 15.9 Estimation

From Chapter 11, the systems estimator of a multivariate regression is least squares. The estimator can be written as

$$\widehat{\mathbf{A}} = \left( \sum_{t=1}^n \mathbf{x}_t \mathbf{x}'_t \right)^{-1} \left( \sum_{t=1}^n \mathbf{x}_t \mathbf{y}'_t \right).$$

Alternatively, the coefficient estimator for the  $j^{th}$  equation is

$$\hat{\boldsymbol{a}}_j = \left( \sum_{t=1}^n \mathbf{x}_t \mathbf{x}'_t \right)^{-1} \left( \sum_{t=1}^n \mathbf{x}_t y_{jt} \right).$$

The least squares residual vector is

$$\hat{\boldsymbol{e}}_t = \mathbf{y}_t - \hat{\mathbf{A}}' \mathbf{x}_t.$$

The estimator of the variance matrix is

$$\hat{\Sigma} = \frac{1}{n} \sum_{t=1}^n \hat{\boldsymbol{e}}_t \hat{\boldsymbol{e}}'_t. \quad (15.8)$$

(This may be adjusted for degrees-of-freedom if desired, but there is no established finite-sample justification for a specific adjustment.)

If  $\mathbf{y}_t$  is strictly stationary and ergodic with finite variances then we can apply the Ergodic Theorem (Theorem 14.13) to deduce that

$$\frac{1}{n} \sum_{t=1}^n \mathbf{x}_t \mathbf{y}'_t \xrightarrow{p} \mathbb{E}(\mathbf{x}_t \mathbf{y}'_t)$$

and

$$\sum_{t=1}^n \mathbf{x}_t \mathbf{x}'_t \xrightarrow{p} \mathbb{E}(\mathbf{x}_t \mathbf{x}'_t).$$

Since the latter is positive definite by Theorem 15.7, we conclude that  $\hat{\mathbf{A}}$  is consistent for  $\mathbf{A}$ . Standard manipulations show that  $\hat{\Sigma}$  is consistent as well.

**Theorem 15.8** If  $\mathbf{y}_t$  is strictly stationary and ergodic and  $0 < \Sigma < \infty$  then  $\hat{\mathbf{A}} \xrightarrow{p} \mathbf{A}$  and  $\hat{\Sigma} \xrightarrow{p} \Sigma$  as  $n \rightarrow \infty$ .

VAR models can be estimated in Stata using the `var` command.

## 15.10 Asymptotic Distribution

Set

$$\boldsymbol{a} = \text{vec}(\mathbf{A}) = \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_m \end{pmatrix}, \quad \hat{\boldsymbol{a}} = \text{vec}(\hat{\mathbf{A}}) = \begin{pmatrix} \hat{\mathbf{a}}_1 \\ \vdots \\ \hat{\mathbf{a}}_m \end{pmatrix}.$$

By the same analysis as in Theorem 14.33 combined with Theorem 11.1 we obtain the following.

**Theorem 15.9** If  $\mathbf{y}_t$  follows the VAR( $p$ ) model with  $\mathbb{E}(\boldsymbol{e}_t | \mathcal{F}_{t-1}) = \mathbf{0}$ ,  $\mathbb{E} \|\mathbf{y}_t\|^4 < \infty$ , and  $\Sigma > 0$ , then as  $n \rightarrow \infty$ ,

$$\sqrt{n} (\hat{\boldsymbol{a}} - \boldsymbol{a}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V})$$

where

$$\begin{aligned} \mathbf{V} &= \bar{\mathbf{Q}}^{-1} \boldsymbol{\Omega} \bar{\mathbf{Q}}^{-1} \\ \bar{\mathbf{Q}} &= \mathbf{I}_m \otimes \mathbf{Q} \\ \mathbf{Q} &= \mathbb{E}(\mathbf{x}_t \mathbf{x}'_t) \\ \boldsymbol{\Omega} &= \mathbb{E}(\boldsymbol{e}_t \boldsymbol{e}'_t \otimes \mathbf{x}_t \mathbf{x}'_t). \end{aligned}$$

Notice that we added the stronger assumption that the innovation is a martingale difference sequence  $\mathbb{E}(\mathbf{e}_t | \mathcal{F}_{t-1}) = 0$ . This means that this distributional result assumes that the VAR(p) model is the correct conditional mean for each variable. In words, these are the correct lags and there is no omitted nonlinearity.

If we further strengthen the MDS assumption to conditional homoskedasticity

$$\mathbb{E}(\mathbf{e}_t \mathbf{e}'_t | \mathcal{F}_{t-1}) = \Sigma$$

then the asymptotic variance simplifies as

$$\begin{aligned}\Omega &= \Sigma \otimes Q \\ V &= \Sigma \otimes Q^{-1}.\end{aligned}$$

In contrast, if the VAR(p) is an approximation then the MDS assumption is not appropriate. In this case the asymptotic distribution can be derived under mixing conditions.

**Theorem 15.10** If  $\mathbf{y}_t$  is strictly stationary, ergodic,  $\Sigma > 0$ , and for some  $r > 4$ ,  $\mathbb{E} \|\mathbf{y}_t\|^r < \infty$  and the mixing coefficients satisfy  $\sum_{\ell=1}^{\infty} \alpha(\ell)^{1-4/r} < \infty$ , then as  $n \rightarrow \infty$ ,

$$\sqrt{n}(\hat{\mathbf{a}} - \mathbf{a}) \xrightarrow{d} N(\mathbf{0}, V)$$

where

$$\begin{aligned}V &= \bar{Q}^{-1} \Omega \bar{Q}^{-1} \\ \bar{Q} &= I_m \otimes Q \\ Q &= \mathbb{E}(\mathbf{x}_t \mathbf{x}'_t) \\ \Omega &= \sum_{\ell=-\infty}^{\infty} \mathbb{E}(\mathbf{e}_{t-\ell} \mathbf{e}'_{t-\ell} \otimes \mathbf{x}_{t-\ell} \mathbf{x}'_{t-\ell}).\end{aligned}$$

This distributional result does not require that the true process is a VAR. Instead, the coefficients are defined as those which produce the best (mean square) approximation, and the only requirements on the true process are general dependence conditions. The distributional result shows that the coefficient estimators are asymptotically normal, with a covariance matrix which takes a “long-run” sandwich form.

## 15.11 Covariance Matrix Estimation

The classic homoskedastic estimator of the covariance matrix for  $\hat{\mathbf{a}}$  equals

$$\hat{V}_{\hat{\mathbf{a}}}^0 = \hat{\Sigma} \otimes (\mathbf{X}' \mathbf{X})^{-1}. \quad (15.9)$$

Estimators adjusted for degree-of-freedom can also be used, though there is no established finite-sample justification. This variance estimator is appropriate under the assumption that the conditional mean is correctly specified as a VAR(p), and the innovations are conditionally homoskedastic.

The heteroskedasticity-robust estimator equals

$$\hat{V}_{\hat{\mathbf{a}}} = \left( I_n \otimes (\mathbf{X}' \mathbf{X})^{-1} \right) \left( \sum_{t=1}^n (\hat{\mathbf{e}}_t \hat{\mathbf{e}}'_t \otimes \mathbf{x}_t \mathbf{x}'_t) \right) \left( I_n \otimes (\mathbf{X}' \mathbf{X})^{-1} \right). \quad (15.10)$$

This variance estimator is appropriate under the assumption that the conditional mean is correctly specified as a VAR(p), but does not require that the innovations are conditionally homoskedastic.

The Newey-West estimator equals

$$\widehat{V}_{\widehat{\alpha}} = \left( \mathbf{I}_n \otimes (\mathbf{X}' \mathbf{X})^{-1} \right) \widehat{\Omega}_M \left( \mathbf{I}_n \otimes (\mathbf{X}' \mathbf{X})^{-1} \right). \quad (15.11)$$

$$\begin{aligned} \widehat{\Omega}_M &= \sum_{\ell=-M}^M w_\ell \sum_{1 \leq t-\ell \leq n} \mathbf{x}_{t-\ell} \widehat{\epsilon}_{t-\ell} \mathbf{x}_{t-\ell}' \widehat{\epsilon}_t \\ w_\ell &= 1 - \frac{|\ell|}{M+1}. \end{aligned}$$

The number  $M$  is called the lag truncation number. An unweighted version sets  $w_\ell = 1$ . The Newey-West estimator does not require that the VAR(p) is correctly specified.

Traditional textbooks have only used the homoskedastic variance estimation formula (15.9) and consequently existing software follows the same convention. For example, the `var` command in Stata displays only homoskedastic standard errors. Some researchers use the heteroskedasticity-robust estimator (15.10). The Newey-West estimator (15.11) is not commonly used for VAR models.

Asymptotic approximations tend to be much less accurate under time series dependence than for independent observations. Therefore bootstrap methods are popular. In Section 14.45 we described several bootstrap methods for time series observations. While Section 14.45 focused on univariate time series, the extension to multivariate observations is straightforward.

## 15.12 Selection of Lag Length in an VAR

For a data-dependent rule to pick the lag length  $p$  in a VAR it is recommended to minimize an information criterion. The formula for the AIC and BIC are

$$\begin{aligned} \text{AIC}(p) &= n \log \det \widehat{\Omega}(p) + 2K(p) \\ \text{BIC}(p) &= n \log \det \widehat{\Omega}(p) + \log(n)K(p) \\ \widehat{\Sigma}(p) &= \frac{1}{n} \sum_{t=1}^n \widehat{\epsilon}_t(p) \widehat{\epsilon}_t(p)' \\ K(p) &= m(pm + 1) \end{aligned}$$

where  $K(p)$  is the number of parameters in the model, and  $\widehat{\epsilon}_t(p)$  is the OLS residual vector from the model with  $p$  lags. The log determinant is the criterion from the multivariate normal likelihood.

In Stata, the AIC for a set of estimated VAR models can be compared using the `varsoc` command. It should be noted, however, that the Stata routine actually displays  $\text{AIC}(p)/n = \log \det \widehat{\Omega}(p) + 2K(p)/n$ . This does not affect the ranking of the models, but makes the differences between models appear smaller than they actually are.

## 15.13 Illustration

We estimate a three-variable system which is a simplified version of a model often used to study the impact of monetary policy. The three variables are quarterly from FRED-QD: real GDP growth rate ( $100\Delta \log(GDP_t)$ ), GDP inflation rate ( $100\Delta \log(P_t)$ ), and the Federal funds interest rate. VARs from lags 1 through 8 were estimated by least squares. The model with the smallest AIC was the VAR(6). The coefficient estimates and (homoskedastic) standard errors for the VAR(6) are reported in Table 15.1.

Examining the coefficients in the table, we can see that GDP displays a moderate degree of serial correlation, and shows a large response to the federal funds rate, especially at lags 2 and 3. Inflation also displays serial correlation, shows minimal response to GDP, and also has meaningful response to the federal funds rate. The federal funds rate has the strongest serial correlation. Overall, it is difficult to read too much meaning into the coefficient estimates, due to the complexity of the interactions. Because of this difficulty, it is typical to focus on other representations of the coefficient estimates such as impulse responses, which we discuss in the upcoming sections.

Table 15.1: Vector Autoregression

	GDP	INF	FF
$GDP_{t-1}$	0.25 (0.07)	0.01 (0.02)	0.08 (0.02)
$GDP_{t-2}$	0.23 (0.07)	-0.02 (0.02)	0.04 (0.02)
$GDP_{t-3}$	0.00 (0.07)	0.03 (0.02)	0.01 (0.02)
$GDP_{t-4}$	0.14 (0.07)	0.04 (0.02)	-0.02 (0.02)
$GDP_{t-5}$	-0.02 (0.07)	-0.03 (0.02)	0.04 (0.02)
$GDP_{t-6}$	0.05 (0.06)	-0.00 (0.02)	-0.01 (0.02)
$INF_{t-1}$	0.11 (0.20)	0.57 (0.07)	0.01 (0.05)
$INF_{t-2}$	-0.17 (0.23)	0.10 (0.08)	0.17 (0.06)
$INF_{t-3}$	0.01 (0.23)	0.09 (0.08)	-0.05 (0.06)
$INF_{t-4}$	0.16 (0.23)	0.14 (0.08)	-0.05 (0.06)
$INF_{t-5}$	0.12 (0.24)	-0.05 (0.08)	-0.05 (0.06)
$INF_{t-6}$	-0.14 (0.21)	0.10 (0.07)	0.09 (0.05)
$FF_{t-1}$	0.13 (0.26)	0.28 (0.08)	1.14 (0.07)
$FF_{t-2}$	-1.50 (0.38)	-0.27 (0.12)	-0.53 (0.10)
$FF_{t-3}$	1.40 (0.40)	0.12 (0.13)	0.53 (0.10)
$FF_{t-4}$	-0.57 (0.41)	-0.13 (0.13)	-0.28 (0.11)
$FF_{t-5}$	0.01 (0.40)	0.25 (0.13)	0.28 (0.10)
$FF_{t-6}$	0.47 (0.26)	-0.27 (0.08)	-0.24 (0.07)
Intercept	1.15 (0.54)	0.22 (0.18)	-0.33 (0.14)

## 15.14 Predictive Regressions

In some contexts (including prediction) it is useful to consider models where the dependent variable is dated multiple periods ahead of the right-hand-side variables. These equations can be single equation or multivariate; we can consider both as special cases of a VAR (as a single equation model can be written as one equation taken from a VAR system). An  $h$ -step predictive VAR(p) takes the form

$$\mathbf{y}_{t+h} = \mathbf{b}_0 + \mathbf{B}_1 \mathbf{y}_t + \cdots + \mathbf{B}_p \mathbf{y}_{t-p+1} + \mathbf{u}_t. \quad (15.12)$$

The integer  $h \geq 1$  is the **horizon**. A one-step predictive VAR equals a standard VAR. The coefficients should be viewed as the best linear predictors of  $\mathbf{y}_{t+h}$  given  $(\mathbf{y}_t, \dots, \mathbf{y}_{t-p+1})$ .

There is an interesting relationship between a VAR model and the corresponding  $h$ -step predictive VAR model.

**Theorem 15.11** If  $\mathbf{y}_t$  is a VAR(p) process, then its  $h$ -step predictive regression is a predictive VAR(p) with  $\mathbf{u}_t$  a  $MA(h-1)$  process and  $\mathbf{B}_1 = \Theta_h = IRF(h)$ .

The proof of Theorem 15.11 is presented in Section 15.31.

There are several implications of this theorem. First, if  $\mathbf{y}_t$  is a VAR(p) process then the correct number of lags for an  $h$ -step predictive regression is also  $p$  lags. Second, the error in a predictive regression is a MA process, and is thus serially correlated. The linear dependence, however, is capped by the horizon. Third, the leading coefficient matrix corresponds to the  $h^{th}$  moving average coefficient matrix, which also equals the  $h^{th}$  impulse response matrix.

The predictive regression (15.12) can be estimated by least-squares. We can write the estimates as

$$\mathbf{y}_{t+h} = \hat{\mathbf{b}}_0 + \hat{\mathbf{B}}_1 \mathbf{y}_t + \cdots + \hat{\mathbf{B}}_p \mathbf{y}_{t-p+1} + \hat{\mathbf{u}}_t. \quad (15.13)$$

For a distribution theory we need to apply Theorem 15.10 since the innovations  $\mathbf{u}_t$  are a moving average and thus clearly violate the MDS assumption. It follows as well that the covariance matrix for the estimators should be estimated by the Newey-West (15.11) estimator. There is a difference, however. Since  $\mathbf{u}_t$  is known to be a  $MA(h-1)$  a reasonable choice is to set  $M = h - 1$  and use the simple weights  $w_\ell = 1$ . Indeed, this was the original suggestion by Hansen and Hodrick (1980).

For a distributional theory we can apply Theorem 15.10. Let  $\mathbf{b}$  be the vector of coefficients in (15.12) and  $\hat{\mathbf{b}}$  the corresponding least squares estimator. Let  $\mathbf{x}_t$  be the vector of regressors in (15.12).

**Theorem 15.12** If  $\mathbf{y}_t$  is strictly stationary, ergodic,  $\Sigma > 0$ , and for some  $r > 4$ ,  $\mathbb{E} \|\mathbf{y}_t\|^r < \infty$  and the mixing coefficients satisfy  $\sum_{\ell=1}^{\infty} \alpha(\ell)^{1-4/r} < \infty$ , then as  $n \rightarrow \infty$ ,

$$\sqrt{n} (\hat{\mathbf{b}} - \mathbf{b}) \xrightarrow{d} N(\mathbf{0}, \mathbf{V})$$

where

$$\begin{aligned} \mathbf{V} &= \bar{\mathbf{Q}}^{-1} \boldsymbol{\Omega} \bar{\mathbf{Q}}^{-1} \\ \bar{\mathbf{Q}} &= \mathbf{I}_m \otimes \mathbf{Q} \\ \mathbf{Q} &= \mathbb{E}(\mathbf{x}_t \mathbf{x}'_t) \\ \boldsymbol{\Omega} &= \sum_{\ell=-\infty}^{\infty} \mathbb{E}(\mathbf{u}_{t-\ell} \mathbf{u}'_{t-\ell} \otimes \mathbf{x}_{t-\ell} \mathbf{x}'_{t-\ell}). \end{aligned}$$

## 15.15 Impulse Response Estimation

Reporting of impulse response estimates is one of the most common applications of vector autoregressive modeling. There are several methods to estimate the impulse response function. In this section we review the most common estimator based on the estimated VAR parameters.

Within a VAR(p) model, the impulse responses are determined by the VAR coefficients. We can write this mapping as  $\Theta_h = g_h(\mathbf{A})$ . The plug-in approach suggests the estimator  $\widehat{\Theta}_h = g_h(\widehat{\mathbf{A}})$  given the VAR(p) coefficient estimator  $\widehat{\mathbf{A}}$ . These are the impulse responses implied by the estimated VAR coefficients. While it is possible to explicitly write the function  $g_h(\mathbf{A})$ , a computationally simple approach is to use Theorem 15.3, which shows that the impulse response matrices can be written as a simple recursion in the VAR coefficients. Thus the impulse response estimator satisfies the recursion

$$\widehat{\Theta}_h = \sum_{\ell=1}^{\min[h,p]} \widehat{\mathbf{A}}_\ell \widehat{\Theta}_{h-\ell}.$$

We then set  $\widehat{\text{IRF}}(h) = \widehat{\Theta}_h$ .

This is the the most commonly used method for impulse response estimation, and it is the method implemented in standard packages.

Since  $\widehat{\mathbf{A}}$  is random, so is  $\widehat{\text{IRF}}(h)$  as it is a nonlinear function of  $\widehat{\mathbf{A}}$ . Using the delta method, we deduce that the elements of  $\widehat{\text{IRF}}(h)$  (the impulse responses) are asymptotically normally distributed. With some messy algebra explicit expressions for the asymptotic variances can be obtained. Sample versions can be used to calculate asymptotic standard errors. These can be used to form asymptotic confidence intervals for the impulse responses.

The asymptotic approximations, however, can be quite poor. As we discussed earlier, the asymptotic approximations for the distribution of the coefficients  $\widehat{\mathbf{A}}$  can be quite poor due to the serial dependence in the observations. The asymptotic approximations for  $\widehat{\text{IRF}}(h)$  can be significantly worse, because the impulse responses are highly nonlinear functions of the coefficients. For example, in the simple AR(1) model with coefficient estimate  $\widehat{\alpha}$ , the  $h^{th}$  impulse response is  $\widehat{\alpha}^h$  which is highly nonlinear for even moderate horizons  $h$ .

Consequently, asymptotic approximations are less popular than bootstrap approximations. The most popular bootstrap approximation uses the recursive bootstrap (see Section 14.45) using the fitted VAR model, and then calculates confidence intervals for the impulse responses with the percentile method. An unfortunate feature of this choice is that the percentile bootstrap confidence interval is quite biased, since the nonlinear impulse response estimates are highly biased and the percentile bootstrap accentuates bias.

Some advantages of the estimation method as described is that it produces impulse response estimates which are directly related to the estimated VAR(p) model, and are internally consistent with one another. The method is also numerically stable. It is efficient when the true process is a true VAR(p) with conditionally homoskedastic MDS innovations. When the true process is not a VAR(p) it can be thought of as a non-parametric estimator of the impulse response if  $p$  is large (or selected appropriately in a data-dependent fashion, such as by the AIC).

A disadvantage of this estimator is that it is a highly non-linear function of the VAR coefficient estimators. Therefore the distribution of the impulse response estimator is unlikely to be well approximated by the normal distribution. When the VAR(p) is not the true process then it is possible that the nonlinear transformation accentuates the misspecification bias.

Impulse response functions can be calculated and displayed in Stata using the `irf` command. The command `irf create` is used to calculate impulse response functions and confidence intervals. The default confidence intervals are asymptotic (delta method). Bootstrap (recursive method) standard errors can be substituted using the `bs` option. The command `irf graph irf` produces graphs of the impulse response function along with 95% asymptotic confidence intervals. The command `irf graph cirf` produces the cumulative impulse response function. It may be useful to know that the impulse response estimates are unscaled, so represent the response due to a one-unit change in the impulse

variable. A limitation of the Stata `irf` command is that there are limited options for standard error and confidence interval construction. The asymptotic standard errors are calculated using the homoskedastic formula, not the correct heteroskedastic formula. The bootstrap confidence intervals are calculated using the normal approximation bootstrap confidence interval, the least reliable bootstrap confidence interval method. Better options such as the bias-corrected percentile confidence interval are not provided as options.

## 15.16 Local Projection Estimator

Jordà (2005) observed that the impulse response can be estimated by a least squares predictive regression. The key is Theorem 15.11, which established that  $\Theta_h = \mathbf{B}_1$ , the leading coefficient matrix in the  $h$ -step predictive regression.

The method is as follows. For each horizon  $h$  estimate a predictive regression (15.12) to obtain the leading coefficient matrix estimator  $\widehat{\mathbf{B}}_1$ . The estimator is  $\widehat{\text{IRF}}(h) = \widehat{\mathbf{B}}_1$ , and is known as the **local projection** estimator.

Theorem 15.12 shows that the local projection impulse response estimator is asymptotically normal. Newey-West methods must be used for calculation of asymptotic standard errors since the regression errors are serially correlated.

Jordà (2005) speculates that the local projection estimator will be less sensitive to misspecification since it is a straightforward linear estimator. This is intuitive but unclear. Theorem 15.11 relies on the assumption that  $\mathbf{y}_t$  is a VAR( $p$ ) process, and fails otherwise. Thus if the true process is not a VAR( $p$ ) then the coefficient matrix  $\mathbf{B}_1$  in (15.12) does not correspond to the desired impulse response matrix  $\Theta_h$ , and hence will be misspecified. The accuracy (in the sense of low bias) of both the conventional and the local projection estimator relies on  $p$  being sufficiently large that the VAR( $p$ ) model is a good approximation to the true infinite-order regression (15.3). Without a formal theory it is difficult to know which estimator is more robust than the other.

One implementation challenge is the choice of  $p$ . While the method allows for  $p$  to vary across horizon  $h$ , there is no well-established method for selection of the VAR order for predictive regressions. (Standard selection criteria such as AIC are inappropriate under serially correlated errors, just as conventional standard errors are inappropriate.) Therefore the seemingly natural choice is to use the same  $p$  for all horizons, and base this choice on the one-step VAR model where AIC can be used for model selection.

An advantage of the local projection method is that it is a direct estimator of the impulse response, and thus possibly more robust than the conventional method. It is a linear estimator and thus likely to have a better-behaved asymptotic distribution.

A disadvantage is that the method relies on a regression (15.12) that has serially correlated errors. The latter are highly correlated at long horizons and this renders the estimator imprecise. Local projection estimators tend to be less smooth and more erratic than those produced by the conventional estimator, reflecting a possible lack of precision.

## 15.17 Regression on Residuals

If the innovations  $\mathbf{e}_t$  were observed it would be natural to directly estimate the coefficients of the multivariate Wold decomposition. We would pick a maximum horizon  $h$  and then estimate the equation

$$\mathbf{y}_t = \boldsymbol{\mu} + \Theta_1 \mathbf{e}_{t-1} + \Theta_2 \mathbf{e}_{t-2} + \cdots + \Theta_h \mathbf{e}_{t-h} + \mathbf{u}_t$$

where

$$\mathbf{u}_t = \mathbf{e}_t + \sum_{\ell=h+1}^{\infty} \Theta_{\ell} \mathbf{e}_{t-\ell}.$$

The variables  $(\mathbf{e}_{t-1}, \dots, \mathbf{e}_{t-h})$  are uncorrelated with  $\mathbf{u}_t$  so the least-squares estimator of the coefficients is consistent and asymptotically normal. Since  $\mathbf{u}_t$  is serially correlated the Newey-West method should be used to calculate standard errors.

In practice the innovations  $\mathbf{e}_t$  are not observed. If they are replaced by the residuals  $\widehat{\mathbf{e}}_t$  from an estimated VAR(p) then we can estimate the coefficients by least squares applied to the equation

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Theta}_1 \widehat{\mathbf{e}}_{t-1} + \boldsymbol{\Theta}_2 \widehat{\mathbf{e}}_{t-2} + \cdots + \boldsymbol{\Theta}_h \widehat{\mathbf{e}}_{t-h} + \widehat{\mathbf{u}}_t.$$

This idea originated with Durbin (1960).

This is a two-step estimator with generated regressors. (See Section 12.26.) The impulse response estimators are consistent and asymptotically normal, but with a covariance matrix which is complicated due to the two-step estimation. Conventional, robust and Newey-West standard errors do not account for this without modification.

Chang and Sakata (2007) proposed a simplified version of the Durbin regression. Notice that for any horizon  $h$  we can rewrite the Wold decomposition as

$$\mathbf{y}_{t+h} = \boldsymbol{\mu} + \boldsymbol{\Theta}_h \mathbf{e}_t + \mathbf{v}_{t+h}$$

where

$$\mathbf{v}_t = \sum_{\ell=0}^{h-1} \boldsymbol{\Theta}_\ell \mathbf{e}_{t-\ell} + \sum_{\ell=h+1}^{\infty} \boldsymbol{\Theta}_\ell \mathbf{e}_{t-\ell}.$$

The regressor  $\mathbf{e}_t$  is uncorrelated with  $\mathbf{v}_{t+h}$ . Thus  $\boldsymbol{\Theta}_h$  can be estimated by a regression of  $\mathbf{y}_{t+h}$  on  $\mathbf{e}_t$ . In practice we can replace  $\mathbf{e}_t$  by the least-squares residual  $\widehat{\mathbf{e}}_t$  from an estimated VAR(p) to estimate the regression

$$\mathbf{y}_{t+h} = \boldsymbol{\mu} + \boldsymbol{\Theta}_h \widehat{\mathbf{e}}_t + \widehat{\mathbf{v}}_{t+h}. \quad (15.14)$$

Similar to the Durbin regression the Chang-Sakata estimator is a two-step estimator with a generated regressor. However, as it takes the form studied in Section 12.27, it can be shown that the Chang-Sakata two-step estimator has the same asymptotic distribution as the idealized one-step estimator as if  $\mathbf{e}_t$  were observed. Thus the standard errors do not need to be adjusted for generated regressors. (The Newey-West method should be used to account for the serial correlation.) This feature is an advantage of their estimator over the Durbin estimator. However, the variance of the error  $\mathbf{v}_{t+h}$  in the Chang-Sakata regression is larger than the variance of the error  $\mathbf{u}_t$  in the Durbin regression, so the Chang-Sakata estimator may be less precise than the Durbin estimator.

Chang and Sakata (2007) also point out the following interesting connection. The least-squares slope estimator in (15.14) is algebraically identical<sup>1</sup> to the slope estimator  $\widehat{\mathbf{B}}_1$  in a predictive regression with  $p-1$  lags. This holds by the FWL Theorem. Thus the Chang-Sakata estimator is similar to a local projection estimator.

## 15.18 Orthogonalized Shocks

We can use the impulse response function to examine how the innovations impact the time-paths of the variables. A difficulty in interpretation, however, is that the elements of the innovation vector  $\mathbf{e}_t$  are contemporaneously correlated. Thus  $e_{jt}$  and  $e_{it}$  are (in general) not independent, so consequently it does not make sense to treat  $e_{jt}$  and  $e_{it}$  as fundamental “shocks”. Another way of describing the problem is that it does not make sense, for example, to describe the impact of  $e_{jt}$  while “holding”  $e_{it}$  constant.

The natural solution is to orthogonalize the innovations so that they are uncorrelated, and then view the orthogonalized errors as the fundamental “shocks”. Recall that  $\mathbf{e}_t$  is mean zero with variance matrix  $\boldsymbol{\Sigma}$ . We can factor  $\boldsymbol{\Sigma}$  into the product of an  $m \times m$  matrix  $\mathbf{B}$  with its transpose

$$\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}'.$$

The matrix  $\mathbf{B}$  is called a “square root” of  $\boldsymbol{\Sigma}$ . (See Section A.13.) Define  $\boldsymbol{\epsilon}_t = \mathbf{B}^{-1} \mathbf{e}_t$ . The random vector  $\boldsymbol{\epsilon}_t$  has mean zero and variance matrix  $\mathbf{B}^{-1} \boldsymbol{\Sigma} \mathbf{B}^{-1'} = \mathbf{B}^{-1} \mathbf{B} \mathbf{B}' \mathbf{B}^{-1'} = \mathbf{I}_m$ . Thus the elements  $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \dots, \epsilon_{mt})$  are mutually uncorrelated. We can write the innovations as a function of the orthogonalized errors as

$$\mathbf{e}_t = \mathbf{B}\boldsymbol{\epsilon}_t. \quad (15.15)$$

---

<sup>1</sup>Technically, if the sample lengths are adjusted.

To distinguish  $\boldsymbol{\varepsilon}_t$  from  $\boldsymbol{e}_t$  we will typically call  $\boldsymbol{\varepsilon}_t$  the “orthogonalized shocks” or more simply as the “shocks” and continue to call  $\boldsymbol{e}_t$  the “innovations”.

When  $m > 1$  there is not a unique square root matrix  $\mathbf{B}$  so there is not a unique orthogonalization. The most common choice (and was originally advocated by Sims (1980)) is to use the Cholesky decomposition (see Section A.16). This sets  $\mathbf{B}$  to be **lower triangular**, meaning that it takes the form

$$\mathbf{B} = \begin{bmatrix} b_{11} & 0 & 0 \\ b_{21} & b_{22} & 0 \\ b_{31} & b_{32} & b_{33} \end{bmatrix}$$

with non-negative diagonal elements. We can write the Cholesky decomposition of the matrix  $\mathbf{A}$  as  $\mathbf{C} = \text{chol}(\mathbf{A})$  which means that  $\mathbf{A} = \mathbf{CC}'$  with  $\mathbf{C}$  lower triangular. We thus set

$$\mathbf{B} = \text{chol}(\boldsymbol{\Sigma}). \quad (15.16)$$

Equivalently, the innovations are related to the orthogonalized shocks by the equations

$$\begin{aligned} e_{1t} &= b_{11}\varepsilon_{1t} \\ e_{2t} &= b_{21}\varepsilon_{1t} + b_{22}\varepsilon_{2t} \\ e_{3t} &= b_{31}\varepsilon_{1t} + b_{32}\varepsilon_{2t} + b_{33}\varepsilon_{3t}. \end{aligned}$$

This structure is **recursive**. The innovation  $e_{1t}$  is a function only of the single shock  $\varepsilon_{1t}$ . The innovation  $e_{2t}$  is a function of the shocks  $\varepsilon_{1t}$  and  $\varepsilon_{2t}$ , and the innovation  $e_{3t}$  is a function of all three shocks. This means that within a single time period  $e_{1t}$  receives no feedback from the other variables, and  $e_{2t}$  is only affected by  $e_{1t}$  (but not  $e_{3t}$ ).

Another way of looking at the structure is that the first shock  $\varepsilon_{1t}$  affects all three innovation, the second shock  $\varepsilon_{2t}$  affects  $e_{2t}$  and  $e_{3t}$ , and the third shock  $\varepsilon_{3t}$  only affects  $e_{3t}$ .

A recursive structure is an exclusion restriction. The recursive structure excludes the possibility of  $\varepsilon_{2t}$  or  $\varepsilon_{3t}$  contemporaneously affecting  $e_{1t}$ , and excludes the possibility of  $\varepsilon_{3t}$  contemporaneously affecting  $e_{2t}$ .

When using the Cholesky decomposition the recursive structure is determined by the ordering of the variables in the system. The order matters, and is the key identifying assumption. We will return to this issue later.

Finally, we mention that the system (15.15) is equivalent to the system

$$\mathbf{A}\boldsymbol{e}_t = \boldsymbol{\varepsilon}_t \quad (15.17)$$

where  $\mathbf{A} = \mathbf{B}^{-1}$  is lower triangular when  $\mathbf{B}$  is lower triangular. The representation (15.15) is more convenient, however, for most of our purposes.

## 15.19 Orthogonalized Impulse Response Function

We have defined the impulse response function as the change in the time  $t$  projection of the variables  $\mathbf{y}_t$  due to the innovation  $\boldsymbol{e}_t$ . As we discussed in the previous section, since the innovations are contemporaneously correlated it makes better sense to focus on changes due to the orthogonalized shocks  $\boldsymbol{\varepsilon}_t$ . Consequently we define the **orthogonalized impulse response function (OIRF)** as

$$\text{OIRF}(h) = \frac{\partial}{\partial \boldsymbol{\varepsilon}_t'} \mathcal{P}_t(\mathbf{y}_{t+h}).$$

We can write the multivariate Wold representation as

$$\mathbf{y}_t = \boldsymbol{\mu} + \sum_{\ell=0}^{\infty} \boldsymbol{\Theta}_{\ell} \boldsymbol{e}_{t-\ell} = \boldsymbol{\mu} + \sum_{\ell=0}^{\infty} \boldsymbol{\Theta}_{\ell} \mathbf{B} \boldsymbol{\varepsilon}_{t-\ell}$$

where  $\mathbf{B}$  is from (15.16). We deduce that

$$\text{OIRF}(h) = \Theta_h \mathbf{B} = \text{IRF}(h) \mathbf{B}.$$

This is the non-orthogonalized impulse response matrix multiplied by the matrix square root  $\mathbf{B}$ .

Write the rows of the matrix  $\Theta_h$  as

$$\Theta_h = \begin{bmatrix} \theta'_{1h} \\ \theta'_{mh} \end{bmatrix}$$

and the columns of the matrix  $\mathbf{B}$  as

$$\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_m].$$

We can see that

$$\text{OIRF}_{ij}(h) = [\Theta_h \mathbf{B}]_{ij} = \theta'_{ih} \mathbf{b}_j.$$

There are  $m^2$  such responses for each horizon  $h$ .

The **cumulative orthogonalized impulse response function (COIRF)** is

$$\text{COIRF}(h) = \sum_{\ell=1}^h \text{OIRF}(\ell) = \sum_{\ell=1}^h \Theta_\ell \mathbf{B}.$$

## 15.20 Orthogonalized Impulse Response Estimation

We have already discussed estimation of the moving average matrices  $\Theta_\ell$ . We need an estimator of  $\mathbf{B}$ .

We first estimate the VAR(p) model by least squares. This gives us the coefficient matrices  $\hat{\mathbf{A}}$  and the error variance matrix  $\hat{\Sigma}$ . From the latter we apply the Cholesky decomposition  $\hat{\mathbf{B}} = \text{chol}(\hat{\Sigma})$  so that  $\hat{\Sigma} = \hat{\mathbf{B}} \hat{\mathbf{B}}'$ . (See Section A.16 for the algorithm.) The orthogonalized impulse response estimators are then

$$\widehat{\text{OIRF}}(h) = \hat{\Theta}_h \hat{\mathbf{B}} = \hat{\theta}'_{ih} \hat{\mathbf{b}}_j.$$

The estimator  $\widehat{\text{OIRF}}(h)$  is a complicated nonlinear function of  $\hat{\mathbf{A}}$  and  $\hat{\Sigma}$ . They are asymptotically normally distributed by the delta method. This allows for explicit calculation of asymptotic standard errors. These can be used to form asymptotic confidence intervals for the impulse responses.

As discussed earlier, the asymptotic approximations can be quite poor. Consequently bootstrap approximations are more widely used than asymptotic methods.

Orthogonalized impulse response functions can be displayed in Stata using the `irf` command. The command `irf graph oirf` produces graphs of the orthogonalized impulse response function along with 95% asymptotic confidence intervals. The command `irf graph coirf` produces the cumulative orthogonalized impulse response function. It may also be useful to know that the OIRF are scaled for a one-standard deviation shock, so the impulse response represents the response due to a one-standard-deviation change in the impulse variable. As discussed earlier, a limitation of the Stata `irf` command is that there are limited options for standard error and confidence interval construction. The asymptotic standard errors are calculated using the homoskedastic formula, not the correct heteroskedastic formula. The bootstrap confidence intervals are calculated using the normal approximation bootstrap confidence interval.

## 15.21 Illustration

To illustrate we use the three-variable system from Section 15.13. We use the ordering (1) real GDP growth rate, (2) inflation rate, (3) Federal funds interest rate. We discuss the choice later when we discuss identification. We use the estimated VAR(6) and calculate the orthogonalized impulse response functions using the standard VAR estimator.

In Figure 15.1 we display the estimated orthogonalized impulse response of the GDP growth rate in response to a one standard deviation increase in the federal funds rate. The left plot shows the impulse response function and the middle plot the cumulative impulse response function. As we discussed earlier, the interpretation of the impulse response and the cumulative impulse response depends on whether the variable enters the VAR in differences or in levels. In this case, GDP growth is the first difference of the natural logarithm. Thus the left plot (the impulse response function) shows the effect of interest rates on the growth rate of GDP. The middle plot (the cumulative impulse response) shows the effect on the log-level of GDP. The left plot shows that the GDP growth rate is negatively affected in the second quarter after an interest rate increase (a drop of about 0.2%, non-annualized), and the negative effects continue for several quarters following. The middle plot shows the effect on the level of GDP, measured as percentage changes. It shows that an interest rate increase causes GDP to fall for about 8 quarters, reducing GDP by about 0.6%.

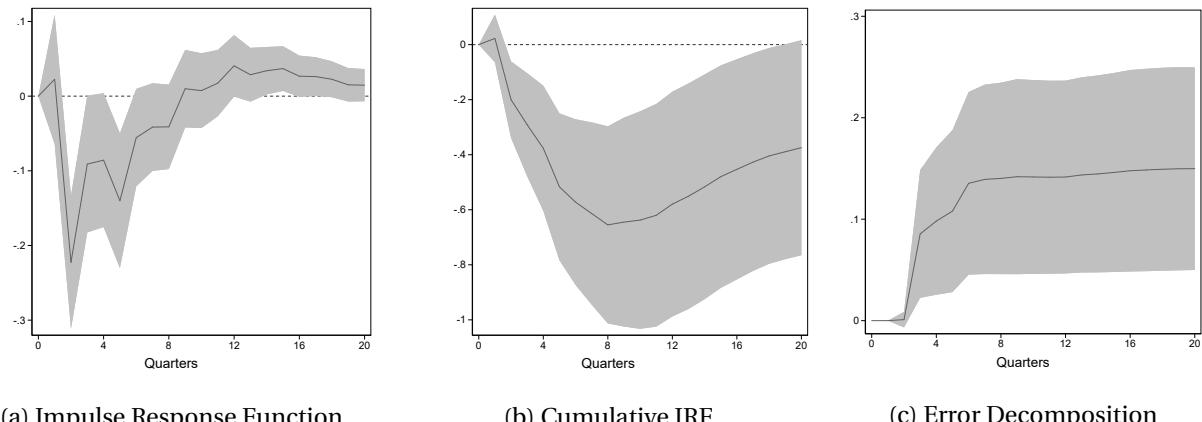


Figure 15.1: Response of GDP Growth to Orthogonalized Fed Funds Shock

## 15.22 Forecast Error Decomposition

An alternative tool to investigate an estimated VAR is the **forecast error decomposition** which decomposes multi-step forecast error variances by the component shocks. The forecast error decomposition indicates which shocks contribute towards the fluctuations of each variable in the system.

It is defined as follows. Take the moving average representation of the  $i^{th}$  variable  $y_{i,t+h}$  written as a function of the orthogonalized shocks

$$y_{i,t+h} = \mu_i + \sum_{\ell=0}^{\infty} \boldsymbol{\theta}_i(\ell)' \mathbf{B} \boldsymbol{\varepsilon}_{t+h-\ell}.$$

The best linear forecast of  $\mathbf{y}_{t+h}$  at time  $t$  is

$$y_{i,t+h|t} = \mu_i + \sum_{\ell=h}^{\infty} \boldsymbol{\theta}_i(\ell)' \mathbf{B} \boldsymbol{\varepsilon}_{t+h-\ell}.$$

Thus the  $h$ -step forecast error is the difference

$$y_{i,t+h} - y_{i,t+h|t} = \sum_{\ell=0}^{h-1} \boldsymbol{\theta}_i(\ell)' \mathbf{B} \boldsymbol{\varepsilon}_{t+h-\ell}.$$

The variance of this forecast error is

$$\begin{aligned}\text{var}(y_{i,t+h} - y_{i,t+h|t}) &= \sum_{\ell=0}^{h-1} \text{var}(\boldsymbol{\theta}_i(\ell)' \mathbf{B} \boldsymbol{\varepsilon}_{t+h-\ell}) \\ &= \sum_{\ell=0}^{h-1} \boldsymbol{\theta}_i(\ell)' \mathbf{B} \mathbf{B}' \boldsymbol{\theta}_i(\ell).\end{aligned}\quad (15.18)$$

To isolate the contribution of the  $j^{th}$  shock, notice that

$$\boldsymbol{\varepsilon}_t = \mathbf{B} \boldsymbol{\varepsilon}_t = \mathbf{b}_1 \varepsilon_{1t} + \cdots + \mathbf{b}_m \varepsilon_{mt}.$$

Thus the contribution of the  $j^{th}$  shock is  $\mathbf{b}_j \varepsilon_{jt}$ . Now imagine replacing  $\mathbf{B} \boldsymbol{\varepsilon}_t$  in the variance calculation by the  $j^{th}$  contribution  $\mathbf{b}_j \varepsilon_{jt}$ . This is

$$\text{var}(y_{i,t+h} - y_{i,t+h|t}) = \sum_{\ell=0}^{h-1} \text{var}(\boldsymbol{\theta}_i(\ell)' \mathbf{b}_j \varepsilon_{jt+h-\ell}) = \sum_{\ell=0}^{h-1} (\boldsymbol{\theta}_i(\ell)' \mathbf{b}_j)^2. \quad (15.19)$$

Examining (15.18) and using  $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_m]$  we can write (15.18) as

$$\text{var}(y_{i,t+h} - y_{i,t+h|t}) = \sum_{j=1}^m \sum_{\ell=0}^{h-1} (\boldsymbol{\theta}_i(\ell)' \mathbf{b}_j)^2. \quad (15.20)$$

The forecast error decomposition is defined as the ratio of the  $j^{th}$  contribution to the total which is the ratio of (15.19) to (15.20):

$$\text{FE}_{ij}(h) = \frac{\sum_{\ell=0}^{h-1} (\boldsymbol{\theta}_i(\ell)' \mathbf{b}_j)^2}{\sum_{j=1}^m \sum_{\ell=0}^{h-1} (\boldsymbol{\theta}_i(\ell)' \mathbf{b}_j)^2}.$$

The  $\text{FE}_{ij}(h)$  lies in  $[0,1]$  and varies across  $h$ . Small values indicate that  $\varepsilon_{jt}$  contributes only a small amount to the variance of  $y_{it}$ . Large values indicate that  $\varepsilon_{jt}$  contributes a major amount of the variance of  $\varepsilon_{it}$ .

A forecast error decomposition requires that orthogonalized innovations. There is no non-orthogonalized version.

The forecast error decomposition can be calculated and displayed in Stata using the `irf` command. The command `irf graph fevd` produces graphs of the forecast error decomposition along with 95% asymptotic confidence intervals.

To illustrate, in Figure 15.1 (right plot) we display the estimated forecast error decomposition of the GDP growth rate due to the federal funds rate. This shows the contribution of movements in the federal funds rate towards fluctuations in GDP growth. The estimated effect is about 15% at long horizons. This is a small but important share of the variance of GDP growth. Combined with the impulse response functions we learn two lessons. That monetary policy (movements in the federal funds rate) can meaningfully affect GDP growth, but monetary policy only accounts for a small component of fluctuations in U.S. GDP.

## 15.23 Identification of Recursive VARs

As we have discussed, a common method to orthogonalize the VAR errors is to use the lower triangular Cholesky decomposition, which implies a recursive structure. The ordering of the variables is critical to a recursive structure. Unless the errors are uncorrelated, different orderings will lead to different impulse response functions and forecast error decompositions. The ordering must be selected by the user; there is no data-dependent choice.

In order for impulse responses and forecast error decompositions to be interpreted causally the orthogonalization must be identified by the user based on a structural economic argument. The choice is

similar to the exclusion restrictions necessary for specification of an instrumental variables regression. By ordering the variables recursively, we are effectively imposing exclusion restrictions. Recall that in our empirical example we used the ordering: (1) real GDP growth rate, (2) inflation rate, (3) Federal funds interest rate. This means that in the equation for GDP we excluded the contemporaneous inflation rate and interest rate, and in the equation for inflation we excluded the contemporaneous interest rate. These are exclusion restrictions. Are they justified?

One approach is to order first the variables which are believed to be contemporaneously affected by the fewest number of shocks. One way of thinking about it is that they are the variables which are “most sticky” within a period. The variables listed last are those which are believed to be contemporaneously affected by the greatest number of shocks. These are the ones which are able to respond within a single period to the shocks, or are most flexible. In our example, we listed output first, prices second and interest rates last. This is consistent with the view that output is effectively pre-determined (within a period) and does not (within a period) respond to price and interest rate movements. Prices are allowed to respond within a period in response to output changes, but not in response to interest rate changes. The latter could be justified if interest rate changes affect investment decisions, but the latter take at least one period to implement. By listing the federal funds rate last, the model allows monetary policy to respond within a period to contemporaneous information about output and prices.

In general, this line of reasoning suggests that production measures should be listed first, goods prices second, and financial prices last. This reasoning is more credible when the time periods are short, and less credible for longer time periods.

Further justifications for possible recursive orderings can include: (1) information delays; (2) implementation delays; (3) institutions; (4) market structure; (5) homogeneity; (6) imposing estimates from other sources. In most cases such arguments can be made, but will be viewed as debatable and restrictive. In any situation it is best to be explicit about the choice and your reasoning for your choice.

Returning to the empirical illustration, it is fairly conventional to order the fed funds rate last. This allows the fed funds rate to respond to contemporaneous information about output and price growth, and identifies the fed funds **policy shock** by the assumption that it does not have a contemporaneous impact on the other variables. It is not clear, however, how to order the other two variables. For simplicity consider a traditional aggregate supply/aggregate demand model of the determination of output and the price level. If the aggregate supply curve is perfectly inelastic in the short run (one quarter), then output is effectively fixed (sticky), so changes in aggregate demand affect prices but not output. Changes in aggregate supply affect both output and prices. Thus we would want to order GDP first and inflation second. This choice would identify the GDP error as the **aggregate supply shock**. This is the ordering choice used in our example.

In contrast, suppose that the aggregate supply curve is perfectly elastic in the short run. Then prices are fixed and output is flexible. Changes in aggregate supply affect both price and output, but changes in aggregate demand only affect output. In this case we would want to order inflation first and GDP second. This choice identifies the inflation error as the aggregate supply shock, the opposite case from the previous assumption!

If the choice between perfectly elastic and perfectly inelastic aggregate supply is not credible then the supply and demand shocks cannot be separately identified based on ordering alone. In this case the full set of impulse responses and error decompositions are not identified. However, a subset may be identified. In general, if the shocks can be ordered in groups, then we can identify any shock for which a group has a single variable. In our example, consider the ordering (1) GDP and inflation; (2) federal funds rate. This means that the model assumes that GDP and inflation do not contemporaneously respond to interest rate movements, but no other restrictions are imposed. In this case the fed funds policy shock is identified. This means that impulse responses of all three variables with respect to the policy shock are identified, and similarly the forecast error composition of the effect of the fed funds shock on each variable is identified. These can be estimated by a VAR using the ordering (GDP, inflation, federal funds rate) as done in our example, or using the ordering (inflation, GDP, federal funds rate). Both choices will lead to the same estimated impulse responses as described. The remaining impulse responses (responses to

GDP and inflation shocks), however, will differ across these two orderings.

## 15.24 Oil Price Shocks

To further illustrate the identification of impulse response functions by recursive structural assumptions we repeat here some of the analysis from Kilian (2009). His paper concerns the identification of the factors affecting crude oil prices, in particular separating supply and demand shocks. The goal is to determine how oil prices respond to economic shocks, and how the responses differ by the type of shock.

To answer this question Kilian uses a three-variable VAR, with monthly measures of global oil production, global economic activity, and the global price of crude oil for 1973m2-2007m12. He uses global variables since the price of crude oil is globally determined. One innovation in the paper is that Kilian develops a new index of global economic activity based on ocean freight rates. His motivation is that shipping rates are directly related to the global demand for industrial commodities. This data set is posted on the text webpage as [Kilian2009](#).

Kilian argues that these three variables are determined by three economic shocks: oil supply, aggregate demand, and oil demand. He suggests that oil supply shocks should be thought of as disruptions in production, processing, or shipping. Aggregate demand is global economic activity. Kilian also argues that oil demand shocks are primarily due to the precautionary demand for oil driven by uncertainty about future oil supply shortfalls.

To identify the shocks, Kilian makes the following exclusion restrictions. First, he assumes that the short-run (one month) supply of crude oil is inelastic with respect to price. Equivalently, oil production takes at least one month to respond to price changes. This restriction is believed to be plausible because of technological factors in crude oil production. It is costly to open new oil fields; and it is nearly impossible to cap an oil well once tapped. Second, Kilian assumes that in the short-run (one month) global real economic activity does not respond to changes in oil prices (due to shocks specific to the oil market), while economic activity is allowed to respond to oil production shocks. This assumption is viewed by Kilian as plausible due to the sluggishness in the response of economic activity to price changes. Crude oil prices, however, are allowed to respond simultaneously to all three shocks.

Kilian's identification strategy is similar to that described in the previous section for the simple aggregate demand/aggregate supply model. The separation of supply and demand shocks is achieved by exclusion restrictions which imply short-run inelasticities. The plausibility of these assumptions rests in part on the monthly frequency of the data. While it is plausible that oil production and economic activity may not respond within one month to price shocks, it is much less plausible that there is no response for a full quarter. The least convincing identifying assumption (in my opinion) is the assumption that economic activity does not respond simultaneously to oil price changes. While much economic activity is pre-planned and hence sluggish to respond, other economic activity (recreational driving, for example) can immediately respond to price changes.

Kilian estimates the three-variable VAR using 24 lags, and calculates the orthogonalized impulse response functions using the ordering implied by these assumptions. He does not discuss the choice of 24 lags, but presumably this is intended to allow for flexible dynamic responses. If the AIC is used for model selection, three lags would be selected. For the analysis reported here, I used 4 lags. The results are qualitatively similar to those obtained using 24 lags. For ease of interpretation, oil supply is entered negatively (multiplied by  $-1$ ) so that all three shocks are scaled to increase oil prices. The impulse response functions for the price of crude oil are displayed in Figure 15.2 for 1-24 months. Panel (a) displays the response of crude oil prices due to an oil supply shock, panel (b) displays the response due to an aggregate demand shock, and panel (c) displays the response due to an oil-demand shock. Notice that all three figures have been displayed using the same y-axis scalings so that the figures are comparable.

What is noticeable about the figures is how differently crude oil prices respond to the three types of shocks. Panel (a) shows that oil prices are only minimally affected by oil production shocks. There is an estimated small short term increase in oil prices, but it is not statistically significant and it reverses within one year. Panel (b) shows that oil prices are significantly affected by aggregate demand shocks, and the

effect cumulatively increases over two years. This is not surprising. Economic activity relies on crude oil, and economic activity is serially correlated. Panel (c) shows that oil prices are strongly immediately affected by oil demand shocks, but the effect attenuates over time. This is a reverse pattern than that found for aggregate demand shocks.

The Kilian (2009) paper is an excellent example of how recursive orderings can be used to identify an orthogonalized VAR through a careful discussion of the causal system and the use of monthly observations.

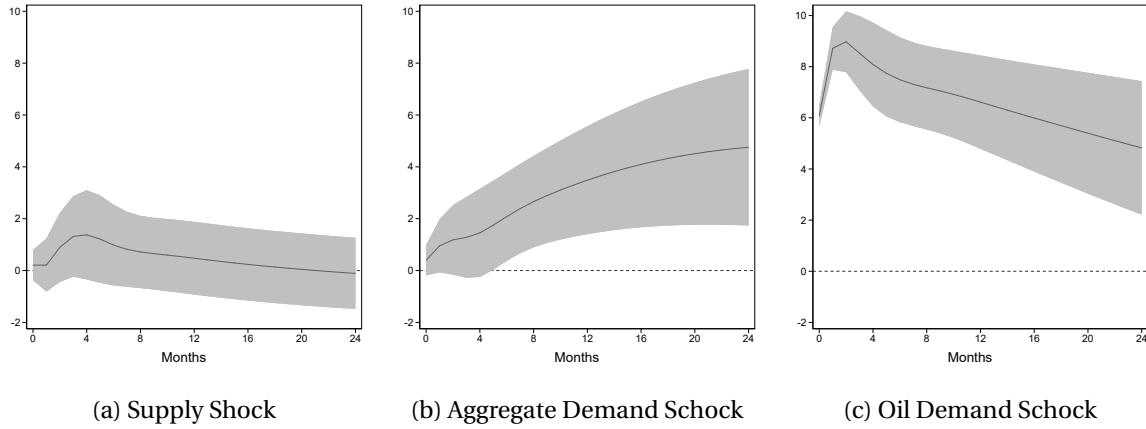


Figure 15.2: Response of Oil Prices to Orthogonalized Shocks

## 15.25 Structural VARs

Recursive models do not allow for simultaneity between the elements of  $\epsilon_t$  and thus the variables  $y_t$  cannot be contemporaneously endogenous. This is highly restrictive, and may not credibly describe many economic systems. There is a general preference in the economics community for **structural vector autoregressive models** (SVARs) which use alternative identification restrictions which do not rely exclusively on recursiveness. Two popular categories of structural VAR models are those based on short-run (contemporaneous) restrictions and those based on long-run (cumulative) restrictions. In this section we review SVARs based on short-run restrictions.

When we introduced methods to orthogonalize the VAR errors we pointed out that we can represent the relationship between the errors and shocks using either the equation  $\epsilon_t = B\epsilon_t$  (15.15) or the equation  $A\epsilon_t = \epsilon_t$  (15.17). Equation (15.15) writes the errors as a function of the shocks. Equation (15.17) writes the errors as a simultaneous system. A broader class of models can be captured by the equation system

$$A\epsilon_t = B\epsilon_t \quad (15.21)$$

where (in the  $3 \times 3$  case)

$$A = \begin{bmatrix} 1 & a_{12} & a_{13} \\ a_{21} & 1 & a_{23} \\ a_{31} & a_{32} & 1 \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}. \quad (15.22)$$

(Note: This matrix  $A$  has nothing to do with the regression coefficient matrix  $A$ . I apologize for the double use of  $A$ , but I use the notation (15.21) to be consistent with the notation elsewhere in the literature.)

Written out,

$$\begin{aligned} e_{1t} &= -a_{12}e_{2t} - a_{13}e_{3t} + b_{11}\epsilon_{1t} + b_{12}\epsilon_{2t} + b_{13}\epsilon_{3t} \\ e_{2t} &= -a_{21}e_{1t} - a_{23}e_{3t} + b_{21}\epsilon_{1t} + b_{22}\epsilon_{2t} + b_{23}\epsilon_{3t} \\ e_{3t} &= -a_{31}e_{1t} - a_{32}e_{2t} + b_{31}\epsilon_{1t} + b_{32}\epsilon_{2t} + b_{33}\epsilon_{3t}. \end{aligned}$$

The diagonal elements of the matrix  $\mathbf{A}$  are set to 1 as normalizations. This normalization allows the shocks  $\varepsilon_{it}$  to have unit variance, which is convenient for impulse response calculations.

The system as written is under-identified. In this three-equation example, the matrix  $\Sigma$  provides only six moments, but the above system has 15 free parameters! To achieve identification we need nine restrictions.

In most applications, it is common to start with the restriction that for each common non-diagonal element of  $\mathbf{A}$  and  $\mathbf{B}$  at most one can be non-zero. That is, for any pair  $i \neq j$ , either  $b_{ji} = 0$  or  $a_{ji} = 0$ .

We will illustrate by using a simplified version of the model employed by Blanchard and Perotti (2002), who were interested in decomposing the effects of government spending and taxes on GDP. They proposed a three-variable system consisting of real government spending (net of transfers), real tax revenues (including transfer payments as negative taxes), and real GDP. All variables are measured in logs. They start with the restrictions  $a_{21} = a_{12} = b_{31} = b_{32} = b_{13} = b_{23} = 0$ , or

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & a_{13} \\ 0 & 1 & a_{23} \\ a_{31} & a_{32} & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & 0 \\ b_{21} & b_{22} & 0 \\ 0 & 0 & b_{33} \end{bmatrix}.$$

This is done so that the relationship between the shocks  $\varepsilon_{1t}$  and  $\varepsilon_{2t}$  is treated as reduced-form, but the coefficients in the  $\mathbf{A}$  matrix can be interpreted as contemporaneous elasticities between the variables. For example,  $a_{23}$  is the within-quarter elasticity of tax revenue with respect to GDP,  $a_{31}$  is the within-quarter elasticity of GDP with respect to government spending, etc.

We just described six restrictions, while nine are required for identification. Blanchard and Perotti (2002) made a strong case for two additional restrictions. First, the within-quarter elasticity of government spending with respect to GDP is zero,  $a_{13} = 0$ . This is because government fiscal policy does not (and cannot) respond to news about GDP within the same quarter. Since the authors defined government spending as net of transfer payments there is no “automatic stabilizer” component of spending. Second, the within-quarter elasticity of tax revenue with respect to GDP can be estimated from existing microeconometric studies. The authors survey the available literature and set  $a_{23} = -2.08$ . To fully identify the model we need one final restriction. The authors argue that there is no clear case for any specific restriction, and so impose a recursive  $\mathbf{B}$  matrix (setting  $b_{12} = 0$ ) and experiment with the alternative  $b_{21} = 0$ , finding that the two specifications are near-equivalent since the two shocks are nearly uncorrelated. In summary the estimated model takes the form

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -2.08 \\ a_{31} & a_{32} & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} b_{11} & 0 & 0 \\ b_{21} & b_{22} & 0 \\ 0 & 0 & b_{33} \end{bmatrix}.$$

Blanchard and Perotti (2002) make use of both matrices  $\mathbf{A}$  and  $\mathbf{B}$ . Other authors use either the simpler structure  $\mathbf{A}\mathbf{e}_t = \boldsymbol{\varepsilon}_t$  or  $\mathbf{e}_t = \mathbf{B}\boldsymbol{\varepsilon}_t$ . In general, either of the two simpler structures are simpler to compute and interpret.

Taking the variance of the variables on each side of (15.21) we find

$$\mathbf{A}\Sigma\mathbf{A}' = \mathbf{B}\mathbf{B}'. \quad (15.23)$$

This is a system of quadratic equations in the free parameters. If the model is just identified it can be solved numerically to find the coefficients of  $\mathbf{A}$  and  $\mathbf{B}$  given  $\Sigma$ . Similarly, given the least-squares error covariance matrix  $\widehat{\Sigma}$  we can numerically solve for the coefficients of  $\widehat{\mathbf{A}}$  and  $\widehat{\mathbf{B}}$ .

While most applications use just-identified models, if the model is over-identified (if there are fewer free parameters than estimated components of  $\Sigma$ ) then the coefficients of  $\widehat{\mathbf{A}}$  and  $\widehat{\mathbf{B}}$  can be found using minimum distance. The implementation in Stata uses MLE (which simultaneously estimates the VAR coefficients). The latter is appropriate when the model is correctly specified (including normality) but otherwise an unclear choice.

Given the parameter estimates the **structural impulse response function** is

$$\widehat{\text{SIRF}}(h) = \widehat{\Theta}(h)\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{B}}.$$

The structural forecast error decompositions are calculated as before, with  $\mathbf{b}_j$  replaced by the  $j^{th}$  column of  $\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{B}}$ .

The structural impulse responses are non-linear functions of the VAR coefficient and variance matrix estimators, so by the delta method are asymptotically normal. Thus asymptotic standard errors can be calculated (using numerical derivatives if convenient). As for orthogonalized impulse responses the asymptotic normal approximation is unlikely to be a good approximation so bootstrap methods are an attractive alternative.

Structural VARs should be interpreted similarly as instrumental variable estimators. Their interpretation relies on valid exclusion restrictions, which can only be justified by external information.

We replicate a simplified version of Blanchard-Perotti (2002). We use<sup>2</sup> quarterly variables from FRED-QD for 1959-2017: real GDP (gdpc1), real tax revenue (fgrectx), and real government spending (gcecc1), all in natural logarithms. Using the AIC for lag length selection, we estimate VARs from one to eight lags and select a VAR(5). The model also includes a linear and quadratic function of time<sup>3</sup>. The estimated structural impulse responses of the three variables with respect to the government spending shock are displayed in Figure 15.3, and the impulse responses with respect to the tax revenue shock are displayed in Figure 15.4. The estimated impulse responses are very similar to those reported in Blanchard and Perotti (2002).

In Figure 15.3 we see that the effect of a government spending shock is persistent, increasing government spending about 1% for the four-year horizon. The effect on tax revenue is minimal. The effect on GDP is positive, small (around 0.25%), but persistent.

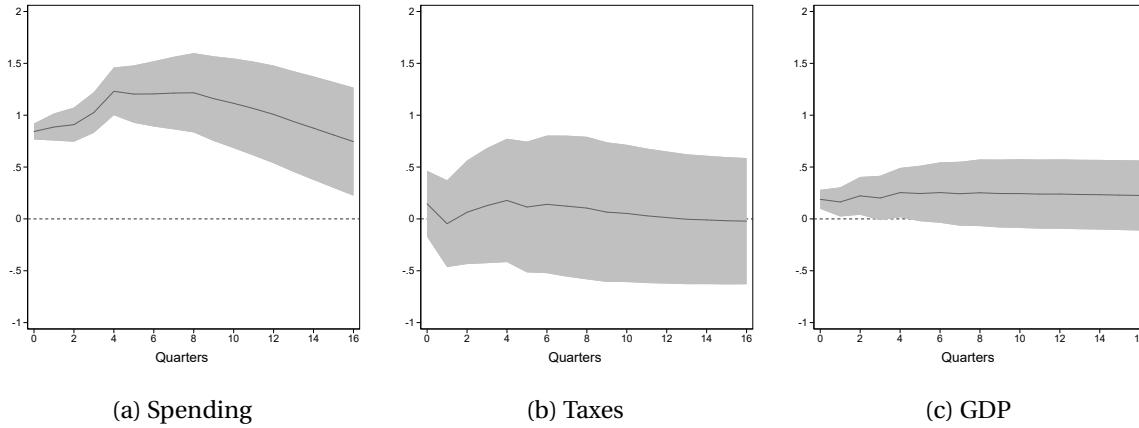


Figure 15.3: Response to a Government Spending Shock

In Figure 15.4 we see that the effect of a tax revenue shock is quite different. The initial effect on tax revenue is high, but diminishes to zero by about two years. The effect on government spending is mildly negative<sup>4</sup>. The effect on GDP is negative and persistent, and more substantial than the effect of a spending shock, reaching about  $-0.5\%$  at six quarters. Together, the impulse response estimates show that changes in government spending and tax revenue have meaningful economic impacts. Increased spending has a positive effect on GDP, while increased taxes has a negative effect.

The Blanchard-Perotti (2002) paper is an excellent example of how credible exclusion restrictions can be used to identify a non-recursive structural system to help answer an important economic question. The within-quarter exogeneity of government spending is compelling, and the use of external information to fix the elasticity of tax revenue with respect to GDP is clever.

Structural vector autoregressions can be estimated in Stata using the `svar` command. Short-run restrictions of the form (15.21) can be imposed using the `aeq` and `beq` options. Structural impulse re-

<sup>2</sup>These are similar to, but not the same as, the variables used by Blanchard and Perotti.

<sup>3</sup>The authors detrend their data using a quadratic function of time. By the FWL Theorem this is equivalent to including a quadratic in time in the regression.

<sup>4</sup>The estimated negative effect is difficult to explain, and was not discussed in Blanchard-Perotti.

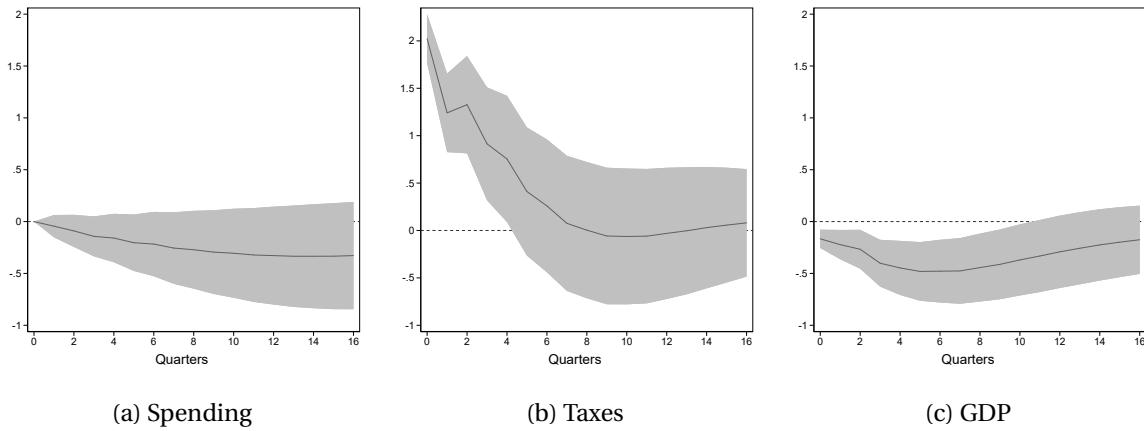


Figure 15.4: Response to a Tax Revenue Shock

sponses can be displayed using `irf graph sirf` and structural forecast error decompositions using `irf graph sfevd`. Unfortunately Stata does not provide a convenient way to display cumulative structural impulse response functions. The same limitations for standard error and confidence interval construction in Stata hold for structural impulse response functions as for non-structural impulse response functions.

## 15.26 Identification of Structural VARs

The coefficient matrices  $\mathbf{A}$  and  $\mathbf{B}$  in (15.21) are identified if they can be uniquely solved from (15.23). This is a set of  $m(m+1)/2$  unique equations so the total number of free coefficients in  $\mathbf{A}$  and  $\mathbf{B}$  cannot be larger than  $m(m+1)/2$ , e.g., 6 when  $m = 3$ . This is the order condition for identification. It is necessary, but not sufficient. It is quite easy to write down restrictions which satisfy the order condition but do not produce an identified system.

It is difficult to see if the system is identified simply by looking at the restrictions (except in the recursive case, which is relatively straightforward to identify). An intuitive way of verifying identification is to use our knowledge of instrumental variables. We can identify the equations sequentially, one at a time, or in blocks, using the metaphor of instrumental variables.

The general technique is as follows. Start by writing out the system imposing all restrictions and absorbing the diagonal elements of  $B$  into the shocks (so that they are still uncorrelated but have non-unit variances). For the Blanchard-Perotti (2002) example, this is

$$\begin{aligned}e_{1t} &= \varepsilon_{1t} \\e_{2t} &= 2.08e_{3t} + b_{21}\varepsilon_{1t} + \varepsilon_{2t} \\e_{3t} &= -a_{31}e_{1t} - a_{32}e_{2t} + \varepsilon_{3t}.\end{aligned}$$

Take the equations one at a time and ask if they can be estimated by instrumental variables using the excluded variables as instruments. Once an equation has been verified as identified, then its shock is identified and can be used as a valid instrument (since it is uncorrelated with the shocks in the other equations).

In this example take the equations as ordered. The first equation is identified as there are no coefficients to estimate. Thus  $\varepsilon_{1t}$  is identified. For the second equation there is one free parameter which can be estimated by least squares of  $e_{2t} - 2.08e_{3t}$  on  $\varepsilon_{1t}$ , which is valid since  $\varepsilon_{1t}$  and  $e_{2t}$  are uncorrelated. This identifies the second equation and the shock  $\varepsilon_{2t}$ . The third equation has two free parameters and two endogenous regressors, so we need two instruments. We can use the shocks  $\varepsilon_{1t}$  and  $\varepsilon_{2t}$ , as they are uncorrelated with  $e_{3t}$  and are correlated with the variables  $e_{1t}$  and  $e_{2t}$ . Thus this equation is identified. We deduce that the system is identified.

Consider another example, based on Keating (1992). He estimated a four-variable system with prices, the fed funds rate, M2, and GDP. His model for the errors takes the form  $\mathbf{A}\boldsymbol{\epsilon}_t = \boldsymbol{\varepsilon}_t$ . Written out explicitly:

$$\begin{aligned} e_P &= \varepsilon_{AS} \\ e_{FF} &= a_{23}e_M + \varepsilon_{MS} \\ e_M &= a_{31}(e_P + e_{GDP}) + a_{32}e_{FF} + \varepsilon_{MD} \\ e_{GDP} &= a_{41}e_P + a_{42}e_{FF} + a_{43}e_M + \varepsilon_{IS} \end{aligned}$$

where the four shocks are “aggregate supply”, “money supply”, “money demand”, and “I-S”. This structure can be based on the following assumptions: An elastic short-run aggregate supply curve (prices do not respond within a quarter); a simple monetary supply policy (the fed funds rate only responds within quarter to the money supply); money demand only responds to nominal output (log price plus log real output) and fed funds rate within quarter; and unrestricted I-S curve.

To analyze conditions for identification we start by checking the order condition. There are 10 coefficients in the system (including the four variances), which equals  $m(m+1)/2$  since  $m = 4$ . Thus the order condition is exactly satisfied.

We then check the equations for identification. We start with the first equation. It has no coefficients so is identified and thus so is  $\varepsilon_{AS}$ . The second equation has one coefficient. We can use  $\varepsilon_{AS}$  as an instrument because it is uncorrelated with  $\varepsilon_{MS}$ . The relevance condition will hold if  $\varepsilon_{AS}$  is correlated with  $e_M$ . From the third equation we see that this will hold if  $a_{31} \neq 0$ . Given this assumption  $a_{23}$  and  $\varepsilon_{MS}$  are identified. The third equation has two coefficients so we can use  $(\varepsilon_{AS}, \varepsilon_{MS})$  as instruments since they are uncorrelated with  $\varepsilon_{MD}$ .  $\varepsilon_{MS}$  is correlated with  $e_{FF}$  and  $\varepsilon_{AS}$  is correlated with  $e_P$ . Thus the relevance condition is satisfied. The final equation has three coefficients, so we use  $(\varepsilon_{AS}, \varepsilon_{MS}, \varepsilon_{MD})$  as instruments. They are uncorrelated with  $\varepsilon_{IS}$  and correlated with the variables  $(e_P, e_{FF}, e_M)$  so this equation is identified.

We find that the system is identified if  $a_{31} \neq 0$ . This requires that money demand responds to nominal GDP, which is a prediction from standard monetary economics. This condition seems reasonable. Regardless, the point of this exercise is to determine specific conditions for identification, and articulate them in your analysis.

## 15.27 Long-Run Restrictions

To review, the algebraic identification problem for impulse response estimation is that we require a square root matrix  $\mathbf{B} = \Sigma^{1/2}$  yet the latter is not unique and the results are sensitive to the choice. The non-uniqueness arises because  $\mathbf{B}$  has  $m^2$  elements while  $\Sigma$  has  $m(m+1)/2$  free elements. The recursive solution is to set  $\mathbf{B}$  to equal the Cholesky decomposition of  $\Sigma$ , or equivalently to specify  $\mathbf{B}$  as lower triangular. Structural VARs based on short-run (contemporaneous) restrictions generalize this idea by allowing general restrictions on  $\mathbf{B}$  based on economic assumptions about contemporaneous causal relations and prior knowledge about  $\mathbf{B}$ . Identification requires  $m(m-1)/2$  restrictions. Even more generally, a structural VAR can be constructed by imposing  $m(m-1)/2$  restrictions due to any known structure or features of the impulse response functions.

One important class of such structural VARs are those based on long-run restrictions. Some economic hypotheses imply restrictions on long-run impulse responses. These can provide a compelling case for identification.

An influential example of a structural VAR based on a long-run restriction is Blanchard and Quah (1989). They were interested in decomposing the effects of demand and supply shocks on output. Their hypothesis is that demand shocks are long-run neutral, meaning that the long-run impact of a demand shock on output is zero. This implies that the long-run impulse response of output with respect to demand is zero. This can be used as an identifying restriction.

The long-run structural impulse response is the cumulative sum of all impulse responses

$$\mathbf{C} = \sum_{\ell=1}^{\infty} \Theta_{\ell} \mathbf{B} = \Theta(1) \mathbf{B} = \mathbf{A}(1)^{-1} \mathbf{B}.$$

A long-run restriction is a restriction placed on the matrix  $\mathbf{C}$ . Since the sum  $\mathbf{A}(1)$  is identified this provides identifying information on the matrix  $\mathbf{B}$ .

Blanchard and Quah (1989) suggest a bivariate VAR for the first-differenced logarithm of real GDP and the unemployment rate. Blanchard-Quah assume that the structural shocks are aggregate supply and aggregate demand. They adopt the hypothesis that aggregate demand has no long-run impact on GDP. This means that the long-run impulse response matrix satisfies

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = \begin{bmatrix} c_{11} & 0 \\ c_{21} & c_{22} \end{bmatrix}. \quad (15.24)$$

Another way of thinking about this is that Blanchard-Quah label “aggregate supply” as the long-run component of GDP and label “aggregate demand” as the transitory component of GDP.

The relations  $\mathbf{C} = \mathbf{A}(1)^{-1} \mathbf{B}$  and  $\mathbf{B}\mathbf{B}' = \Sigma$  imply

$$\mathbf{C}\mathbf{C}' = \mathbf{A}(1)^{-1} \mathbf{B}\mathbf{B}' \mathbf{A}(1)^{-1'} = \mathbf{A}(1)^{-1} \Sigma \mathbf{A}(1)^{-1'}. \quad (15.25)$$

This is a set of  $m^2$  equations but because the matrices are positive semi-definite there are  $m(m+1)/2$  independent equations. If the matrix  $\mathbf{C}$  has  $m(m+1)/2$  free coefficients then the system is identified. This requires  $m(m-1)/2$  restrictions. In the Blanchard-Quah example,  $m = 2$  so one restriction is sufficient for identification.

In many applications, including Blanchard-Quah, the matrix  $\mathbf{C}$  is lower triangular which permits the following elegant solution. Examining (15.25) we see that  $\mathbf{C}$  is a matrix square root of  $\mathbf{A}(1)^{-1} \Sigma \mathbf{A}(1)^{-1'}$ , and since  $\mathbf{C}$  is lower triangular it must be the Cholesky decomposition for which simple algorithms are available. We can then write

$$\mathbf{C} = \text{chol}(\mathbf{A}(1)^{-1} \Sigma \mathbf{A}(1)^{-1}).$$

The plug-in estimator for  $\mathbf{C}$  is

$$\widehat{\mathbf{C}} = \text{chol}(\widehat{\mathbf{A}}(1)^{-1} \widehat{\Sigma} \widehat{\mathbf{A}}(1)^{-1'})$$

where

$$\widehat{\mathbf{A}}(1) = \mathbf{I}_m - \widehat{\mathbf{A}}_1 - \cdots - \widehat{\mathbf{A}}_p.$$

By construction, the solution  $\widehat{\mathbf{C}}$  will be lower triangular and satisfy the desired restriction.

More generally if the restrictions on  $\mathbf{C}$  do not take a lower triangular form then the estimator can be found by numerically solving the system of quadratic equations

$$\widehat{\mathbf{C}}\widehat{\mathbf{C}}' = \widehat{\mathbf{A}}(1)^{-1} \widehat{\Sigma} \widehat{\mathbf{A}}(1)^{-1'}.$$

In either case the estimator for  $\mathbf{B}$  is

$$\widehat{\mathbf{B}} = \widehat{\mathbf{A}}(1)\widehat{\mathbf{C}}$$

and the estimator of the structural impulse response is

$$\widehat{\text{SIRF}}(h) = \widehat{\Theta}_h \widehat{\mathbf{B}} = \widehat{\Theta}_h \widehat{\mathbf{A}}(1)\widehat{\mathbf{C}}.$$

Notice that by construction the long-run impulse response is

$$\sum_{\ell=1}^{\infty} \widehat{\text{SIRF}}(h) = \sum_{\ell=1}^{\infty} \widehat{\Theta}_h \widehat{\mathbf{A}}(1)\widehat{\mathbf{C}} = \widehat{\mathbf{A}}(1)^{-1} \widehat{\mathbf{A}}(1)\widehat{\mathbf{C}} = \widehat{\mathbf{C}}$$

so indeed  $\widehat{\mathbf{C}}$  is the estimated long-run impulse response, and satisfies the desired restriction.

Long-run structural vector autoregressions can be estimated in Stata using the `svar` command using the `lreq` option. Structural impulse responses can be displayed using `irf graph sirf` and structural forecast error decompositions using `irf graph sfevd`. This Stata option does not produce asymptotic standard errors when imposing long-run restrictions, so for confidence intervals bootstrapping is recommended. The same limitations for such intervals constructed in Stata hold for structural impulse response functions as the other cases discussed.

Unfortunately a major limitation of the Stata `svar` command is that it does not provide a way to display cumulative structural impulse response functions. In order to display these, one needs to cumulate the impulse response estimates. This can be done, but then standard errors and confidence intervals are not available. This means that for serious applied work the programming needs to be done outside of Stata.

## 15.28 Blanchard and Quah (1989) Illustration

As we described in the previous section, Blanchard and Quah (1989) estimated a bivariate VAR in GDP growth and the unemployment rate assuming that the the structural shocks are aggregate supply and aggregate demand, imposing that that the long-run response of GDP with respect to aggregate demand is zero. Their original application used U.S. data for 1950-1987. We revisit using FRED-QD (1959-2017). While Blanchard and Quah used a VAR(8) model, the AIC selects a VAR(3). We use a VAR(4). To ease the interpretation of the impulse responses the unemployment rate is entered negatively (multiplied by  $-1$ ) so that both series are pro-cyclical and positive shocks increase output. Blanchard and Quah used a careful detrending method; instead we including a linear time trend in the estimated VAR.

The fitted reduced form model coefficients satisfy

$$\widehat{\mathbf{A}}(1) = \mathbf{I}_m - \sum_{j=1}^4 \widehat{\mathbf{A}}_j = \begin{pmatrix} 0.42 & 0.05 \\ -0.15 & 0.04 \end{pmatrix}$$

and the residual covariance matrix is

$$\widehat{\Sigma} = \begin{pmatrix} 0.531 & 0.095 \\ 0.095 & 0.053 \end{pmatrix}.$$

We calculate

$$\begin{aligned} \widehat{\mathbf{C}} &= \text{chol}(\widehat{\mathbf{A}}(1)^{-1} \widehat{\Sigma} \widehat{\mathbf{A}}(1)^{-1'}) = \begin{pmatrix} 1.00 & 0 \\ 4.75 & 5.42 \end{pmatrix} \\ \widehat{\mathbf{B}} &= \widehat{\mathbf{A}}(1) \widehat{\mathbf{C}} = \begin{pmatrix} 0.67 & 0.28 \\ 0.05 & 0.23 \end{pmatrix}. \end{aligned}$$

Examining  $\widehat{\mathbf{B}}$ , the unemployment rate is contemporaneously mostly affected by the aggregate demand shock, while GDP growth is affected by both shocks.

Using this square root of  $\widehat{\Sigma}$ , we construct the structural impulse response functions for GDP and the unemployment rate as a function of the two shocks (aggregate supply and aggregate demand). The calculations were done in Stata. Unfortunately the Stata `svar` command is highly limited and does not produce cumulative structural impulse responses, which are needed for GDP (as it is estimated in growth rates). We calculated the impulse responses for GDP by cumulating the impulse responses for GDP growth. This can be done for the point estimates but does not produce standard errors. For confidence intervals explicit programming of the estimation would be required.

In Figures 15.5 and 15.6 we display the estimated structural impulse response functions. Figure 15.5 displays the impulse responses of GDP and Figure 15.6 displays the impulse responses of the (negative) unemployment rate. The left panels display the impulse responses with respect to the aggregate supply shock, and the right panels the impulse responses with respect to the aggregate demand shock. Figure

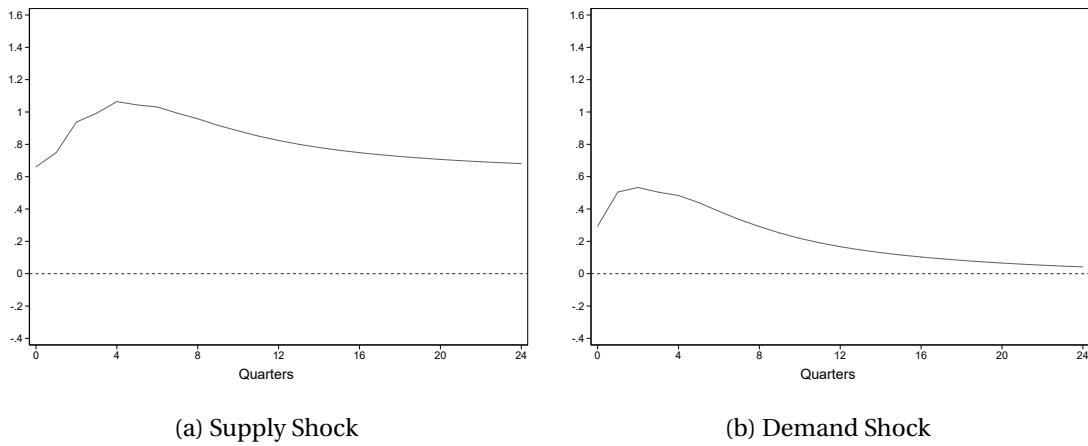


Figure 15.5: Response of GDP

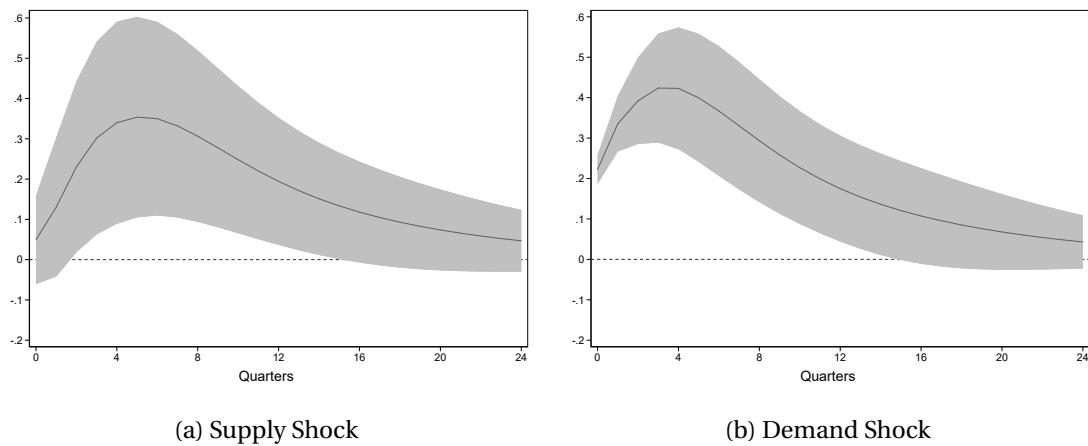


Figure 15.6: Response of Unemployment Rate

15.6 displays 95% normal approximation bootstrap intervals, calculated from 10,00 bootstrap replications. The four estimated impulse responses have similar hump shapes with a peak around four quarters. The estimated functions are similar to those found by Blanchard and Quah (1989).

Let's examine and contrast panels (a) and (b) of Figure 15.5. These are the responses of GDP to aggregate supply and demand shocks, respectively. We can see in panel (a) that the impulse response due to a supply shock is immediate, strong, and persistent. The effect peaks around four quarters, and then flattens, with an effect at 24 quarters similar to the immediate effect. In contrast we can see in panel (b) that the effect of a demand shock is more modest, peaks sooner, and decays, with the effect near zero by 24 quarters. The decay reflects the long-run neutrality of demand shocks. While the estimated effect is transitory the duration of the effect is still meaningful out to three years.

Figure 15.6 displays the responses of the unemployment rate. Its response to a supply shock (panel (a)) takes several quarters to take effect, peaks around 5 quarters, and then decays. The response of the unemployment rate to a demand shock (panel (b)) is more immediate, peaks around 4 quarters, and then decays. Both are near zero by 6 years. The confidence intervals for the supply shock impulse responses are much wider than those for the demand shocks, indicating that the estimates of the impulse responses due to supply shocks are not precisely estimated.

Figure 15.7 displays the estimated structural forecast error decompositions. Since there are only two errors we only display the percentage squared error due to the supply shock. In panel (a) we display the forecast error decomposition for GDP and in panel (b) the forecast error decomposition for the unemployment rate. We can see that about 80% of the fluctuations in GDP are attributed to the supply shock.

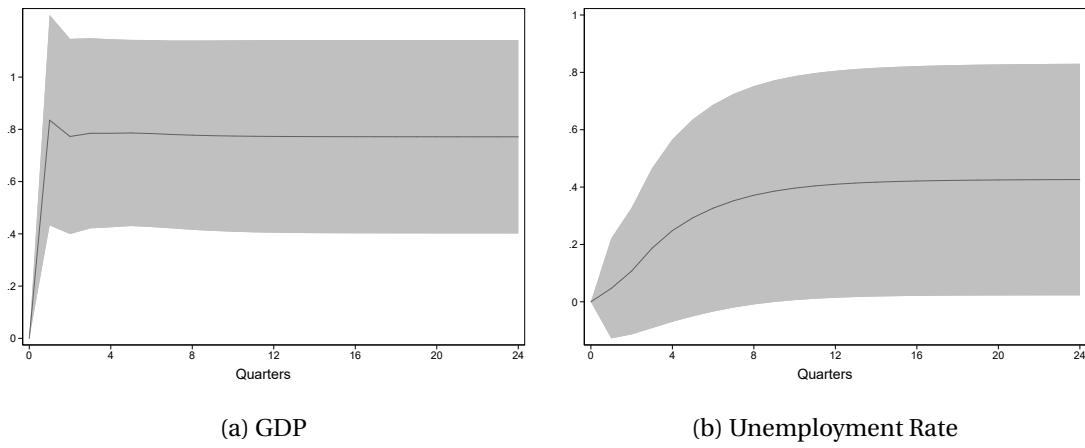


Figure 15.7: Forecast Error Decomposition, % due to Supply Shock

For the unemployment rate, the short-term fluctuations are mostly attributed to the demand shock, but the long-run impact is about 40% due to the supply shock. The confidence intervals are very wide, however, indicating that these estimates are not precise.

It is fascinating that the structural impulse response estimates shown here are nearly identical in form to those found by Blanchard and Quah (1989), despite the fact that we have used a much different sample period.

## 15.29 External Instruments

Structural VARs can also be identified and estimated using **external instrumental variables**. This method is also called **Proxy SVARs**. Consider the three-variable simultaneous equation system for the innovations

$$e_{1t} + a_{12}e_{2t} + a_{13}e_{3t} = \varepsilon_{1t} \quad (15.26)$$

$$a_{21}e_{1t} + e_{2t} = \varepsilon_{2t} + b_{23}\varepsilon_{3t} = u_{2t} \quad (15.27)$$

$$a_{31}e_{1t} + e_{3t} = b_{32}\varepsilon_{2t} + \varepsilon_{3t} = u_{3t}. \quad (15.28)$$

In this system we have used the normalization  $b_{11} = b_{22} = b_{33} = 1$  rather than normalizing the variances of the shocks.

Suppose we have an external instrumental variable  $z_t$  which satisfies the properties

$$\mathbb{E}(z_t \varepsilon_{1t}) \neq 0 \quad (15.29)$$

$$\mathbb{E}(z_t \varepsilon_{2t}) = 0 \quad (15.30)$$

$$\mathbb{E}(z_t \varepsilon_{3t}) = 0. \quad (15.31)$$

Equation (15.29) is the relevance condition, that the instrument and the shock  $\varepsilon_{1t}$  are correlated. Equations (15.30)-(15.31) are the exogeneity condition, that the instrument is uncorrelated with the shocks  $\varepsilon_{2t}$  and  $\varepsilon_{3t}$ . Identification rests on the validity of these assumptions.

Suppose  $e_{1t}$ ,  $e_{2t}$  and  $e_{3t}$  were observed. Then the coefficient  $a_{21}$  in (15.27) can be estimated by instrumental variables regression of  $e_{2t}$  on  $e_{1t}$  using the instrumental variable  $z_t$ . This is valid because  $z_t$  is uncorrelated with  $u_{2t} = \varepsilon_{2t} + b_{23}\varepsilon_{3t}$  under the assumptions (15.30)-(15.31) yet is correlated with  $e_{1t}$  under (15.29). Given this estimator we obtain a residual  $\hat{u}_{2t}$ . Similarly we can estimate  $a_{31}$  in (15.27) by instrumental variables regression of  $e_{3t}$  on  $e_{1t}$  using the instrumental variable  $z_t$ , obtaining a residual  $\hat{u}_{3t}$ . We can then estimate  $a_{12}$  and  $a_{13}$  in (15.26) by instrumental variables regression of  $e_{1t}$  on  $(e_{2t}, e_{3t})$  using the instrumental variables  $(\hat{u}_{2t}, \hat{u}_{3t})$ . The latter are valid instruments since  $\mathbb{E}(u_{2t}\varepsilon_{1t}) = 0$

and  $\mathbb{E}(u_{3t}\varepsilon_{1t}) = 0$  since the structural errors are uncorrelated, and because  $(u_{2t}, u_{3t})$  is correlated with  $(e_{2t}, e_{3t})$  by construction. This regression also produces a residual  $\hat{\varepsilon}_{1t}$  which is an appropriate estimator for the shock  $\varepsilon_{1t}$ .

This estimation method is not special for a three-variable system; it can be applied for any  $m$ . The identified coefficients are those in the first equation (15.26), the structural shock  $\varepsilon_{1t}$ , and the impacts ( $a_{21}$  and  $a_{31}$ ) of this shock on the other variables. The other shocks  $\varepsilon_{2t}$  and  $\varepsilon_{3t}$  are not separately identified, and their correlation structure ( $b_{23}$  and  $b_{32}$ ) is not identified. An exception arises when  $m = 2$ , in which case all coefficients and shocks are identified.

While  $e_{1t}$ ,  $e_{2t}$  and  $e_{3t}$  are not observed we can replace their values by the residuals  $\hat{e}_{1t}$ ,  $\hat{e}_{2t}$  and  $\hat{e}_{3t}$  from the estimated VAR(p) model. All of the coefficient estimates are then two-step estimators with generated regressors. This affects the asymptotic distribution so conventional asymptotic standard errors should not be used. Bootstrap confidence intervals are appropriate.

The structure (15.26)-(15.28) is convenient as four coefficients can be identified. Other structures can also be used. Consider the structure

$$\begin{aligned} e_{1t} &= \varepsilon_{1t} + b_{12}\varepsilon_{2t} + b_{23}\varepsilon_{3t} \\ e_{2t} &= b_{21}\varepsilon_{1t} + \varepsilon_{2t} + b_{23}\varepsilon_{3t} \\ e_{3t} &= b_{31}\varepsilon_{1t} + b_{32}\varepsilon_{2t} + \varepsilon_{3t} \end{aligned}$$

If the same procedure is applied, we can identify the coefficients  $b_{21}$  and  $b_{31}$  and the shock  $\varepsilon_{1t}$  but no other coefficients or shocks. In this structure the coefficients  $b_{12}$  and  $b_{23}$  cannot be separately identified because the shocks  $\varepsilon_{2t}$  and  $\varepsilon_{3t}$  are not separately identified.

For more details see Stock and Watson (2012) and Mertens and Ravn (2013).

### 15.30 Dynamic Factor Models

Dynamic factor models are increasingly popular in applied time series, in particular for forecasting. For a recent detailed review of the methods see Stock and Watson (2016) and the references therein. For some of the foundational theory see Bai (2003) and Bai and Ng (2002, 2006).

In Section 11.12 we introduced the standard multi-factor model (11.23):

$$\mathbf{x}_t = \mathbf{H}\mathbf{f}_t + \mathbf{u}_t \quad (15.32)$$

where  $\mathbf{x}_t$  and  $\mathbf{u}_t$  are  $k \times 1$ ,  $\mathbf{H}$  is  $k \times r$  with  $r < k$ , and  $\mathbf{f}_t$  is  $r \times 1$ . The elements of  $\mathbf{f}_t$  are called the common factors as they affect all elements of  $\mathbf{x}_t$ . The columns of  $\mathbf{H}$  are called the factor loadings. The variables  $\mathbf{u}_t$  are called the individual errors. It is assumed that the elements of  $\mathbf{x}_t$  have been transformed to be mean zero and have common variances.

In the time-series case it is natural to augment the model to allow for dynamic relationships. In particular we would like to allow  $\mathbf{f}_t$  and  $\mathbf{u}_t$  to be serially correlated. It is convenient to consider vector autoregressive models which can be written using lag operator notation as

$$\mathbf{A}(\mathbf{L})\mathbf{f}_t = \mathbf{v}_t \quad (15.33)$$

$$\mathbf{B}(\mathbf{L})\mathbf{u}_t = \mathbf{e}_t \quad (15.34)$$

where  $\mathbf{A}(\mathbf{L})$  and  $\mathbf{B}(\mathbf{L})$  are lag polynomials with  $p$  and  $q$  lags, respectively. Equations (15.32)-(15.33)-(15.34) together make the standard **dynamic factor model**. To simplify the model and aid identification, further restrictions are often imposed, in particular that the lag polynomial  $\mathbf{B}(\mathbf{L})$  is diagonal.

Furthermore we may wish to generalize (15.32) to allow  $\mathbf{f}_t$  to impact  $\mathbf{x}_t$  via a distributed lag relationship. This generalization can be written as

$$\mathbf{x}_t = \mathbf{H}(\mathbf{L})\mathbf{f}_t + \mathbf{u}_t \quad (15.35)$$

where  $\mathbf{H}(L)$  is an  $\ell^{th}$  order distributed lag of dimension  $k \times r$ . Equation (15.35), however, is not fundamentally different from (15.32). That is, if we define the stacked factor vector  $\mathbf{F}_t = (\mathbf{f}'_t, \mathbf{f}'_{t-1}, \dots, \mathbf{f}'_{t-\ell})'$  then (15.35) can be written in the form (15.32) with  $\mathbf{F}_t$  replacing  $\mathbf{f}_t$  and the matrix  $\mathbf{H}$  replaced by  $(\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_\ell)$ . Hence we will focus on the standard model (15.32)-(15.33)-(15.34).

Define the inverse lag operators  $\mathbf{D}(L) = \mathbf{A}(L)^{-1}$  and  $\mathbf{C}(L) = \mathbf{B}(L)^{-1}$ . Then by applying  $\mathbf{C}(L)$  to (15.32) and  $\mathbf{D}(L)$  to (15.33) we obtain

$$\begin{aligned}\mathbf{C}(L)\mathbf{x}_t &= \mathbf{C}(L)\mathbf{H}\mathbf{f}_t + \mathbf{C}(L)\mathbf{u}_t \\ &= \mathbf{C}(L)\mathbf{H}\mathbf{D}(L)\mathbf{v}_t + \mathbf{e}_t \\ &= \mathbf{H}(L)\mathbf{v}_t + \mathbf{e}_t\end{aligned}$$

where  $\mathbf{H}(L) = \mathbf{C}(L)\mathbf{H}\mathbf{D}(L)$ . For simplicity treat this lag polynomial as if it has  $\ell$  lags. Using the same stacking trick from the previous paragraph and defining  $\mathbf{V}_t = (\mathbf{v}'_t, \mathbf{v}'_{t-1}, \dots, \mathbf{v}'_{t-\ell})'$  we find that this model can be written as

$$\mathbf{C}(L)\mathbf{x}_t = \mathbf{H}\mathbf{V}_t + \mathbf{e}_t \quad (15.36)$$

for some  $k \times r\ell$  matrix  $\mathbf{H}$ . This is known as the **static form** of the dynamic factor model. It shows that  $\mathbf{x}_t$  can be written as a function of its own lags plus a linear function of the serially uncorrelated factors  $\mathbf{V}_t$  and a serially uncorrelated error  $\mathbf{e}_t$ .

The static form (15.36) is convenient as PCA methods can be used for estimation. The model is identical to PCA with additional regressors as described in Section 11.13. (The additional regressors are the lagged values of  $\mathbf{x}_t$ .) In that section it is described how to estimate the coefficients and factors by iterating between multivariate least squares and PCA.

To estimate the explicit dynamic model (15.32)-(15.33)-(15.34) state-space methods are convenient. For details and references see Stock and Watson (2016).

The dynamic factor model (15.32)-(15.33)-(15.34) can be estimated in Stata using the `dfactor` command.

### 15.31 Technical Proofs\*

**Proof of Theorem 15.4.** First, observe that if we write  $\mathbf{A}^\ell = [B_{ij,\ell}]$ ,  $\mathbf{x}_t = (x_{1t}, \dots, x_{mt})'$  and  $\mathbf{u}_t = (u_{1t}, \dots, u_{mt})'$  then  $\mathbf{x}_t = \sum_{\ell=0}^{\infty} \mathbf{A}^\ell \mathbf{u}_{t-\ell}$  is the same as

$$x_{it} = \sum_{j=1}^m \sum_{\ell=0}^{\infty} B_{ij,\ell} u_{jt}.$$

Applying Theorem 14.10, this is convergent, strictly stationary, and ergodic if  $\sum_{\ell=0}^{\infty} |B_{ij,\ell}| < \infty$  for each  $i$  and  $j$ .

By the Jordan matrix decomposition,  $\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$  where  $\mathbf{J} = \text{diag}\{\mathbf{J}_1, \dots, \mathbf{J}_r\}$  is in Jordan normal form. Thus

$$\mathbf{x}_t = \sum_{\ell=0}^{\infty} \mathbf{A}^\ell \mathbf{u}_{t-\ell} = \mathbf{P} \sum_{\ell=0}^{\infty} \mathbf{J}^\ell \mathbf{v}_{t-\ell} \quad (15.37)$$

where  $\mathbf{v}_t = \mathbf{P}^{-1} \mathbf{u}_t$  is strictly stationary and ergodic and satisfies  $\mathbb{E} \|\mathbf{v}_t\| < \infty$ . Since  $\mathbf{J}$  is block diagonal the sum in (15.37) converges if and only if each block converges.

The dimension of each Jordan block  $\mathbf{J}_i$  is determined by the multiplicity of the eigenvalues of  $\mathbf{A}$ . For unique eigenvalues  $\lambda$ ,  $\mathbf{J}_i = \lambda$  so  $\mathbf{J}_i^\ell = \lambda^\ell$  which is absolutely summable if  $|\lambda| < 1$ .

For eigenvalues with double multiplicity the Jordan blocks take the form

$$\mathbf{J}_i = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}$$

where  $\lambda$  is an eigenvalue of  $\mathbf{A}$ . We calculate that

$$\mathbf{J}_i^\ell = \begin{bmatrix} \lambda^\ell & \ell\lambda^{\ell-1} \\ 0 & \lambda^\ell \end{bmatrix}.$$

If  $|\lambda| < 1$  these elements are absolutely summable by Theorem 14.4.

For eigenvalues with higher multiplicity  $s$  the Jordan blocks are  $s \times s$  with a similar form. Similar calculations show that the elements of  $\mathbf{J}_i^\ell$  are absolutely summable if  $|\lambda| < 1$ . This verifies the conditions for summability as required. ■

**Proof of Theorem 15.6.** Factor the autoregressive polynomial as

$$\mathbf{A}(z) = \mathbf{I}_m - \mathbf{A}_1 z - \cdots - \mathbf{A}_p z^p = \prod_{j=1}^p (\mathbf{I}_m - \mathbf{G}_j z).$$

Then

$$\det(\mathbf{A}(z)) = \prod_{j=1}^p \det(\mathbf{I}_m - \mathbf{G}_j z).$$

If  $\lambda$  is a solution to  $\det(\mathbf{A}(z)) = 0$  this means  $\det(\mathbf{I}_m - \mathbf{G}_j \lambda) = 0$  for some  $j$ , which means  $\lambda^{-1}$  is an eigenvalue of  $\mathbf{G}_j$ . The assumption  $|\lambda| > 1$  means the eigenvalues of  $\mathbf{G}_1, \dots, \mathbf{G}_m$  are less than one. By Theorem 15.5 the processes

$$\begin{aligned}\mathbf{u}_{1t} &= \mathbf{G}_1 \mathbf{u}_{1,t-1} + \mathbf{e}_t \\ \mathbf{u}_{2t} &= \mathbf{G}_2 \mathbf{u}_{2,t-1} + \mathbf{u}_{1t} \\ &\vdots \\ \mathbf{u}_{mt} &= \mathbf{G}_m \mathbf{u}_{m,t-1} + \mathbf{u}_{m-1,t}\end{aligned}$$

are all strictly stationary and ergodic. But

$$\begin{aligned}\mathbf{u}_{mt} &= (\mathbf{I}_m - \mathbf{G}_m L)^{-1} \mathbf{u}_{m-1,t} \\ &= (\mathbf{I}_m - \mathbf{G}_m L)^{-1} (\mathbf{I}_m - \mathbf{G}_{m-1} L)^{-1} \mathbf{u}_{m-2,t} \\ &= \prod_{j=1}^p (\mathbf{I}_m - \mathbf{G}_j L)^{-1} \mathbf{e}_t \\ &= \mathbf{A}(L)^{-1} \mathbf{e}_t \\ &= \mathbf{y}_t.\end{aligned}$$

Thus  $\mathbf{y}_t$  is strictly stationary and ergodic. ■

**Proof of Theorem 15.7.** The assumption that  $\Sigma > 0$  means that if we regress  $y_{1t}$  on  $y_{2t}, \dots, y_{pt}$  and  $\mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-p}$  that the error will have positive variance. If  $\mathbf{Q}$  is singular then there is some  $\boldsymbol{\gamma}$  such that  $\boldsymbol{\gamma}' \mathbf{Q} \boldsymbol{\gamma} = 0$ . As in the proof of Theorem 14.31 This means that the regression of  $y_{1t}$  on  $y_{2t}, \dots, y_{pt}, \mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-p+1}$  has a zero variance. This is a contradiction. We conclude that  $\mathbf{Q}$  is not singular. ■

**Proof of Theorem 15.11.** The first part of the theorem is established by back-substitution. Since  $\mathbf{y}_t$  is a VAR(p) process,

$$\mathbf{y}_{t+h} = \mathbf{a}_0 + \mathbf{A}_1 \mathbf{y}_{t+h-1} + \mathbf{A}_2 \mathbf{y}_{t+h-2} + \cdots + \mathbf{A}_p \mathbf{y}_{t+h-p} + \mathbf{e}_t.$$

We then substitute out the first lag. We find

$$\begin{aligned}\mathbf{y}_{t+h} &= \mathbf{a}_0 + \mathbf{A}_1 (\mathbf{a}_0 + \mathbf{A}_1 \mathbf{y}_{t+h-2} + \mathbf{A}_2 \mathbf{y}_{t+h-3} + \cdots + \mathbf{A}_p \mathbf{y}_{t+h-p-1} + \mathbf{e}_{t-1}) + \mathbf{A}_2 \mathbf{y}_{t+h-2} + \cdots + \mathbf{A}_p \mathbf{y}_{t+h-p} + \mathbf{e}_t \\ &= \mathbf{a}_0 + \mathbf{A}_1 \mathbf{a}_0 + (\mathbf{A}_1 \mathbf{A}_1 + \mathbf{A}_2) \mathbf{y}_{t+h-2} + (\mathbf{A}_1 \mathbf{A}_2 + \mathbf{A}_3) \mathbf{y}_{t+h-3} + \cdots + \mathbf{A}_p \mathbf{A}_p \mathbf{y}_{t+h-p-1} + \mathbf{A}_1 \mathbf{e}_{t-1} + \mathbf{e}_t.\end{aligned}$$

We continue making substitutions. With each substitution the error increases its MA order. After  $h-1$  substitutions the equation takes the form (15.12) with  $\mathbf{u}_t$  an MA(h-1) process.

To recognize that  $\mathbf{B}_1 = \Theta_h$ , notice that the deduction that  $\mathbf{u}_t$  is an MA(h-1) process means that we can equivalently write (15.12) as

$$\mathbf{y}_{t+h} = \mathbf{b}_0 + \sum_{j=1}^{\infty} \mathbf{B}_j \mathbf{y}_{t+1-j} + \mathbf{u}_t$$

with  $\mathbf{B}_j = 0$  for  $j > p$ . That is, the equation (15.12) includes all relevant lags. By the projection properties of regression coefficients, this means that the coefficient  $\mathbf{B}_1$  is invariant to replacing the regressor  $\mathbf{y}_t$  by the innovation from its regression on the other lags. This is the VAR(p) model itself which has innovation  $\mathbf{e}_t$ . We have deduced that the coefficient  $\mathbf{B}_1$  is equivalent to that in the regression

$$\mathbf{y}_{t+h} = \mathbf{b}_0 + \mathbf{B}_1 \mathbf{e}_t + \sum_{j=2}^{\infty} \mathbf{B}_j \mathbf{y}_{t+1-j} + \mathbf{u}_t.$$

Notice that  $\mathbf{e}_t$  is uncorrelated with the other regressors. Thus  $\mathbf{B}_1 = \frac{\partial}{\partial \mathbf{e}_t'} \mathcal{P}_t(\mathbf{y}_{t+h}) = \Theta_h$  as claimed. This completes the proof. ■

## 15.32 Exercises

**Exercise 15.1** Take the VAR(1) model  $\mathbf{y}_t = \mathbf{A}\mathbf{y}_{t-1} + \mathbf{e}_t$ . Assume  $\mathbf{e}_t$  is i.i.d. For each specified matrix  $\mathbf{A}$  below, check if  $\mathbf{y}_t$  is strictly stationary. Feel free to use mathematical software to compute eigenvalues if needed.

$$(a) \mathbf{A} = \begin{bmatrix} 0.7 & 0.2 \\ 0.2 & 0.7 \end{bmatrix}$$

$$(b) \mathbf{A} = \begin{bmatrix} 0.8 & 0.4 \\ 0.4 & 0.8 \end{bmatrix}$$

$$(c) \mathbf{A} = \begin{bmatrix} 0.8 & 0.4 \\ -0.4 & 0.8 \end{bmatrix}$$

**Exercise 15.2** Take the VAR(2) model  $\mathbf{y}_t = \mathbf{A}_1\mathbf{y}_{t-1} + \mathbf{A}_2\mathbf{y}_{t-2} + \mathbf{e}_t$  with  $\mathbf{A}_1 = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$  and  $\mathbf{A}_2 = \begin{bmatrix} 0.4 & -0.1 \\ -0.1 & 0.4 \end{bmatrix}$ . Assume  $\mathbf{e}_t$  is i.i.d. Is  $\mathbf{y}_t$  strictly stationary? Feel free to use mathematical software if needed.

**Exercise 15.3** Suppose  $\mathbf{y}_t = \mathbf{A}\mathbf{y}_{t-1} + \mathbf{u}_t$  and  $\mathbf{u}_t = \mathbf{B}\mathbf{u}_{t-1} + \mathbf{e}_t$ . Show that  $\mathbf{y}_t$  is a VAR(2) and derive the coefficient matrices and equation error.

**Exercise 15.4** Suppose  $y_{it}$ ,  $i = 1, \dots, m$ , are independent AR(p) processes. Derive the form of their joint VAR representation.

**Exercise 15.5** In the VAR(1) model  $\mathbf{y}_t = \mathbf{A}_1\mathbf{y}_{t-1} + \mathbf{e}_t$  find an explicit expression for the  $h$ -step moving average matrix  $\Theta_h$  from (15.4).

**Exercise 15.6** In the VAR(2) model  $\mathbf{y}_t = \mathbf{A}_1\mathbf{y}_{t-1} + \mathbf{A}_2\mathbf{y}_{t-2} + \mathbf{e}_t$  find explicit expressions for the moving average matrix  $\Theta_h$  from (15.4) for  $h = 1, \dots, 4$ .

**Exercise 15.7** Derive a VAR(1) representation of a VAR(p) process analogously to equation (14.40) for autoregressions. Use this to derive an explicit formula for the  $h$ -step impulse response  $IRF(h)$  analogously to (14.41).

**Exercise 15.8** Let  $\mathbf{y}_t = (y_{1t}, y_{2t})'$  be  $2 \times 1$  and consider a VAR(2) model. Suppose  $y_{2t}$  does not Granger-cause  $y_{1t}$ . What are the implications for the VAR coefficient matrices  $\mathbf{A}_1$  and  $\mathbf{A}_2$ ?

**Exercise 15.9** Continuing the previous exercise, suppose that both  $y_{2t}$  does not Granger-cause  $y_{1t}$ , and  $y_{1t}$  does not Granger-cause  $y_{2t}$ . What are the implications for the VAR coefficient matrices  $\mathbf{A}_1$  and  $\mathbf{A}_2$ ?

**Exercise 15.10** Suppose that you have 20 years of monthly observations on  $m = 8$  variables. Your advisor generally recommends  $p = 12$  lags to account for annual patterns. How many coefficients per equation will you be estimating? How many observations do you have? In this context does it make sense to you to estimate a VAR(12) with all eight variables?

**Exercise 15.11** Let  $\hat{\mathbf{e}}_t$  be the least squares residuals from an estimated VAR,  $\hat{\Sigma}$  be the residual covariance matrix, and  $\hat{\mathbf{B}} = \text{chol}(\hat{\Sigma})$ . Show that  $\hat{\mathbf{B}}$  can be calculated by recursive least squares using the residuals.

**Exercise 15.12** Cholesky factorization

$$(a) \text{Derive the Cholesky decomposition of the covariance matrix } \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

$$(b) \text{Write the answer for the correlation matrix (the special case } \sigma_1^2 = 1 \text{ and } \sigma_2^2 = 1\text{).}$$

- (c) Find an upper triangular decomposition for the correlation matrix. That is, an upper-triangular matrix  $R$  which satisfies  $RR' = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ .
- (d) Suppose  $\Theta_h = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ ,  $\sigma_1^2 = 1$ , and  $\sigma_2^2 = 1$ , and  $\rho = 0.8$ . Find the orthogonalized impulse response  $OIRF(h)$  using the Cholesky decomposition.
- (e) Suppose that the ordering of the variables is reversed. This is equivalent to using the upper triangular decomposition from part (c). Calculate the orthogonalized impulse response  $OIRF(h)$ .
- (f) Compare the two orthogonalized impulse responses.

**Exercise 15.13** You read an empirical paper which estimates a VAR in a listed set of variables, and displays estimated orthogonalized impulse response functions. No comment is made in the paper about the ordering or the identification of the system, and you have no reason to believe that the order used is “standard” in the literature. How should you interpret the estimated impulse response functions?

**Exercise 15.14** Take the quarterly series *gdpc1* (real GDP), *gdpc1pi* (GDP price deflator), and *fedfunds* (Fed funds interest rate) from FRED-QD. Transform the first two into growth rates as in Section 15.13. Estimate the same three-variable VAR(6) using the same ordering. The identification strategy discussed in Section 15.23 specifies the supply shock as the orthogonalized shock to the GDP equation. Calculate the impulse response function of GDP, the price level, and the Fed funds rate with respect to this supply shock. For the first two this will require calculating the cumulative impulse response function. (Explain why.) Comment on the estimated functions.

**Exercise 15.15** Take the Kilian2009 dataset which has the variables *oil* (oil production), *output* (global economic activity), and *price* (price of crude oil). Estimate an orthogonalized VAR(4) using the same ordering as in Kilian (2009) as described in Section 15.24. (As described in that section, multiply “oil” by  $-1$  so that all shocks increase prices.) Estimate the impulse response of output with respect to the three shocks. Comment on the estimated functions.

**Exercise 15.16** Take the monthly series *permit* (building permits), *houst* (housing starts), and *realln* (real estate loans) from FRED-MD. The listed ordering is motivated by transaction timing. A developer is required to obtain a building permit before they start building a house (the latter is known as a “housing start”). A real estate loan is obtained when the house is purchased.

- Transform *realln* into growth rates (first difference of logs).
- Select an appropriate lag order for the three-variable system by comparing the AIC of VARs of order 1 through 8.
- Estimate the VAR model and plot the impulse response functions of housing starts with respect to the three shocks.
- Interpret your findings.

**Exercise 15.17** Take the quarterly series *gpdic1* (Real Gross Private Domestic Investment), *gdpc1pi* (GDP price deflator), *gdpc1* (real GDP), and *fedfunds* (Fed funds interest rate) from FRED-QD. Transform the first three into logs, e.g.  $gdp = 100\log(gdpc1)$ . Consider a structural VAR based on short-run restrictions. Use a structure of the form  $A\epsilon_t = \epsilon_t$ . Impose the restrictions that the first three variables do not react to the fed funds rate, that investment does not respond to prices, and that prices do not respond to investment. Finally, impose that investment is short-run unit elastic with respect to GDP (in the equation for investment, the  $A$  coefficient on GDP is  $-1$ ).

- (a) Write down the matrix  $A$  similar to (15.22), imposing the identifying constraints as defined above.
- (b) Is the model identified? Is there a condition for identification? Explain.
- (c) In this model are output and price simultaneous, or recursive as in the example described in Section 15.23.
- (d) Estimate the structural VAR using 6 lags, or a different number of your choosing (justify your choice), and include an exogenous time trend. Report your estimates of the  $A$  matrix. Can you interpret the coefficients?
- (e) Estimate and report the following three impulse response functions:
  - 1. The effect of the fed funds rate on GDP.
  - 2. The effect of the GDP shock on GDP.
  - 3. The effect of the GDP shock on Prices.

**Exercise 15.18** Take the Kilian2009 dataset which has the variables *oil* (oil production), *output* (global economic activity), and *price* (price of crude oil). Consider a structural VAR based on short-run restrictions. Use a structure of the form  $A\epsilon_t = \epsilon_t$ . Impose the restrictions that oil production does not respond to output or oil prices, and that output does not respond to oil production. The last restriction can be motivated by the observation that supply disruptions take more than a month to reach the retail market, so the effect on economic activity is similarly delayed by one month.

- (a) Write down the matrix  $A$  similar to (15.22), imposing the identifying constraints as defined above.
- (b) Is the model identified? Is there a condition for identification? Explain.
- (c) Estimate the structural VAR using 4 lags, or a different number of your choosing (justify your choice). (As described in that section, multiply “oil” by  $-1$  so that all shocks increase prices.) Report your estimates of the  $A$  matrix. Can you interpret the coefficients?
- (d) Estimate the impulse response of oil price with respect to the three shocks. Comment on the estimated functions.

**Exercise 15.19** Take the quarterly series *gdpc1* (real GDP), *m1realx* (real M1 money stock), and *cpiaucsl* (CPI) from FRED-QD. Create nominal M1 (multiply *m1realx* times *cpiaucsl*), and transform real GDP and nominal M1 to growth rates. The hypothesis of monetary neutrality is that the nominal money supply has no effect on real outcomes such as GDP. Strict monetary neutrality states that there is no short or long-term effect. Long-run neutrality states that there is no long-term effect.

- (a) To test strict neutrality use a Granger-causality test. Regress GDP growth on four lags of GDP growth and four lags of money growth. Test the hypothesis that the four money lags jointly have zero coefficients. Use robust standard errors. Interpret the results.
- (b) To test long-run neutrality test if the sum of the four coefficients on money growth equals zero. Interpret the results.
- (c) Estimate a structural VAR in real GDP growth and nominal money growth, imposing the long-run neutrality of money. Explain your method.
- (d) Report estimates of the impulse responses of the levels of GDP and nominal money to the two shocks. Interpret the results.

**Exercise 15.20** Shapiro and Watson (1988) estimated a structural VAR imposing long-run constraints. We will replicate a simplified version of their model. Take the quarterly series *hoanbs* (hours worked, nonfarm business sector), *gdpc1* (real GDP), and *gdpcpti* (GDP deflator) from FRED-QD. Transform the first two to growth rates, and for the third (GDP deflator), take the second difference of the logarithm (differenced inflation). Shapiro and Watson estimated a structural model imposing the constraints that labor supply hours are long-run unaffected by output and inflation, and GDP is long-run unaffected by demand shocks. This implies a recursive ordering in the variables for a long-run restriction.

- (a) Write down the matrix  $\mathbf{C}$  as in (15.24), imposing the identifying constraints as defined above.
- (b) Is the model identified?
- (c) Use the AIC to select the number of lags for a VAR.
- (d) Estimate the structural VAR. Report the estimated  $\mathbf{C}$  matrix. Can you interpret the coefficients?
- (e) Estimate the structural impulse responses of the level of GDP with respect to the three shocks.  
Interpret the results.

# Chapter 16

## Non Stationary Time Series

### 16.1 Introduction

This chapter is preliminary.

### 16.2 Trend Stationarity

$$y_t = \mu_0 + \mu_1 t + S_t \quad (16.1)$$

$$S_t = \alpha_1 S_{t-1} + \alpha_2 S_{t-2} + \cdots + \alpha_p S_{t-p} + e_t, \quad (16.2)$$

or

$$y_t = \alpha_0 + \gamma t + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + e_t. \quad (16.3)$$

There are two essentially equivalent ways to estimate the autoregressive parameters  $(\alpha_1, \dots, \alpha_p)$ .

- You can estimate (16.3) by OLS.
- You can estimate (16.1)-(16.2) sequentially by OLS. That is, first estimate (16.1), get the residual  $\hat{S}_t$ , and then perform regression (16.2) replacing  $S_t$  with  $\hat{S}_t$ . This procedure is sometimes called *Detrending*.

The reason why these two procedures are (essentially) the same is the Frisch-Waugh-Lovell theorem.

### 16.3 Autoregressive Unit Roots

The AR(p) model is

$$\begin{aligned} \alpha(L)y_t &= \alpha_0 + e_t \\ \alpha(L) &= 1 - \alpha_1 L - \cdots - \alpha_p L^p. \end{aligned}$$

As we discussed before,  $y_t$  has a unit root when  $\alpha(1) = 0$ , or

$$\alpha_1 + \alpha_2 + \cdots + \alpha_p = 1.$$

In this case,  $y_t$  is non-stationary. The ergodic theorem and MDS CLT do not apply, and test statistics are asymptotically non-normal.

A helpful way to write the equation is the so-called Dickey-Fuller reparameterization:

$$\Delta y_t = \rho_0 y_{t-1} + \rho_1 \Delta y_{t-1} + \cdots + \rho_{p-1} \Delta y_{t-(p-1)} + e_t. \quad (16.4)$$

These models are equivalent linear transformations of one another. The DF parameterization is convenient because the parameter  $\rho_0$  summarizes the information about the unit root, since  $\alpha(1) = -\rho_0$ . To see this, observe that the lag polynomial for the  $y_t$  computed from (16.4) is

$$(1 - L) - \rho_0 L - \rho_1(L - L^2) - \cdots - \rho_{p-1}(L^{p-1} - L^p)$$

But this must equal  $\rho(L)$ , as the models are equivalent. Thus

$$\alpha(1) = (1 - 1) - \rho_0 - (1 - 1) - \cdots - (1 - 1) = -\rho_0.$$

Hence, the hypothesis of a unit root in  $y_t$  can be stated as

$$\mathbb{H}_0 : \rho_0 = 0.$$

Note that the model is stationary if  $\rho_0 < 0$ . So the natural alternative is

$$\mathbb{H}_1 : \rho_0 < 0.$$

Under  $\mathbb{H}_0$ , the model for  $y_t$  is

$$\Delta y_t = \mu + \rho_1 \Delta y_{t-1} + \cdots + \rho_{p-1} \Delta y_{t-(p-1)} + e_t,$$

which is an AR(p-1) in the first-difference  $\Delta y_t$ . Thus if  $y_t$  has a (single) unit root, then  $\Delta y_t$  is a stationary AR process. Because of this property, we say that if  $y_t$  is non-stationary but  $\Delta^d y_t$  is stationary, then  $y_t$  is “integrated of order  $d$ ”, or  $I(d)$ . Thus a time series with unit root is  $I(1)$ .

Since  $\alpha_0$  is the parameter of a linear regression, the natural test statistic is the t-statistic for  $\mathbb{H}_0$  from OLS estimation of (16.4). Indeed, this is the most popular unit root test, and is called the Augmented Dickey-Fuller (ADF) test for a unit root.

It would seem natural to assess the significance of the ADF statistic using the normal table. However, under  $\mathbb{H}_0$ ,  $y_t$  is non-stationary, so conventional normal asymptotics are invalid. An alternative asymptotic framework has been developed to deal with non-stationary data. We do not have the time to develop this theory in detail, but simply assert the main results.

**Theorem 16.1 Dickey-Fuller Theorem.**

If  $\rho_0 = 0$  then as  $n \rightarrow \infty$ ,

$$n\hat{\rho}_0 \xrightarrow{d} (1 - \rho_1 - \rho_2 - \cdots - \rho_{p-1}) \text{DF}_\alpha$$

$$\text{ADF} = \frac{\hat{\rho}_0}{s(\hat{\rho}_0)} \rightarrow \text{DF}_t.$$

The limit distributions  $\text{DF}_\alpha$  and  $\text{DF}_t$  are non-normal. They are skewed to the left, and have negative means.

The first result states that  $\hat{\rho}_0$  converges to its true value (of zero) at rate  $n$ , rather than the conventional rate of  $n^{1/2}$ . This is called a “super-consistent” rate of convergence.

The second result states that the t-statistic for  $\hat{\rho}_0$  converges to a limit distribution which is non-normal, but does not depend on the parameters  $\rho$ . This distribution has been extensively tabulated, and may be used for testing the hypothesis  $\mathbb{H}_0$ . Note: The standard error  $s(\hat{\rho}_0)$  is the conventional (“homoskedastic”) standard error. But the theorem does not require an assumption of homoskedasticity. Thus the Dickey-Fuller test is robust to heteroskedasticity.

Since the alternative hypothesis is one-sided, the ADF test rejects  $\mathbb{H}_0$  in favor of  $\mathbb{H}_1$  when  $\text{ADF} < c$ , where  $c$  is the critical value from the ADF table. If the test rejects  $\mathbb{H}_0$ , this means that the evidence points

to  $y_t$  being stationary. If the test does not reject  $\mathbb{H}_0$ , a common conclusion is that the data suggests that  $y_t$  is non-stationary. This is not really a correct conclusion, however. All we can say is that there is insufficient evidence to conclude whether the data are stationary or not.

We have described the test for the setting of with an intercept. Another popular setting includes as well a linear time trend. This model is

$$\Delta y_t = \mu_1 + \mu_2 t + \rho_0 y_{t-1} + \rho_1 \Delta y_{t-1} + \cdots + \rho_{p-1} \Delta y_{t-(p-1)} + e_t. \quad (16.5)$$

This is natural when the alternative hypothesis is that the series is stationary about a linear time trend. If the series has a linear trend (e.g. GDP, Stock Prices), then the series itself is non-stationary, but it may be stationary around the linear time trend. In this context, it is a silly waste of time to fit an AR model to the level of the series without a time trend, as the AR model cannot conceivably describe this data. The natural solution is to include a time trend in the fitted OLS equation. When conducting the ADF test, this means that it is computed as the t-ratio for  $\rho_0$  from OLS estimation of (16.5).

If a time trend is included, the test procedure is the same, but different critical values are required. The ADF test has a different distribution when the time trend has been included, and a different table should be consulted.

Most texts include as well the critical values for the extreme polar case where the intercept has been omitted from the model. These are included for completeness (from a pedagogical perspective) but have no relevance for empirical practice where intercepts are always included.

## 16.4 Cointegration

The idea of cointegration is due to Granger (1981), and was articulated in detail by Engle and Granger (1987).

**Definition 16.1** The  $m \times 1$  series  $\mathbf{y}_t$  is **cointegrated** if  $\mathbf{y}_t$  is  $I(1)$  yet there exists  $\boldsymbol{\beta}$ ,  $m \times r$ , of rank  $r$ , such that  $\mathbf{z}_t = \boldsymbol{\beta}' \mathbf{y}_t$  is  $I(0)$ . The  $r$  vectors in  $\boldsymbol{\beta}$  are called the **cointegrating vectors**.

If the series  $\mathbf{y}_t$  is not cointegrated, then  $r = 0$ . If  $r = m$ , then  $\mathbf{y}_t$  is  $I(0)$ . For  $0 < r < m$ ,  $\mathbf{y}_t$  is  $I(1)$  and cointegrated.

In some cases, it may be believed that  $\boldsymbol{\beta}$  is known a priori. Often,  $\boldsymbol{\beta} = (1 \ -1)'$ . For example, if  $\mathbf{y}_t$  is a pair of interest rates, then  $\boldsymbol{\beta} = (1 \ -1)'$  specifies that the spread (the difference in returns) is stationary. If  $\mathbf{y} = (\log(C) \ \log(I))'$ , then  $\boldsymbol{\beta} = (1 \ -1)'$  specifies that  $\log(C/I)$  is stationary.

In other cases,  $\boldsymbol{\beta}$  may not be known.

If  $\mathbf{y}_t$  is cointegrated with a single cointegrating vector ( $r = 1$ ), then it turns out that  $\boldsymbol{\beta}$  can be consistently estimated by an OLS regression of one component of  $\mathbf{y}_t$  on the others. Thus  $\mathbf{y}_t = (Y_{1t}, Y_{2t})$  and  $\boldsymbol{\beta} = (\beta_1 \ \beta_2)$  and normalize  $\beta_1 = 1$ . Then  $\hat{\beta}_2 = (\mathbf{y}'_2 \mathbf{y}_2)^{-1} \mathbf{y}'_2 \mathbf{y}_1 \xrightarrow{P} \beta_2$ . Furthermore this estimator is super-consistent:  $T(\hat{\beta}_2 - \beta_2) = O_p(1)$ , as first shown by Stock (1987). While OLS is not, in general, a good method to estimate  $\boldsymbol{\beta}$ , it is useful in the construction of alternative estimators and tests.

We are often interested in testing the hypothesis of no cointegration:

$$\mathbb{H}_0 : r = 0$$

$$\mathbb{H}_1 : r > 0.$$

Suppose that  $\boldsymbol{\beta}$  is known, so  $\mathbf{z}_t = \boldsymbol{\beta}' \mathbf{y}_t$  is known. Then under  $\mathbb{H}_0$   $\mathbf{z}_t$  is  $I(1)$ , yet under  $\mathbb{H}_1$   $\mathbf{z}_t$  is  $I(0)$ . Thus  $\mathbb{H}_0$  can be tested using a univariate ADF test on  $\mathbf{z}_t$ .

When  $\boldsymbol{\beta}$  is unknown, Engle and Granger (1987) suggested using an ADF test on the estimated residual  $\hat{z}_t = \hat{\boldsymbol{\beta}}' \mathbf{y}_t$ , from OLS of  $y_{1t}$  on  $y_{2t}$ . Their justification was Stock's result that  $\hat{\boldsymbol{\beta}}$  is super-consistent under  $\mathbb{H}_1$ .

Under  $H_0$ , however,  $\hat{\beta}$  is not consistent, so the ADF critical values are not appropriate. The asymptotic distribution was worked out by Phillips and Ouliaris (1990).

When the data have time trends, it may be necessary to include a time trend in the estimated cointegrating regression. Whether or not the time trend is included, the asymptotic distribution of the test is affected by the presence of the time trend.

## 16.5 Cointegrated VARs

We can write a VAR as

$$\begin{aligned} \mathbf{A}(L)\mathbf{y}_t &= \mathbf{e}_t \\ \mathbf{A}(L) &= \mathbf{I} - \mathbf{A}_1 L - \mathbf{A}_2 L^2 - \cdots - \mathbf{A}_k L^k \end{aligned}$$

or alternatively as

$$\Delta \mathbf{y}_t = \boldsymbol{\Pi} \mathbf{y}_{t-1} + \mathbf{D}(L) \Delta \mathbf{y}_{t-1} + \mathbf{e}_t$$

where

$$\begin{aligned} \boldsymbol{\Pi} &= -\mathbf{A}(1) \\ &= -\mathbf{I} + \mathbf{A}_1 + \mathbf{A}_2 + \cdots + \mathbf{A}_k. \end{aligned}$$

**Theorem 16.2 Granger Representation Theorem**

$\mathbf{y}_t$  is cointegrated with  $m \times r$   $\beta$  if and only if  $\text{rank}(\boldsymbol{\Pi}) = r$  and  $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$  where  $\boldsymbol{\alpha}$  is  $m \times r$ ,  $\text{rank}(\boldsymbol{\alpha}) = r$ .

Thus cointegration imposes a restriction upon the parameters of a VAR. The restricted model can be written as

$$\begin{aligned} \Delta \mathbf{y}_t &= \boldsymbol{\alpha}\boldsymbol{\beta}' \mathbf{y}_{t-1} + \mathbf{D}(L) \Delta \mathbf{y}_{t-1} + \mathbf{e}_t \\ \Delta \mathbf{y}_t &= \boldsymbol{\alpha} \mathbf{z}_{t-1} + \mathbf{D}(L) \Delta \mathbf{y}_{t-1} + \mathbf{e}_t. \end{aligned}$$

If  $\boldsymbol{\beta}$  is known, this can be estimated by OLS of  $\Delta \mathbf{y}_t$  on  $\mathbf{z}_{t-1}$  and the lags of  $\Delta \mathbf{y}_t$ .

If  $\boldsymbol{\beta}$  is unknown, then estimation is done by “reduced rank regression”, which is least-squares subject to the stated restriction. Equivalently, this is the MLE of the restricted parameters under the assumption that  $\mathbf{e}_t$  is iid  $N(\mathbf{0}, \boldsymbol{\Omega})$ .

One difficulty is that  $\boldsymbol{\beta}$  is not identified without normalization. When  $r = 1$ , we typically just normalize one element to equal unity. When  $r > 1$ , this does not work, and different authors have adopted different identification schemes.

In the context of a cointegrated VAR estimated by reduced rank regression, it is simple to test for cointegration by testing the rank of  $\boldsymbol{\Pi}$ . These tests are constructed as likelihood ratio (LR) tests. As they were discovered by Johansen (1988, 1991, 1995), they are typically called the “Johansen Max and Trace” tests. Their asymptotic distributions are non-standard, and are similar to the Dickey-Fuller distributions.

# Chapter 17

## Panel Data

### 17.1 Introduction

Economists traditionally use the term **panel data** to refer to data structures consisting of observations on individuals for multiple time periods. Other fields such as statistics typically call this structure **longitudinal data**. The observed “individuals” can be, for example, people, households, workers, firms, schools, production plants, industries, regions, states, or countries. The distinguishing feature relative to cross-sectional data sets is the presence of multiple observations for each individual. More broadly, panel data methods can be applied to any context with cluster-type dependence.

There are several distinct advantages of panel data relative to cross-section data. One is the possibility of controlling for unobserved time-invariant endogeneity without the use of instrumental variables. A second is the possibility of allowing for broader forms of heterogeneity. A third is modeling dynamic relationships and effects.

There are two broad categories of panel data sets in economic applications: micro panels and macro panels. Micro panels are typically surveys or administrative records on individuals and are characterized by a large number of individuals (often in the 1000's or higher) and a relatively small number of time periods (often 2 to 20 years). Macro panels are typically national or regional macroeconomic variables and are characterized by a moderate number of individuals (e.g. 7-20) and a moderate number of time periods (20-60 years).

Panel data was once relatively esoteric in applied economic practice. Now, it is a dominant feature of applied research.

A typical maintained assumption for micro panels (which we follow in this chapter) is that the individuals are mutually independent while the observations for a given individual are correlated across time periods. This means that the observations follow a clustered dependence structure. Because of this, current econometric practice is to use cluster-robust covariance matrix estimators when possible. Similar assumptions are often used for macro panels, though the assumption of independence across individuals (e.g. countries) is much less compelling.

The application of panel data methods in econometrics started with the pioneering work of Mundlak (1961) and Balestra and Nerlove (1966).

Several excellent monographs and textbooks have been written on panel econometrics, including Arellano (2003), Hsiao (2003), Wooldridge (2010), and Baltagi (2013). This chapter will summarize some of the main themes, but for a more in-depth treatment see these references.

One challenge arising in panel data applications is that the computational methods can require meticulous attention to detail. It is therefore advised to use established packages for routine applications. For most panel data applications in economics, Stata is the standard application package.

## 17.2 Time Indexing and Unbalanced Panels

It is typical to index observations by both the individual  $i$  and the time period  $t$ , thus  $y_{it}$  denotes a variable for individual  $i$  in period  $t$ . We will index individuals as  $i = 1, \dots, N$  and time periods as  $t = 1, \dots, T$ . Thus  $N$  is the number of individuals in the panel and  $T$  is the number of time series periods.

Panel data sets can involve data at any time series frequency, though the typical application involves annual data. The observations in a data set will be indexed by calendar time, which for the case of annual observations is the year. However, for notational convenience it is customary to denote the time periods as  $t = 1, \dots, T$ , so that  $t = 1$  is the first time period observed and  $T$  is the final time period.

When observations are available on all individuals for the same time periods we say that the panel is **balanced**. In this case there are an equal number  $T$  of observations for each individual, and the total number of observations is  $n = NT$ .

When different time periods are available for the individuals in the sample we say that the panel is **unbalanced**. This is the most common type of panel data set. It does not pose a problem for applications, but does make the notation a bit cumbersome and can also considerably complicate computer programming.

To illustrate, consider the data set `Invest1993` on the textbook webpage. This is a sample of 1962 U.S. firms extracted from Compustat and assembled by Bronwyn Hall, and used in the empirical work in Hall and Hall (1993). In Table 17.1 we display a set of variables from the data set for the first 13 observations. The first variable is the firm code number. The second variable is the year of the observation. These two variables are essential for any panel data analysis. In Table 17.1 you can see that the first firm (#32) is observed for the years 1970 through 1977. The second firm (#209) is observed for 1987 through 1991. You can see that the years vary considerably across the firms, so this is an unbalanced panel.

For unbalanced panels the time index  $t = 1, \dots, T$  denotes the full set of time periods. For example, in the data set `Invest1993` there are observations for the years 1960 through 1991, so the total number of time periods is  $T = 32$ . Each individual is observed for a subset of  $T_i$  periods. The set of time periods for individual  $i$  is denoted as  $S_i$  so that individual-specific sums (over time periods) are written as  $\sum_{t \in S_i}$ .

The observed time periods for a given individual are typically contiguous (for example, in Table 17.1, firm #32 is observed for each year from 1970 through 1977) but in some cases are non-contiguous (if, for example, 1973 was missing for firm #32). The total number of observations in the sample is  $n = \sum_{i=1}^N T_i$ .

Table 17.1: Observations from Investment Data Set

Firm Code Number	Year	$I_{it}$	$\bar{I}_i$	$\dot{I}_{it}$	$Q_{it}$	$\bar{Q}_i$	$\dot{Q}_{it}$	$\hat{e}_{it}$
32	1970	0.122	0.155	-0.033	1.17	0.62	0.55	.
32	1971	0.092	0.155	-0.063	0.79	0.62	0.17	-0.005
32	1972	0.094	0.155	-0.061	0.91	0.62	0.29	-0.005
32	1973	0.116	0.155	-0.039	0.29	0.62	-0.33	0.014
32	1974	0.099	0.155	-0.057	0.30	0.62	-0.32	-0.002
32	1975	0.187	0.155	0.032	0.56	0.62	-0.06	0.086
32	1976	0.349	0.155	0.194	0.38	0.62	-0.24	0.248
32	1977	0.182	0.155	0.027	0.57	0.62	-0.05	0.081
209	1987	0.095	0.071	0.024	9.06	21.57	-12.51	.
209	1988	0.044	0.071	-0.027	16.90	21.57	-4.67	-0.244
209	1989	0.069	0.071	-0.002	25.14	21.57	3.57	-0.257
209	1990	0.113	0.071	0.042	25.60	21.57	4.03	-0.226
209	1991	0.034	0.071	-0.037	31.14	21.57	9.57	-0.283

## 17.3 Notation

This chapter focuses on panel data regression models whose observations are pairs  $(y_{it}, \mathbf{x}_{it})$  where  $y_{it}$  is the dependent variable and  $\mathbf{x}_{it}$  is a  $k$ -vector of regressors. These are the observations on individual  $i$  for time period  $t$ .

It will be useful to cluster the observations at the level of the individual. We borrow the notation from Section 4.21 to write  $\mathbf{y}_i$  as the  $T_i \times 1$  stacked observations on  $y_{it}$  for  $t \in S_i$ , stacked in chronological order. Similarly, we write  $\mathbf{X}_i$  as the  $T_i \times k$  matrix of stacked  $\mathbf{x}'_{it}$  for  $t \in S_i$ , stacked in chronological order.

We will also sometimes use matrix notation for the full sample. To do so, let  $\mathbf{y} = (\mathbf{y}'_1, \dots, \mathbf{y}'_N)'$  denote the  $n \times 1$  vector of stacked  $\mathbf{y}_i$ , and set  $\mathbf{X} = (\mathbf{X}'_1, \dots, \mathbf{X}'_N)'$  similarly.

## 17.4 Pooled Regression

The simplest model in panel regression is pooled regression

$$\begin{aligned} y_{it} &= \mathbf{x}'_{it}\boldsymbol{\beta} + e_{it} \\ \mathbb{E}(\mathbf{x}_{it}e_{it}) &= 0. \end{aligned} \tag{17.1}$$

where  $\boldsymbol{\beta}$  is a  $k \times 1$  coefficient vector and  $e_{it}$  is an error. The model can be written at the level of the individual as

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{e}_i \\ \mathbb{E}(\mathbf{X}'_i\mathbf{e}_i) &= 0 \end{aligned}$$

where  $\mathbf{e}_i$  is  $T_i \times 1$ . The equation for the full sample is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

where  $\mathbf{e}$  is  $n \times 1$ .

The standard estimator of  $\boldsymbol{\beta}$  in the pooled regression model is least squares, which can be written as

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{pool}} &= \left( \sum_{i=1}^N \sum_{t \in S_i} \mathbf{x}_{it} \mathbf{x}'_{it} \right)^{-1} \left( \sum_{i=1}^N \sum_{t \in S_i} \mathbf{x}_{it} y_{it} \right) \\ &= \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{y}_i \right) \\ &= (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{y}). \end{aligned}$$

The vector of least-squares residuals for the  $i^{th}$  individual is  $\hat{\mathbf{e}}_i = \mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}_{\text{pool}}$ . While it is the conventional least-squares estimator, in the context of panel data  $\hat{\boldsymbol{\beta}}_{\text{pool}}$  is called the **pooled regression estimator**.

The pooled regression model is ideally suited for the context where the errors  $e_{it}$  satisfy **strict mean independence**:

$$\mathbb{E}(e_{it} | \mathbf{X}_i) = 0. \tag{17.2}$$

This occurs when the errors  $e_{it}$  are mean independent of all regressors  $\mathbf{x}_{ij}$  for all time periods  $j = 1, \dots, T$ . Strict mean independence is stronger than pairwise mean independence  $\mathbb{E}(e_{it} | \mathbf{x}_{it}) = 0$  as well the projection assumption (17.1). Strict mean independence requires that neither lagged nor future values of  $\mathbf{x}_{it}$  help to forecast  $e_{it}$ . It excludes lagged dependent variables (such as  $y_{it-1}$ ) from  $\mathbf{x}_{it}$  (otherwise  $e_{it}$  would be predictable given  $\mathbf{x}_{it+1}$ ). It also requires that  $\mathbf{x}_{it}$  is exogenous in the sense discussed in Chapter 12.

We now describe some statistical properties of  $\hat{\beta}_{\text{pool}}$  under (17.2). First, notice that by linearity and the cluster-level notation we can write the estimator as

$$\begin{aligned}\hat{\beta}_{\text{pool}} &= \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{X}'_i (\mathbf{X}_i \boldsymbol{\beta} + \mathbf{e}_i) \right) \\ &= \boldsymbol{\beta} + \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{e}_i \right).\end{aligned}$$

Then using (17.2)

$$\mathbb{E}(\hat{\beta}_{\text{pool}} | \mathbf{X}) = \boldsymbol{\beta} + \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{X}'_i \mathbb{E}(\mathbf{e}_i | \mathbf{X}_i) \right) = 0$$

so  $\hat{\beta}_{\text{pool}}$  is unbiased for  $\boldsymbol{\beta}$ .

Under the additional assumption that the error  $e_{it}$  is serially uncorrelated and homoskedastic, the covariance estimator takes a classical form and the classical homoskedastic variance estimator can be used. If the error  $e_{it}$  is heteroskedastic but serially uncorrelated then a heteroskedasticity-robust covariance matrix estimator can be used.

In general, however, we expect the errors  $e_{it}$  to be correlated across time  $t$  for a given individual. This does not necessarily violate (17.2) but invalidates classical covariance matrix estimation. The conventional solution is to use a cluster-robust covariance matrix estimator which allows arbitrary within-cluster dependence. Cluster-robust covariance matrix estimators for pooled regression take the form

$$\hat{\mathbf{V}}_{\text{pool}} = (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^N \mathbf{X}'_i \hat{\mathbf{e}}_i \hat{\mathbf{e}}'_i \mathbf{X}_i \right) (\mathbf{X}' \mathbf{X})^{-1}.$$

As in (4.50) this can be multiplied by a degree-of-freedom adjustment. The adjustment used by the Stata `regress` command is

$$\hat{\mathbf{V}}_{\text{pool}} = \left( \frac{n-1}{n-k} \right) \left( \frac{N}{N-1} \right) (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^N \mathbf{X}'_i \hat{\mathbf{e}}_i \hat{\mathbf{e}}'_i \mathbf{X}_i \right) (\mathbf{X}' \mathbf{X})^{-1}.$$

The pooled regression estimator with cluster-robust standard errors can be obtained using the Stata command `regress cluster(id)` where `id` indicates the individual.

When strict mean independence (17.2) fails, however, the pooled least-squares estimator  $\hat{\beta}_{\text{pool}}$  is not necessarily consistent for  $\boldsymbol{\beta}$ . Since strict mean independence is a strong and typically undesirable restriction, it is typically preferred to adopt one of the alternative estimation approaches described in the following sections.

To illustrate the pooled regression estimator, consider the data set `Invest1993` described earlier. We consider a simple investment model

$$I_{it} = \beta_1 Q_{it-1} + \beta_2 D_{it-1} + \beta_3 CF_{it-1} + \beta_4 T_i + e_{it} \quad (17.3)$$

where  $I$  is investment/assets,  $Q$  is market value/assets,  $CF$  is cash flow/assets,  $D$  is long term debt/assets, and  $T$  is a dummy variable indicating if the corporation's stock is traded on the NYSE or AMEX. The regression also includes 19 dummy variables indicating an industry code. The  $Q$  theory of investment suggests that  $\beta_1 > 0$  while  $\beta_2 = \beta_3 = 0$ . Theories of liquidity constraints suggest that  $\beta_2 < 0$  and  $\beta_3 > 0$ . We will be using this example throughout this chapter. The values of  $I$  and  $Q$  for the first 13 observations are also displayed in Table 17.1.

In Table 17.2 we present the pooled regression estimates of (17.3) in the first column with cluster-robust standard errors.

Table 17.2: Estimates of Investment Equation

	Pooled	Random Effects	Fixed Effects	Two-Way	Hausman-Taylor
$Q_{it-1}$	0.0024 (0.0010)	0.0019 (0.0009)	0.0017 (0.0008)	0.0016 (0.0008)	0.0017 (0.0008)
	0.0096 (0.0041)	-0.0092 (0.0039)	-0.0139 (0.0049)	-0.0140 (0.0051)	0.0132 (0.0050)
$CF_{it-1}$	0.0261 (0.0111)	0.0412 (0.0125)	0.0491 (0.0132)	0.0476 (0.0129)	0.0408 (0.0119)
	-0.0167 (0.0024)	-0.0181 (0.0028)			-0.0348 (0.0048)
Industry Dummies	Yes	Yes	No	No	Yes
Time Effects	No	No	No	Yes	Yes

Cluster-robust standard errors in parenthesis.

## 17.5 One-Way Error Component Model

One approach to panel data regression is to model the correlation structure of the regression error  $e_{it}$ . The most common choice is an error-components structure. The simplest takes the form

$$e_{it} = u_i + \varepsilon_{it} \quad (17.4)$$

where  $u_i$  is an individual-specific effect and  $\varepsilon_{it}$  are idiosyncratic (i.i.d.) errors. This is known as a **one-way error component model**.

In vector notation we can write

$$\mathbf{e}_i = \mathbf{1}_i u_i + \boldsymbol{\varepsilon}_i$$

where  $\mathbf{1}_i$  is a  $T_i \times 1$  vector of 1's.

The one-way error component regression model is

$$y_{it} = \mathbf{x}'_{it} \boldsymbol{\beta} + u_i + \varepsilon_{it}$$

written at the level of the observation, or

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{1}_i u_i + \boldsymbol{\varepsilon}_i$$

written at the level of the individual.

To illustrate why an error-component structure such as (17.4) might be appropriate, examine Table 17.1. In the final column we have included the pooled regression residuals  $\hat{e}_{it}$  for these observations. (There is no residual for the first year for each firm due to the lack of lagged regressors for this observation.) What is quite striking is that the residuals for the second firm (#209) are all highly negative, clustering around -0.25. While informal, this suggests that it may be appropriate to model these errors using (17.4), expecting that firm #209 has a large negative value for its individual effect  $u$ .

## 17.6 Random Effects

The random effects model assumes that the errors  $u_i$  and  $\varepsilon_{it}$  in (17.4) are conditionally mean zero, uncorrelated, and homoskedastic.

**Assumption 17.1** (Random Effects). Model (17.4) holds with

$$\mathbb{E}(\varepsilon_{it} | \mathbf{X}_i) = 0 \quad (17.5)$$

$$\mathbb{E}(\varepsilon_{it}^2 | \mathbf{X}_i) = \sigma_\varepsilon^2 \quad (17.6)$$

$$\mathbb{E}(\varepsilon_{ij}\varepsilon_{it} | \mathbf{X}_i) = 0 \quad (17.7)$$

$$\mathbb{E}(u_i | \mathbf{X}_i) = 0 \quad (17.8)$$

$$\mathbb{E}(u_i^2 | \mathbf{X}_i) = \sigma_u^2 \quad (17.9)$$

$$\mathbb{E}(u_i\varepsilon_{it} | \mathbf{X}_i) = 0 \quad (17.10)$$

where (17.7) holds for all  $j \neq t$ .

Assumption 17.1 is known as a **random effects** specification. It implies that the vector of errors  $\mathbf{e}_i$  for individual  $i$  has the covariance structure

$$\begin{aligned} \mathbb{E}(\mathbf{e}_i | \mathbf{X}_i) &= 0 \\ \mathbb{E}(\mathbf{e}_i \mathbf{e}'_i | \mathbf{X}_i) &= \mathbf{1}_i \mathbf{1}'_i \sigma_u^2 + \mathbf{I}_i \sigma_\varepsilon^2 \\ &= \begin{pmatrix} \sigma_u^2 + \sigma_\varepsilon^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 + \sigma_\varepsilon^2 & \cdots & \sigma_u^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 + \sigma_\varepsilon^2 \end{pmatrix} \\ &= \sigma_\varepsilon^2 \mathbf{\Omega}_i, \end{aligned}$$

say, where  $\mathbf{I}_i$  is an identity matrix of dimension  $T_i$ . Note  $\mathbf{\Omega}_i = \mathbf{I}_i + \mathbf{1}_i \mathbf{1}'_i \sigma_u^2 / \sigma_\varepsilon^2$ .

Observe that Assumptions 17.1.1 and 17.1.4 state that the idiosyncratic error  $\varepsilon_{it}$  and individual-specific error  $u_i$  are strictly mean independent, so the total error  $e_{it}$  is strictly mean independent as well.

The random effects model is equivalent to an **equi-correlation** model. That is, suppose that the error  $e_{it}$  satisfies

$$\mathbb{E}(e_{it} | \mathbf{X}_i) = 0$$

$$\mathbb{E}(e_{it}^2 | \mathbf{X}_i) = \sigma^2$$

and

$$\mathbb{E}(e_{ij}e_{it} | \mathbf{X}_i) = \rho\sigma^2$$

for  $j \neq t$ . These conditions imply that  $e_{it}$  can be written as (17.4) with the components satisfying Assumption 17.1, with  $\sigma_u^2 = \rho\sigma^2$  and  $\sigma_\varepsilon^2 = (1-\rho)\sigma^2$ . Thus random effects and equi-correlation are identical models.

The random effects regression model is

$$y_{it} = \mathbf{x}'_{it} \boldsymbol{\beta} + u_i + \varepsilon_{it}$$

or

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{1}_i u_i + \mathbf{\varepsilon}_i$$

where the errors satisfy Assumption 17.1.

Given the error structure, the natural estimator for  $\boldsymbol{\beta}$  is GLS. Suppose  $\sigma_u^2$  and  $\sigma_\varepsilon^2$  are known. The GLS estimator of  $\boldsymbol{\beta}$  is

$$\hat{\boldsymbol{\beta}}_{\text{gls}} = \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{\Omega}_i^{-1} \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{\Omega}_i^{-1} \mathbf{y}_i \right).$$

A feasible GLS estimator replaces the unknown  $\sigma_u^2$  and  $\sigma_\varepsilon^2$  with estimates. We discuss this in Section 17.15.

We now describe some statistical properties of the estimator under Assumption 17.1. By linearity we can write

$$\hat{\boldsymbol{\beta}}_{\text{gls}} - \boldsymbol{\beta} = \left( \sum_{i=1}^N \mathbf{X}'_i \boldsymbol{\Omega}_i^{-1} \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{X}'_i \boldsymbol{\Omega}_i^{-1} \mathbf{e}_i \right).$$

Thus

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_{\text{gls}} - \boldsymbol{\beta} | \mathbf{X}) = \left( \sum_{i=1}^N \mathbf{X}'_i \boldsymbol{\Omega}_i^{-1} \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{X}'_i \boldsymbol{\Omega}_i^{-1} \mathbb{E}(\mathbf{e}_i | \mathbf{X}_i) \right) = 0.$$

Thus  $\hat{\boldsymbol{\beta}}_{\text{gls}}$  is unbiased for  $\boldsymbol{\beta}$ . The variance of  $\hat{\boldsymbol{\beta}}_{\text{gls}}$  is

$$\mathbf{V}_{\text{gls}} = \left( \sum_{i=1}^n \mathbf{X}'_i \boldsymbol{\Omega}_i^{-1} \mathbf{X}_i \right)^{-1}. \quad (17.11)$$

Now let's compare  $\hat{\boldsymbol{\beta}}_{\text{gls}}$  with the pooled estimator  $\hat{\boldsymbol{\beta}}_{\text{pool}}$ . Under Assumption 17.1 the latter is also unbiased for  $\boldsymbol{\beta}$  and has variance

$$\mathbf{V}_{\text{pool}} = \left( \sum_{i=1}^n \mathbf{X}'_i \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{X}'_i \boldsymbol{\Omega}_i \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^n \mathbf{X}'_i \mathbf{X}_i \right)^{-1}. \quad (17.12)$$

Using the algebra of the Gauss-Markov Theorem, we can deduce that

$$\mathbf{V}_{\text{gls}} \leq \mathbf{V}_{\text{pool}} \quad (17.13)$$

and thus the random effects estimator  $\hat{\boldsymbol{\beta}}_{\text{gls}}$  is more efficient than the pooled estimator  $\hat{\boldsymbol{\beta}}_{\text{pool}}$  under Assumption 17.1. (See Exercise 17.1.) The two variance matrices are identical when there is no individual-specific effect (when  $\sigma_u^2 = 0$ ) for then

$$\mathbf{V}_{\text{gls}} = \mathbf{V}_{\text{pool}} = \left( \sum_{i=1}^n \mathbf{X}'_i \mathbf{X}_i \right)^{-1} \sigma_\varepsilon^2.$$

Under the assumption that the random effects model is a useful approximation but not literally true, then we may consider a cluster-robust covariance matrix estimator such as

$$\hat{\mathbf{V}}_{\text{gls}} = \left( \sum_{i=1}^N \mathbf{X}'_i \boldsymbol{\Omega}_i^{-1} \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{X}'_i \boldsymbol{\Omega}_i^{-1} \hat{\mathbf{e}}_i \hat{\mathbf{e}}'_i \boldsymbol{\Omega}_i^{-1} \mathbf{X}_i \right) \left( \sum_{i=1}^N \mathbf{X}'_i \boldsymbol{\Omega}_i^{-1} \mathbf{X}_i \right)^{-1} \quad (17.14)$$

where  $\hat{\mathbf{e}}_i = \mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}_{\text{gls}}$ . This may be re-scaled by a degree of freedom adjustment if desired.

The random effects estimator  $\hat{\boldsymbol{\beta}}_{\text{gls}}$  can be obtained using the Stata command `xtreg`. The default covariance matrix estimator is (17.11). For the cluster-robust covariance matrix estimator (17.14) use the command `xtreg vce(robust)`. (The `xtset` command must be used first to declare the group identifier. For example, `cusip` is the group identifier in Table 17.1.)

To illustrate, in Table 17.2 we present the random effect regression estimates of the investment model (17.3) in the second column with cluster-robust standard errors (17.14). The point estimates are reasonably different from the pooled regression estimator. The coefficient on debt switches from positive to negative (the latter consistent with theories of liquidity constraints) and the coefficient on cash flow increases significantly in magnitude.

## 17.7 Fixed Effect Model

Consider the one-way error component regression model

$$y_{it} = \mathbf{x}'_{it} \boldsymbol{\beta} + u_i + \varepsilon_{it} \quad (17.15)$$

or

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{1}_i u_i + \boldsymbol{\varepsilon}_i. \quad (17.16)$$

In many applications it is useful to interpret the individual-specific effect  $u_i$  as a time-invariant unobserved missing variable. For example, in a wage regression  $u_i$  may be the unobserved ability of individual  $i$ . In the investment model (17.3)  $u_i$  may be a firm-specific productivity factor.

When  $u_i$  is interpreted as a missing variable it is natural to expect it to be correlated with the regressors  $\mathbf{x}_{it}$ . This is especially the case when  $\mathbf{x}_{it}$  includes choice variables.

To illustrate, consider the entries in Table 17.1. The final column displays the pooled regression residuals  $\hat{e}_{it}$  for the first 13 observations, which we interpret as estimates of the error  $e_{it} = u_i + \varepsilon_{it}$ . As described before, what is particularly striking about the residuals is that they are all strongly negative for firm #209, clustering around  $-0.25$ . We can interpret this as an estimate of  $u_i$  for this firm. Examining the values of the regressor  $Q$  for the two firms, we can also see that firm #209 has very large values (in all time periods) for  $Q$ . (The average value  $\bar{Q}_i$  for the two firms appears in the seventh column.) Thus it appears (though we are only looking at two observations) that  $u_i$  and  $Q_{it}$  may be negatively correlated. It is not reasonable to infer too much from these limited observations (indeed the correlation between  $u_i$  and  $\bar{Q}_i$  is positive in the full sample), but the point is that it seems reasonable that the unobserved common effect  $u_i$  may be correlated with the regressors  $\mathbf{x}_{it}$ .

In the econometrics literature, if the stochastic structure of  $u_i$  is treated as unknown and possibly correlated with  $\mathbf{x}_{it}$  then  $u_i$  is called a **fixed effect**.

Correlation between  $u_i$  and  $\mathbf{x}_{it}$  will cause both pooled and random effect estimators to be biased. This is due to the classic problems of omitted variables bias and endogeneity. To see this in a generated example, view Figure 17.1. This shows a scatter plot of three observations  $(y_{it}, x_{it})$  from each of three firms. The true model is  $y_{it} = 9 - x_{it} + u_i$ . (Thus the true slope coefficient is  $-1$ .) The variables  $u_i$  and  $x_{it}$  are highly correlated, so the fitted pooled regression line through the nine observations has a slope close to  $+1$ . (The random effects estimator is identical.) The apparent positive relationship between  $y$  and  $x$  is driven entirely by the positive correlation between  $x$  and  $u$ . Conditional on  $u$ , however, the slope is  $-1$ . Thus regression techniques which do not control for  $u_i$  will produce biased and inconsistent estimates.

The presence of the unstructured individual effect  $u_i$  means that it is not possible to identify  $\boldsymbol{\beta}$  under a simple projection assumption such as  $\mathbb{E}(\mathbf{x}_{it}\varepsilon_t) = 0$ . It turns out that a sufficient condition for identification is the following.

**Definition 17.1** The regressor  $\mathbf{x}_{it}$  is **strictly exogenous** for the error  $\varepsilon_{it}$  if

$$\mathbb{E}(\mathbf{x}_{is}\varepsilon_{it}) = 0 \quad (17.17)$$

for all  $s = 1, \dots, T$ .

Strict exogeneity is a strong projection condition, meaning that if  $\mathbf{x}_{is}$  for any  $s \neq t$  is added to (17.15) it will have a zero coefficient. Strict exogeneity is a projection analog of strict mean independence

$$\mathbb{E}(\varepsilon_{it} | \mathbf{X}_i) = 0. \quad (17.18)$$

(17.18) implies (17.17), but not conversely. While (17.17) is sufficient for identification and asymptotic theory, we will also use the stronger condition (17.18) for finite sample analysis.

While (17.17) and (17.18) are strong assumptions they are much weaker than (17.2) or Assumption 17.1, which require that the individual effect  $u_i$  is also strictly mean independent. In contrast, (17.17) and (17.18) make no assumptions about  $u_i$ .

Strict exogeneity (17.17) is typically inappropriate in dynamic models. In Section 17.41 we discuss estimation under the weaker assumption of predetermined regressors.

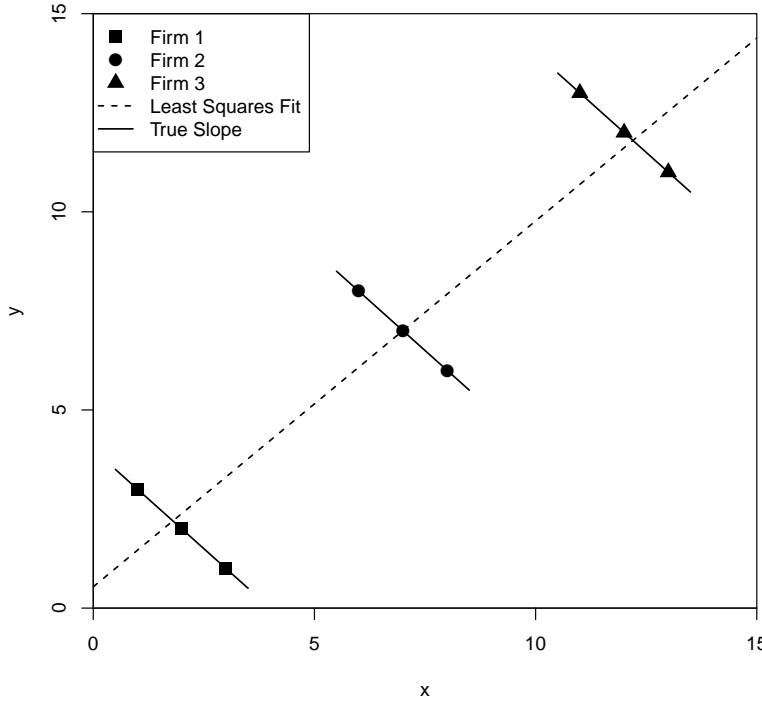


Figure 17.1: Scatter Plot and Pooled Regression Line

## 17.8 Within Transformation

In the previous section we showed that if  $u_i$  and  $\mathbf{x}_{it}$  are correlated, then pooled and random-effects estimators will be biased and inconsistent. If we leave the relationship between  $u_i$  and  $\mathbf{x}_{it}$  fully unstructured, then the only way to consistently estimate the coefficient  $\beta$  is by an estimator which is invariant to  $u_i$ . This can be achieved by transformations which eliminate  $u_i$ .

One such transformation is the **within transformation**. In this section we describe this transformation in detail.

Define the mean of a variable for a given individual as

$$\bar{y}_i = \frac{1}{T_i} \sum_{t \in S_i} y_{it}.$$

We call this the **individual-specific mean**, since it is the mean of a given individual. Contrarywise, some authors call this the **time-average** or **time-mean** since it is the average over the time periods.

Subtracting the individual-specific mean from the variable we obtain the deviations

$$\dot{y}_{it} = y_{it} - \bar{y}_i.$$

This is known as the **within transformation**. We also refer to  $\dot{y}_{it}$  as the **demeaned values** or **deviations from individual means**. Some authors refer to  $\dot{y}_{it}$  as **deviations from time means**. What is important is that the demeaning has occurred at the individual level.

Some algebra may also be useful. We can write the individual-specific mean as  $\bar{y}_i = (\mathbf{1}'_i \mathbf{1}_i)^{-1} \mathbf{1}'_i \mathbf{y}_i$ . Stacking the observations for individual  $i$  we can write the within transformation using the notation

$$\begin{aligned}\dot{\mathbf{y}}_i &= \mathbf{y}_i - \mathbf{1}_i \bar{y}_i \\ &= \mathbf{y}_i - \mathbf{1}_i (\mathbf{1}'_i \mathbf{1}_i)^{-1} \mathbf{1}'_i \mathbf{y}_i \\ &= \mathbf{M}_i \mathbf{y}_i\end{aligned}$$

where

$$\mathbf{M}_i = \mathbf{I}_i - \mathbf{1}_i (\mathbf{1}'_i \mathbf{1}_i)^{-1} \mathbf{1}'_i$$

is the individual-specific demeaning operator. Notice that  $\mathbf{M}_i$  is an idempotent matrix.

Similarly for the regressors we define the individual-specific means and demeaned values:

$$\begin{aligned}\bar{\mathbf{x}}_i &= \frac{1}{T_i} \sum_{t \in S_i} \mathbf{x}_{it} \\ \dot{\mathbf{x}}_{it} &= \mathbf{x}_{it} - \bar{\mathbf{x}}_i \\ \dot{\mathbf{X}}_i &= \mathbf{M}_i \mathbf{X}_i.\end{aligned}$$

We illustrate demeaning in Table 17.1. In the fourth and seventh columns we display the firm-specific means  $\bar{I}_i$  and  $\bar{Q}_i$  and in the fifth and eighth columns the demeaned values  $\dot{I}_{it}$  and  $\dot{Q}_{it}$ .

We can also define the full-sample within operator. Define  $\mathbf{D} = \text{diag}\{\mathbf{1}_{T_1}, \dots, \mathbf{1}_{T_N}\}$  and  $\mathbf{M}_D = \mathbf{I}_n - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'$ . Note that  $\mathbf{M}_D = \text{diag}\{\mathbf{M}_1, \dots, \mathbf{M}_N\}$ . Thus

$$\mathbf{M}_D \mathbf{y} = \dot{\mathbf{y}} = \begin{pmatrix} \dot{y}_1 \\ \vdots \\ \dot{y}_N \end{pmatrix}, \quad \mathbf{M}_D \mathbf{X} = \dot{\mathbf{X}} = \begin{pmatrix} \dot{\mathbf{X}}_1 \\ \vdots \\ \dot{\mathbf{X}}_N \end{pmatrix}. \quad (17.19)$$

Now apply these operations to equation (17.15). Taking individual-specific averages we obtain

$$\bar{y}_i = \bar{\mathbf{x}}'_i \boldsymbol{\beta} + u_i + \bar{\varepsilon}_i \quad (17.20)$$

where  $\bar{\varepsilon}_i = \frac{1}{T_i} \sum_{t \in S_i} \varepsilon_{it}$ . Subtracting from (17.15) we obtain

$$\dot{y}_{it} = \dot{\mathbf{x}}'_{it} \boldsymbol{\beta} + \dot{\varepsilon}_{it} \quad (17.21)$$

where  $\dot{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_{it}$ . The individual effect  $u_i$  has been eliminated!

We can alternatively write this in vector notation. Applying the demeaning operator  $\mathbf{M}_i$  to (17.16) we obtain

$$\dot{\mathbf{y}}_i = \dot{\mathbf{X}}_i \boldsymbol{\beta} + \dot{\varepsilon}_i. \quad (17.22)$$

The individual-effect  $u_i$  is eliminated since  $\mathbf{M}_i \mathbf{1}_i = 0$ . Equation (17.22) is a vector version of (17.21).

The equation (17.21) is a linear equation in the transformed (demeaned) variables. As desired, the individual effect  $u_i$  has been eliminated. Consequently estimators constructed from (17.21) (or equivalently (17.22)) will be invariant to the values of  $u_i$ . This means that the endogeneity bias described in the previous section will be eliminated.

Another consequence, however, is that all time-invariant regressors are also eliminated. That is, if the original model (17.15) had included any regressors  $\mathbf{x}_{it} = \mathbf{x}_i$  which are constant over time for each individual, then for these regressors the demeaned values are identically 0. What this means is that if equation (17.21) is used to estimate  $\boldsymbol{\beta}$  it will be impossible to estimate (or identify) a coefficient on any regressor which is time invariant. This is not a consequence of the estimation method but rather a consequence of the model assumptions. In other words, if the individual effect  $u_i$  has no known structure then it is impossible to disentangle the effect of any time-invariant regressor  $x_i$ . The two have observationally equivalent effects and cannot be separately identified.

The within transformation can greatly reduce the variance of the regressors. This can be seen in Table 17.1, where you can see that the variation between the elements of the transformed variables  $\dot{I}_{it}$  and  $\dot{Q}_{it}$  is less than that of the untransformed variables since much of the variation is captured by the firm-specific means.

It is not typically needed to directly program the within transformation, but if it is desired the following Stata commands easily do so.

### Stata Commands for Within Transformation

- \* x is the original variable
- \* id is the group identifier
- \* xdot is the within-transformed variable

```
egen xmean = mean(x), by(id)
gen xdot = x - xmean
```

## 17.9 Fixed Effects Estimator

Consider least-squares applied to the demeaned equation (17.21) or equivalently (17.22). This is

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{fe}} &= \left( \sum_{i=1}^N \sum_{t \in S_i} \dot{\mathbf{x}}_{it} \dot{\mathbf{x}}'_{it} \right)^{-1} \left( \sum_{i=1}^N \sum_{t \in S_i} \dot{\mathbf{x}}_{it} \dot{y}_{it} \right) \\ &= \left( \sum_{i=1}^N \dot{\mathbf{X}}'_i \dot{\mathbf{X}}_i \right)^{-1} \left( \sum_{i=1}^N \dot{\mathbf{X}}'_i \dot{\mathbf{y}}_i \right) \\ &= \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{M}_i \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{M}_i \mathbf{y}_i \right).\end{aligned}$$

This is known as the **fixed-effects** or **within** estimator of  $\boldsymbol{\beta}$ . It is called the fixed-effects estimator because it is appropriate for the fixed effects model (17.15). It is called the within estimator because it is based on the variation of the data within each individual.

The above definition implicitly assumes that the matrix  $\sum_{i=1}^N \dot{\mathbf{X}}'_i \dot{\mathbf{X}}_i$  is full rank. This requires that all components of  $\mathbf{x}_{it}$  have time variation for at least some individuals in the sample.

The fixed effects residuals are

$$\begin{aligned}\hat{\varepsilon}_{it} &= \dot{y}_{it} - \dot{\mathbf{x}}'_{it} \hat{\boldsymbol{\beta}}_{\text{fe}} \\ \hat{\boldsymbol{\varepsilon}}_i &= \dot{\mathbf{y}}_i - \dot{\mathbf{X}}_i \hat{\boldsymbol{\beta}}_{\text{fe}}.\end{aligned}\tag{17.23}$$

Let us describe some of the statistical properties of the estimator under strict mean independence (17.18). By linearity and the fact  $\mathbf{M}_i \mathbf{1}_i = 0$ , we can write

$$\hat{\boldsymbol{\beta}}_{\text{fe}} - \boldsymbol{\beta} = \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{M}_i \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{M}_i \boldsymbol{\varepsilon}_i \right).$$

Then (17.18) implies

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_{\text{fe}} - \boldsymbol{\beta} | \mathbf{X}) = \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{M}_i \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{M}_i \mathbb{E}(\boldsymbol{\varepsilon}_i | \mathbf{X}_i) \right) = 0.$$

Thus  $\hat{\boldsymbol{\beta}}_{\text{fe}}$  is unbiased for  $\boldsymbol{\beta}$  under (17.18).

Let

$$\boldsymbol{\Sigma}_i = \mathbb{E}(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}'_i | \mathbf{X}_i)$$

denote the  $T_i \times T_i$  conditional covariance matrix of the idiosyncratic errors. The variance of  $\hat{\boldsymbol{\beta}}_{\text{fe}}$  is

$$V_{\text{fe}} = \text{var}(\hat{\boldsymbol{\beta}}_{\text{fe}} | \mathbf{X}) = \left( \sum_{i=1}^N \dot{\mathbf{X}}'_i \dot{\mathbf{X}}_i \right)^{-1} \left( \sum_{i=1}^N \dot{\mathbf{X}}'_i \boldsymbol{\Sigma}_i \dot{\mathbf{X}}_i \right) \left( \sum_{i=1}^N \dot{\mathbf{X}}'_i \dot{\mathbf{X}}_i \right)^{-1}. \tag{17.24}$$

This expression simplifies when the idiosyncratic errors are homoskedastic and serially uncorrelated:

$$\mathbb{E}(\varepsilon_{it}^2 | \mathbf{X}_i) = \sigma_\varepsilon^2 \quad (17.25)$$

$$\mathbb{E}(\varepsilon_{ij}\varepsilon_{it} | \mathbf{X}_i) = 0 \quad (17.26)$$

for all  $j \neq t$ . In this case,  $\Sigma_i = \mathbf{I}_i \sigma_\varepsilon^2$  and (17.24) simplifies to

$$\mathbf{V}_{\text{fe}}^0 = \sigma_\varepsilon^2 \left( \sum_{i=1}^N \dot{\mathbf{X}}_i' \dot{\mathbf{X}}_i \right)^{-1}. \quad (17.27)$$

It is instructive to compare the variances of the fixed-effects estimator and the pooled estimator under (17.25)-(17.26) and the assumption that there is no individual-specific effect  $u_i = 0$ . In this case we see that

$$\mathbf{V}_{\text{fe}}^0 = \sigma_\varepsilon^2 \left( \sum_{i=1}^N \dot{\mathbf{X}}_i' \dot{\mathbf{X}}_i \right)^{-1} \geq \sigma_\varepsilon^2 \left( \sum_{i=1}^N \mathbf{X}_i' \mathbf{X}_i \right)^{-1} = \mathbf{V}_{\text{pool}}. \quad (17.28)$$

The inequality holds since the demeaned variables  $\dot{\mathbf{X}}_i$  have reduced variation relative to the original observations  $\mathbf{X}_i$ . (See Exercise 17.28.) This shows the cost of using fixed effects relative to pooled estimation. The estimation variance increases due to reduced variation in the regressors. This reduction in efficiency is a necessary by-product of the robustness of the estimator to the individual effects  $u_i$ .

## 17.10 Differenced Estimator

The within transformation is not the only transformation which eliminates the individual-specific effect. Another important transformation which does the same is first-differencing.

The **first-differencing** transformation is

$$\Delta y_{it} = y_{it} - y_{it-1}.$$

This can be applied to all but the first observation (which is essentially lost). At the level of the individual this can be written as

$$\Delta \mathbf{y}_i = \mathbf{D}_i \mathbf{y}_i$$

where  $\mathbf{D}_i$  is the  $(T_i - 1) \times T_i$  matrix differencing operator

$$\mathbf{D}_i = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}.$$

Applying the transformation  $\Delta$  to (17.15) or (17.16) we obtain

$$\Delta y_{it} = \Delta \mathbf{x}'_{it} \boldsymbol{\beta} + \Delta \varepsilon_{it}$$

or

$$\Delta \mathbf{y}_i = \Delta \mathbf{X}_i \boldsymbol{\beta} + \Delta \varepsilon_i. \quad (17.29)$$

Least squares applied to the differenced equation is

$$\begin{aligned} \hat{\boldsymbol{\beta}}_\Delta &= \left( \sum_{i=1}^N \sum_{t \geq 2} \Delta \mathbf{x}_{it} \Delta \mathbf{x}'_{it} \right)^{-1} \left( \sum_{i=1}^N \sum_{t \geq 2} \Delta \mathbf{x}_{it} \Delta y_{it} \right) \\ &= \left( \sum_{i=1}^N \Delta \mathbf{X}'_i \Delta \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \Delta \mathbf{X}'_i \Delta \mathbf{y}_i \right) \\ &= \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{D}'_i \mathbf{D}_i \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{D}'_i \mathbf{D}_i \mathbf{y}_i \right). \end{aligned} \quad (17.30)$$

(17.30) is called the **differenced estimator**. For  $T = 2$ ,  $\hat{\beta}_{\Delta} = \hat{\beta}_{\text{fe}}$  equals the fixed effects estimator. See Exercise 17.6. They differ, however, for  $T > 2$ .

When the errors  $\varepsilon_{it}$  are serially uncorrelated and homoskedastic, then the error  $\Delta\varepsilon_i = \mathbf{D}_i\varepsilon_i$  in (17.29) has variance matrix  $\mathbf{H}\sigma_{\varepsilon}^2$  where

$$\mathbf{H} = \mathbf{D}_i \mathbf{D}'_i = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & \ddots & 0 \\ 0 & \ddots & \ddots & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}. \quad (17.31)$$

We can reduce estimation variance by using GLS, which is

$$\begin{aligned} \hat{\beta}_{\bar{\Delta}} &= \left( \sum_{i=1}^N \Delta \mathbf{X}'_i \mathbf{H}^{-1} \Delta \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \Delta \mathbf{X}'_i \mathbf{H}^{-1} \Delta \mathbf{y}_i \right) \\ &= \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{D}'_i (\mathbf{D}_i \mathbf{D}'_i)^{-1} \mathbf{D}_i \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{D}'_i (\mathbf{D}_i \mathbf{D}'_i)^{-1} \mathbf{D}_i \mathbf{y}_i \right) \\ &= \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{P}_D \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{P}_D \mathbf{y}_i \right) \end{aligned}$$

where  $\mathbf{P}_D = \mathbf{D}'_i (\mathbf{D}_i \mathbf{D}'_i)^{-1} \mathbf{D}_i$ . Recall, the matrix  $\mathbf{D}_i$  is  $(T_i - 1) \times T_i$  with rank  $T_i - 1$  and is orthogonal to the vector of ones  $\mathbf{1}_i$ . This means  $\mathbf{P}_D$  projects orthogonally to  $\mathbf{1}_i$  and thus equals  $\mathbf{P}_D = \mathbf{M}_i$ , the within transformation matrix. Hence  $\hat{\beta}_{\bar{\Delta}} = \hat{\beta}_{\text{fe}}$ , the fixed effects estimator!

What we have shown is that GLS applied to the first-differenced equation precisely equals the fixed effects estimator. Since the Gauss-Markov theorem shows that GLS has lower variance than least-squares, this means that the fixed effects estimator is more efficient than first differencing under the assumption that  $\varepsilon_{it}$  is i.i.d.

This argument extends to any other transformation which eliminates the fixed effect. GLS applied after such a transformation is equal to the fixed effects estimator, and is more efficient than least-squares applied after the same transformation. This shows that the fixed effects estimator is Gauss-Markov efficient in the class of estimators which eliminate the fixed effect.

## 17.11 Dummy Variables Regression

An alternative way to estimate the fixed effects model is by least squares of  $y_i$  on  $\mathbf{x}_{it}$  and a full set of dummy variables, one for each individual in the sample. It turns out that this is algebraically equivalent to the within estimator.

To see this, start with the error-component model without a regressor:

$$y_{it} = u_i + \varepsilon_{it}. \quad (17.32)$$

Consider least-squares estimation of the vector of fixed effects  $\mathbf{u} = (u_1, \dots, u_N)'$ . Since each fixed effect  $u_i$  is an individual-specific mean, and the least-squares estimate of the intercept is the sample mean, it follows that the least-squares estimate of  $u_i$  is  $\hat{u}_i = \bar{y}_i$ . The least-squares residual is then  $\hat{\varepsilon}_{it} = y_{it} - \bar{y}_i = \hat{y}_{it}$ , the within transformation.

If you would prefer an algebraic argument, let  $\mathbf{d}_i$  be a vector of  $N$  dummy variables where the  $i^{th}$  element indicates the  $i^{th}$  individual. Thus the  $i^{th}$  element of  $\mathbf{d}_i$  is 1 and the remaining elements are zero. Notice that  $u_i = \mathbf{d}'_i \mathbf{u}$  and (17.32) equals

$$y_{it} = \mathbf{d}'_i \mathbf{u} + \varepsilon_{it}.$$

This is a regression with the regressors  $\mathbf{d}_i$  and coefficients  $\mathbf{u}$ . We can also write this in vector notation at the level of the individual as

$$\mathbf{y}_i = \mathbf{1}_i \mathbf{d}'_i \mathbf{u} + \boldsymbol{\epsilon}_i$$

or using full matrix notation as

$$\mathbf{y} = \mathbf{D}\mathbf{u} + \boldsymbol{\epsilon}$$

where  $\mathbf{D} = \text{diag}\{\mathbf{1}_{T_1}, \dots, \mathbf{1}_{T_N}\}$ .

The least-squares estimate of  $\mathbf{u}$  is

$$\begin{aligned}\hat{\mathbf{u}} &= (\mathbf{D}'\mathbf{D})^{-1}(\mathbf{D}'\mathbf{y}) \\ &= \text{diag}(\mathbf{1}'_i \mathbf{1}_i)^{-1} \text{vec}(\mathbf{1}'_i \mathbf{y}_i) \\ &= \text{vec}\left((\mathbf{1}'_i \mathbf{1}_i)^{-1} \mathbf{1}'_i \mathbf{y}_i\right) \\ &= \text{vec}(\bar{\mathbf{y}}_i).\end{aligned}$$

The least-squares residuals are

$$\hat{\boldsymbol{\epsilon}} = \left(\mathbf{I}_n - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\right)\mathbf{y} = \dot{\mathbf{y}}$$

as shown in (17.19). Thus the least-squares residuals from the simple error-component model are the within transformed variables.

Now consider the error-component model with regressors, which can be written as

$$y_{it} = \mathbf{x}'_{it} \boldsymbol{\beta} + \mathbf{d}'_i \mathbf{u} + \varepsilon_{it} \quad (17.33)$$

since  $u_i = \mathbf{d}'_i \mathbf{u}$  as discussed above. In matrix notation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\mathbf{u} + \boldsymbol{\epsilon}. \quad (17.34)$$

We consider estimation of  $(\boldsymbol{\beta}, \mathbf{u})$  by least-squares, and write the estimates as

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{D}\hat{\mathbf{u}} + \hat{\boldsymbol{\epsilon}}.$$

We call this the **dummy variable estimator** of the fixed effects model.

By the Frisch-Waugh-Lovell Theorem (Theorem 3.5), the dummy variable estimator  $\hat{\boldsymbol{\beta}}$  and residuals  $\hat{\boldsymbol{\epsilon}}$  may be obtained by the least-squares regression of the residuals from the regression of  $\mathbf{y}$  on  $\mathbf{D}$  on the residuals from the regression of  $\mathbf{X}$  on  $\mathbf{D}$ . We learned above that residuals from the regression on  $\mathbf{D}$  are the within transformations. Thus the dummy variable estimator  $\hat{\boldsymbol{\beta}}$  and residuals  $\hat{\boldsymbol{\epsilon}}$  may be obtained from least-squares regression of the within transformed  $\dot{\mathbf{y}}$  on the within transformed  $\dot{\mathbf{X}}$ . This is exactly the fixed effects estimator  $\hat{\boldsymbol{\beta}}_{fe}$ . Thus the dummy variable and fixed effects estimators of  $\boldsymbol{\beta}$  are identical.

This is sufficiently important that we state this result as a theorem.

**Theorem 17.1** The fixed effects estimator of  $\boldsymbol{\beta}$  is algebraically identical to the dummy variable estimator of  $\boldsymbol{\beta}$ . The two estimators also yield the same residuals.

This may be the most important practical application of the Frisch-Waugh-Lovell Theorem. It shows that we can estimate the coefficients either by applying the within transformation, or by inclusion of dummy variables (one for each individual in the sample). This is important because in some cases one approach is more convenient than the other, and it is important to know that the two methods are algebraically equivalent.

When  $N$  is large it is advisable to use the within transformation rather than the dummy variable approach. This is because the latter requires considerably more computer memory. To see this, consider the matrix  $\mathbf{D}$  in (17.34) in the balanced case. It has  $TN^2$  elements, and for dummy variable estimation it must be created and stored in memory. When  $N$  is large this can be excessive. For example, if  $T = 10$  and  $N = 10,000$ , the matrix  $\mathbf{D}$  has one billion elements! Whether or not a package can technically handle a matrix of this dimension depends on several particulars (system RAM, operating system, package version), but even if it can execute the calculation the computation time will be considerably slowed. Hence for fixed effects estimation with large  $N$  it is recommended to use the within transformation rather than dummy variable regression.

The dummy variable formulation may add insight about how the fixed effects estimator achieves invariance to the fixed effects. Given the regression equation (17.34) we can write the least-squares estimator of  $\boldsymbol{\beta}$  using the residual regression formula:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{fe}} &= (\mathbf{X}' \mathbf{M}_D \mathbf{X})^{-1} (\mathbf{X}' \mathbf{M}_D \mathbf{y}) \\ &= (\mathbf{X}' \mathbf{M}_D \mathbf{X})^{-1} (\mathbf{X}' \mathbf{M}_D (\mathbf{X}\boldsymbol{\beta} + \mathbf{D}\mathbf{u} + \boldsymbol{\epsilon})) \\ &= \boldsymbol{\beta} + (\mathbf{X}' \mathbf{M}_D \mathbf{X})^{-1} (\mathbf{X}' \mathbf{M}_D \boldsymbol{\epsilon})\end{aligned}\quad (17.35)$$

since  $\mathbf{M}_D \mathbf{D} = 0$ . The expression (17.35) is free of the vector  $\mathbf{u}$  and thus  $\hat{\boldsymbol{\beta}}_{\text{fe}}$  is invariant to  $\mathbf{u}$ . This is another demonstration that the fixed effects estimator is invariant to the actual values of the fixed effects, and thus its statistical properties do not rely on assumptions about  $u_i$ .

## 17.12 Fixed Effects Covariance Matrix Estimation

First consider estimation of the classical covariance matrix  $\mathbf{V}_{\text{fe}}^0$  as defined in (17.27). This is

$$\hat{\mathbf{V}}_{\text{fe}}^0 = \hat{\sigma}_\epsilon^2 (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1} \quad (17.36)$$

with

$$\hat{\sigma}_\epsilon^2 = \frac{1}{n - N - k} \sum_{i=1}^n \sum_{t \in S_i} \hat{\epsilon}_{it}^2 = \frac{1}{n - N - k} \sum_{i=1}^n \hat{\boldsymbol{\epsilon}}_i' \hat{\boldsymbol{\epsilon}}_i. \quad (17.37)$$

The  $N + k$  degree of freedom adjustment is motivated by the dummy variable representation. Indeed, you can verify that  $\hat{\sigma}_\epsilon^2$  is unbiased for  $\sigma_\epsilon^2$  under assumptions (17.18), (17.25) and (17.26). See Exercise 17.8.

Notice that the assumptions (17.18), (17.25) and (17.26) are identical to (17.5)-(17.7) of Assumption 17.1. The assumptions (17.8)-(17.10) are not needed. Thus the fixed effect model weakens the random effects model by eliminating the assumptions on  $u_i$  but retaining those on  $\epsilon_{it}$ .

The classical covariance matrix estimator (17.36) for the fixed effects estimator is valid when the errors  $\epsilon_{it}$  are homoskedastic and serially uncorrelated but is invalid otherwise. A covariance matrix estimator which allows  $\epsilon_{it}$  to be heteroskedastic and serially correlated across  $t$  is the cluster-robust covariance matrix estimator, clustered by individual

$$\hat{\mathbf{V}}_{\text{fe}}^{\text{cluster}} = (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1} \left( \sum_{i=1}^N \dot{\mathbf{X}}_i' \hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i' \dot{\mathbf{X}}_i \right) (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1} \quad (17.38)$$

where  $\hat{\boldsymbol{\epsilon}}_i$  as the fixed effects residuals as defined in (17.23). (17.38) was first proposed by Arellano (1987). As in (4.50)  $\hat{\mathbf{V}}_{\text{fe}}^{\text{cluster}}$  can be multiplied by a degree-of-freedom adjustment. The adjustment recommended by the theory of C. Hansen (2007) is

$$\hat{\mathbf{V}}_{\text{fe}}^{\text{cluster}} = \left( \frac{N}{N-1} \right) (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1} \left( \sum_{i=1}^N \dot{\mathbf{X}}_i' \hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i' \dot{\mathbf{X}}_i \right) (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1} \quad (17.39)$$

and that corresponding to (4.50) is

$$\hat{V}_{\text{fe}}^{\text{cluster}} = \left( \frac{n-1}{n-N-k} \right) \left( \frac{N}{N-1} \right) (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1} \left( \sum_{i=1}^N \dot{\mathbf{X}}_i' \hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i' \dot{\mathbf{X}}_i \right) (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1}. \quad (17.40)$$

These estimators are convenient because they are simple to apply and allow for unbalanced panels.

In typical micropanel applications,  $N$  is very large and  $k$  is modest. Thus the adjustment in (17.39) is minor, while that in (17.40) is approximately  $\bar{T}/(\bar{T}-1)$  where  $\bar{T} = n/N$  is the average number of time periods per individual. When  $\bar{T}$  is small this can be a very large adjustment. Hence the choice between (17.38), (17.39), or (17.40) can be substantial.

To understand if the degree of freedom adjustment in (17.40) is appropriate, consider the simplified setting where the residuals are constructed with the true  $\boldsymbol{\beta}$ . This is a useful approximation since the number of estimated slope coefficients  $\boldsymbol{\beta}$  is small relative to the sample size  $n$ . Then  $\hat{\boldsymbol{\epsilon}}_i = \dot{\boldsymbol{\epsilon}}_i = \mathbf{M}_i \boldsymbol{\epsilon}_i$  so  $\dot{\mathbf{X}}_i' \hat{\boldsymbol{\epsilon}}_i = \dot{\mathbf{X}}_i' \boldsymbol{\epsilon}_i$  and (17.38) equals

$$\hat{V}_{\text{fe}}^{\text{cluster}} = (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1} \left( \sum_{i=1}^N \dot{\mathbf{X}}_i' \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i' \dot{\mathbf{X}}_i \right) (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1}$$

which is the idealized estimator with the true errors rather than the residuals. Since  $\mathbb{E}(\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i' | \mathbf{X}_i) = \boldsymbol{\Sigma}_i$  it follows that  $\mathbb{E}(\hat{V}_{\text{fe}}^{\text{cluster}} | \mathbf{X}) = V_{\text{fe}}$  and  $\hat{V}_{\text{fe}}^{\text{cluster}}$  is unbiased for  $V_{\text{fe}}$ ! Thus no degree of freedom adjustment is required. This is despite the fact that  $N$  fixed effects have been estimated. While this analysis concerns the idealized case where the residuals have been constructed with the true coefficients  $\boldsymbol{\beta}$ , so does not translate into a direct recommendation for the feasible estimator, it still suggests that the strong *ad hoc* adjustment in (17.40) is unwarranted.

This (crude) analysis suggests that for the cluster robust covariance estimator for fixed effects regression, the adjustment recommended by C. Hansen (17.39) may be the most appropriate, and is typically well approximated by the unadjusted estimator (17.38). Based on current theory, there is no justification for the *ad hoc* adjustment (17.40). The main argument for the latter is that it produces the largest standard errors, and is thus the most conservative choice.

In current practice the estimators (17.38) and (17.40) are the most commonly used covariance matrix estimators for fixed effects estimation.

In Sections 17.22 and 17.23 we discuss covariance matrix estimation under heteroskedasticity but no serial correlation.

To illustrate, in Table 17.2 we present the fixed effect regression estimates of the investment model (17.3) in the third column with cluster-robust standard errors. The trading indicator  $T_i$  and the industry dummies cannot be included as they are time-invariant. The point estimates are similar to the random effects estimates, though the coefficients on debt and cash flow increase in magnitude.

### 17.13 Fixed Effects Estimation in Stata

There are several methods to obtain the fixed effects estimator  $\hat{\boldsymbol{\beta}}_{\text{fe}}$  in Stata.

The first method is to use full dummy variable regression, which can be obtained using the Stata `regress` command, for example `reg y x i.id, cluster(id)` where `id` is the group (individual) identifier. In most cases, as discussed in Section 17.11, this is not recommended due to the excessive computer memory requirements and slow computation.

The second method is to manually create the within transformed variables as described in Section 17.8, and then use `regress`.

The third method is `xtreg fe`, which is specifically written for panel data. This estimates the slope coefficients using the partialling out approach. The default covariance matrix estimator is classical, as defined in (17.36). The cluster-robust covariance matrix (17.38) can be obtained using the option `vce(robust)`, or simply `r`.

The fourth method is `areg absorb(id)`, where `id` is the group (individual) identifier. This command is more general than panel data, also implementing the partialling out regression estimator. The

default covariance matrix estimator is the classical (17.36). The cluster-robust covariance matrix estimator (17.40) can be obtained using the `cluster(id)` option. The heteroskedasticity-robust covariance matrix is obtained when `r` or `vce(robust)` is specified, but this is not recommended unless  $T_i$  is large, as will be discussed in Section 17.22.

An important difference between the Stata `xtreg` and `areg` commands is that they implement different cluster-robust covariance matrix estimators: (17.38) in the case of `xtreg`, and (17.40) in the case of `areg`. As discussed in the previous section, the adjustment used by `areg` is *ad hoc* and not well-justified, but produces the largest and hence most conservative standard errors.

In current econometric practice, both `xtreg` and `areg` are used, though `areg` appears to be the more popular choice.

## 17.14 Between Estimator

The **between estimator** is calculated from the individual-mean equation (17.20)

$$\bar{y}_i = \bar{\mathbf{x}}_i' \boldsymbol{\beta} + u_i + \bar{\varepsilon}_i. \quad (17.41)$$

Estimation can be done at the level of individuals or at the level of observations. Least squares applied to (17.41) at the level of the  $N$  individuals is

$$\hat{\boldsymbol{\beta}}_{\text{be}} = \left( \sum_{i=1}^N \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i' \right)^{-1} \left( \sum_{i=1}^N \bar{\mathbf{x}}_i \bar{y}_i \right).$$

Least squares applied to (17.41) at the level of observations is

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_{\text{be}} &= \left( \sum_{i=1}^N \sum_{t \in S_i} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i' \right)^{-1} \left( \sum_{i=1}^N \sum_{t \in S_i} \bar{\mathbf{x}}_i \bar{y}_i \right) \\ &= \left( \sum_{i=1}^N T_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i' \right)^{-1} \left( \sum_{i=1}^N T_i \bar{\mathbf{x}}_i \bar{y}_i \right). \end{aligned}$$

In balanced panels  $\tilde{\boldsymbol{\beta}}_{\text{be}} = \hat{\boldsymbol{\beta}}_{\text{be}}$  but they differ on unbalanced panels.  $\tilde{\boldsymbol{\beta}}_{\text{be}}$  equals weighted least squares applied at the level of individuals with weight  $T_i$ .

Under the random effects assumptions (Assumption 17.1),  $\hat{\boldsymbol{\beta}}_{\text{be}}$  is unbiased for  $\boldsymbol{\beta}$  and has variance

$$\begin{aligned} V_{\text{be}} &= \text{var}(\hat{\boldsymbol{\beta}}_{\text{be}} | \mathbf{X}) \\ &= \left( \sum_{i=1}^N \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i' \right)^{-1} \left( \sum_{i=1}^N \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i' \sigma_u^2 \right) \left( \sum_{i=1}^N \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i' \right)^{-1} \end{aligned}$$

where

$$\sigma_i^2 = \text{var}(u_i + \bar{\varepsilon}_i) = \sigma_u^2 + \frac{\sigma_\varepsilon^2}{T_i}$$

is the variance of the error in (17.41). When the panel is balanced the variance formula simplifies to

$$V_{\text{be}} = \text{var}(\hat{\boldsymbol{\beta}}_{\text{be}} | \mathbf{X}) = \left( \sum_{i=1}^N \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i' \right)^{-1} \left( \sigma_u^2 + \frac{\sigma_\varepsilon^2}{T} \right).$$

Under the random effects assumption the between estimator  $\hat{\boldsymbol{\beta}}_{\text{be}}$  is unbiased for  $\boldsymbol{\beta}$  but is less efficient than the random effects estimator  $\hat{\boldsymbol{\beta}}_{\text{gls}}$ . Consequently there seems little direct use for the between estimator in linear panel data applications.

Instead, its primary application is to construct an estimate of  $\sigma_u^2$ . First, consider estimation of

$$\begin{aligned}\sigma_b^2 &= \frac{1}{N} \sum_{i=1}^N \sigma_i^2 \\ &= \sigma_u^2 + \frac{1}{N} \sum_{i=1}^N \frac{\sigma_\epsilon^2}{T_i} \\ &= \sigma_u^2 + \frac{\sigma_\epsilon^2}{\bar{T}}\end{aligned}$$

where  $\bar{T} = N / \sum_{i=1}^N \frac{1}{T_i}$  is the harmonic mean of  $T_i$ . (In the case of a balanced panel  $\bar{T} = T$ .) A natural estimator of  $\sigma_b^2$  is

$$\hat{\sigma}_b^2 = \frac{1}{N-k} \sum_{i=1}^N \hat{e}_{bi}^2. \quad (17.42)$$

where  $\hat{e}_{bi} = \bar{y}_i - \bar{x}'_i \hat{\beta}_{be}$  are the between residuals. (Either  $\hat{\beta}_{be}$  or  $\tilde{\beta}_{be}$  can be used.)

From the relation  $\sigma_b^2 = \sigma_u^2 + \sigma_\epsilon^2 / \bar{T}$  and (17.42) we can deduce an estimator for  $\sigma_u^2$ . We have already described an estimator  $\hat{\sigma}_\epsilon^2$  for  $\sigma_\epsilon^2$  in (17.37) for the fixed effects model. Since the fixed effects model holds under weaker conditions than the random effects model,  $\hat{\sigma}_\epsilon^2$  is valid for the latter as well. This suggests the following estimator for  $\sigma_u^2$

$$\hat{\sigma}_u^2 = \hat{\sigma}_b^2 - \frac{\hat{\sigma}_\epsilon^2}{\bar{T}}. \quad (17.43)$$

To summarize, the fixed effect estimator is used for  $\hat{\sigma}_\epsilon^2$ , the between estimator for  $\hat{\sigma}_b^2$ , and  $\hat{\sigma}_u^2$  is constructed from the two.

It is possible for (17.43) to be negative. It is typical to use the constrained estimator

$$\hat{\sigma}_u^2 = \max \left[ 0, \hat{\sigma}_b^2 - \frac{\hat{\sigma}_\epsilon^2}{\bar{T}} \right]. \quad (17.44)$$

(17.44) is the most common estimator for  $\sigma_u^2$  in the random effects model.

The between estimator  $\hat{\beta}_{be}$  can be obtained using Stata command `xtreg be`. The estimator  $\tilde{\beta}_{be}$  can be obtained by `xtreg be wls`.

## 17.15 Feasible GLS

The random effects estimator can be written as

$$\begin{aligned}\hat{\beta}_{re} &= \left( \sum_{i=1}^N \mathbf{X}'_i \boldsymbol{\Omega}_i^{-1} \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{X}'_i \boldsymbol{\Omega}_i^{-1} \mathbf{y}_i \right) \\ &= \left( \sum_{i=1}^N \tilde{\mathbf{X}}'_i \tilde{\mathbf{X}}_i \right)^{-1} \left( \sum_{i=1}^N \tilde{\mathbf{X}}'_i \tilde{\mathbf{y}}_i \right)\end{aligned} \quad (17.45)$$

where  $\tilde{\mathbf{X}}_i = \boldsymbol{\Omega}_i^{-1/2} \mathbf{X}_i$  and  $\tilde{\mathbf{y}}_i = \boldsymbol{\Omega}_i^{-1/2} \mathbf{y}_i$ . It is instructive to study these transformations.

Define  $\mathbf{P}_i = \mathbf{1}_i (\mathbf{1}'_i \mathbf{1}_i)^{-1} \mathbf{1}'_i$  so that  $\mathbf{M}_i = \mathbf{I}_i - \mathbf{P}_i$ . Thus while  $\mathbf{M}_i$  is the within operator,  $\mathbf{P}_i$  can be called the individual-mean operator, since  $\mathbf{P}_i \mathbf{y}_i = \mathbf{1}_i \bar{y}_i$ . We can write

$$\begin{aligned}\boldsymbol{\Omega}_i &= \mathbf{I}_i + \mathbf{1}_i \mathbf{1}'_i \sigma_u^2 / \sigma_\epsilon^2 \\ &= \mathbf{I}_i + \frac{T_i \sigma_u^2}{\sigma_\epsilon^2} \mathbf{P}_i \\ &= \mathbf{M}_i + \rho_i^{-2} \mathbf{P}_i\end{aligned}$$

where

$$\rho_i = \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T_i \sigma_u^2}}. \quad (17.46)$$

Since the matrices  $\mathbf{M}_i$  and  $\mathbf{P}_i$  are idempotent and orthogonal, we find that

$$\boldsymbol{\Omega}_i^{-1} = \mathbf{M}_i + \rho_i^2 \mathbf{P}_i$$

and

$$\boldsymbol{\Omega}_i^{-1/2} = \mathbf{M}_i + \rho_i \mathbf{P}_i = \mathbf{I}_i - (1 - \rho_i) \mathbf{P}_i. \quad (17.47)$$

Therefore the transformation used by the GLS estimator is

$$\begin{aligned} \tilde{\mathbf{y}}_i &= (\mathbf{I}_i - (1 - \rho_i) \mathbf{P}_i) \mathbf{y}_i \\ &= \mathbf{y}_i - (1 - \rho_i) \mathbf{1}_i \bar{\mathbf{y}}_i, \end{aligned}$$

which is a partial within transformation.

The transformation as written depends on  $\rho_i$  which is unknown. It can be replaced by the estimate

$$\hat{\rho}_i = \frac{\hat{\sigma}_\varepsilon}{\sqrt{\hat{\sigma}_\varepsilon^2 + T_i \hat{\sigma}_u^2}} \quad (17.48)$$

where the estimators  $\hat{\sigma}_\varepsilon^2$  and  $\hat{\sigma}_u^2$  are given in (17.37) and (17.44). We thus obtain the feasible transformations

$$\tilde{\mathbf{y}}_i = \mathbf{y}_i - (1 - \hat{\rho}_i) \mathbf{1}_i \bar{\mathbf{y}}_i \quad (17.49)$$

and

$$\tilde{\mathbf{X}}_i = \mathbf{X}_i - (1 - \hat{\rho}_i) \mathbf{1}_i \bar{\mathbf{x}}'_i. \quad (17.50)$$

The feasible random effects estimator is (17.45) using (17.49) and (17.50).

In the previous section we noted that it is possible for  $\hat{\sigma}_u^2 = 0$ . In this case  $\hat{\rho}_i = 1$  and  $\hat{\beta}_{re} = \hat{\beta}_{pool}$ .

What this shows is the following. The random effects estimator (17.45) is least-squares applied to the transformed variables  $\tilde{\mathbf{X}}_i$  and  $\tilde{\mathbf{y}}_i$  defined in (17.50) and (17.49). When  $\hat{\rho}_i = 0$  these are the within transformations, so  $\tilde{\mathbf{X}}_i = \dot{\mathbf{X}}_i$ ,  $\tilde{\mathbf{y}}_i = \dot{\mathbf{y}}_i$ , and  $\hat{\beta}_{re} = \hat{\beta}_{fe}$  is the fixed effects estimator. When  $\hat{\rho}_i = 1$  the data are untransformed  $\tilde{\mathbf{X}}_i = \mathbf{X}_i$ ,  $\tilde{\mathbf{y}}_i = \mathbf{y}_i$ , and  $\hat{\beta}_{re} = \hat{\beta}_{pool}$  is the pooled estimator. In general,  $\tilde{\mathbf{X}}_i$  and  $\tilde{\mathbf{y}}_i$  can be viewed as partial within transformations.

Recalling the definition  $\hat{\rho}_i = \hat{\sigma}_\varepsilon / \sqrt{\hat{\sigma}_\varepsilon^2 + T_i \hat{\sigma}_u^2}$ , we see that when the idiosyncratic error variance  $\hat{\sigma}_\varepsilon^2$  is large relative to  $T_i \hat{\sigma}_u^2$  then  $\hat{\rho}_i \approx 1$  and  $\hat{\beta}_{re} \approx \hat{\beta}_{pool}$ . Thus when the variance estimates suggest that the individual effect is relatively small, the random effect estimator simplifies to the pooled estimator. On the other hand when the individual effect error variance  $\hat{\sigma}_u^2$  is large relative to  $\hat{\sigma}_\varepsilon^2$  then  $\hat{\rho}_i \approx 0$  and  $\hat{\beta}_{re} \approx \hat{\beta}_{fe}$ . Thus when the variance estimates suggest that the individual effect is relatively large, the random effect estimator is close to the fixed effects estimator.

## 17.16 Intercept in Fixed Effects Regression

The fixed effect estimator does not apply to any regressor which is time-invariant for all individuals. This includes an intercept. Yet some authors and packages (e.g. Amemiya (1971) and `xtreg` in Stata) report an intercept. To see how to construct an estimator of an intercept, take the components regression equation adding an explicit intercept

$$y_{it} = \alpha + \mathbf{x}'_{it} \boldsymbol{\beta} + u_i + \varepsilon_{it}.$$

We have already discussed estimation of  $\boldsymbol{\beta}$  by  $\hat{\beta}_{fe}$ . Replacing  $\boldsymbol{\beta}$  in this equation with  $\hat{\beta}_{fe}$  and then estimating  $\alpha$  by least-squares, we obtain

$$\hat{\alpha}_{fe} = \bar{y} - \bar{\mathbf{x}}' \hat{\beta}_{fe}$$

where  $\bar{y}$  and  $\bar{\mathbf{x}}$  are averages from the full sample. This is the estimator reported by `xtreg`.

It is unclear if  $\hat{\alpha}_{fe}$  is particularly useful. It may be best to ignore the reported intercepts and focus on the slope coefficients.

## 17.17 Estimation of Fixed Effects

For most applications researchers are interested in the coefficients  $\beta$ , not the fixed effects  $u_i$ . But in some cases the fixed effects themselves are interesting. This arises when we want to measure the distribution of  $u_i$  to understand its heterogeneity. It also arises in the context of prediction. As discussed in Section 17.11 the fixed effects estimate  $\hat{u}$  is obtained by least-squares applied to the regression (17.33). To find their solution, replace  $\beta$  in (17.33) with the least squares minimizer  $\hat{\beta}_{\text{fe}}$  and apply least-squares. Since this is the individual-specific intercept, the solution is

$$\hat{u}_i = \frac{1}{T_i} \sum_{t=1}^N (y_{it} - \bar{x}'_i \hat{\beta}_{\text{fe}}) = \bar{y}_i - \bar{x}'_i \hat{\beta}_{\text{fe}}. \quad (17.51)$$

Alternatively, using (17.34), this is

$$\begin{aligned} \hat{u} &= (\mathbf{D}' \mathbf{D})^{-1} \mathbf{D}' (\mathbf{y} - \mathbf{X} \hat{\beta}_{\text{fe}}) \\ &= \text{diag}\{T_i^{-1}\} \sum_{i=1}^N d_i \mathbf{1}'_i (\mathbf{y}_i - \mathbf{X}_i \hat{\beta}_{\text{fe}}) \\ &= \sum_{i=1}^N d_i (\bar{y}_i - \bar{x}'_i \hat{\beta}_{\text{fe}}) \\ &= (\hat{u}_1, \dots, \hat{u}_N)'. \end{aligned}$$

Thus the least-squares estimates of the fixed effects can be obtained from the individual-specific means, and does not require a regression with  $N + k$  regressors.

If an intercept has been estimated (as discussed in the previous section) it should be subtracted from (17.51). In this case the estimated fixed effects are

$$\hat{u}_i = \bar{y}_i - \bar{x}'_i \hat{\beta}_{\text{fe}} - \hat{\alpha}_{\text{fe}}. \quad (17.52)$$

With either estimator, when the number of time series observations  $T_i$  is small,  $\hat{u}_i$  will be an imprecise estimator of  $u_i$ . Thus calculations based on  $\hat{u}_i$  should be interpreted cautiously.

The fixed effects (17.52) may be obtained in Stata after `ivreg`, `fe` using the `predict u` command, or after `areg` using the `predict d` command.

## 17.18 GMM Interpretation of Fixed Effects

We can also interpret the fixed effects estimator through the generalized method of moments.

Take the fixed effects model after applying the within transformation (17.21). We can view this as a system of  $T$  equations, one for each time period  $t$ . This is a multivariate regression model. Using the notation of Chapter 11, define the  $T \times kT$  regressor matrix

$$\bar{\mathbf{X}}_i = \begin{pmatrix} \dot{\mathbf{x}}'_{i1} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \dot{\mathbf{x}}'_{i2} & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \dot{\mathbf{x}}'_{iT} \end{pmatrix}. \quad (17.53)$$

If we treat each time period as a separate equation, we have the  $kT$  moment conditions

$$\mathbb{E}(\bar{\mathbf{X}}'_i (\dot{\mathbf{y}}_i - \dot{\mathbf{X}}'_i \beta)) = 0.$$

This is an overidentified system of equations when  $T \geq 3$  as there are  $k$  coefficients and  $kT$  moments. (However, the moments are collinear due to the within transformation. There are  $k(T-1)$  effective moments.) Interpreting this model in the context of multivariate regression, overidentification is achieved by the restriction that the coefficient vector  $\beta$  is constant across time periods.

This model can be interpreted as a regression of  $\dot{\mathbf{y}}_i$  on  $\dot{\mathbf{X}}_i$  using the instruments  $\bar{\mathbf{X}}_i$ . The 2SLS estimator, using matrix notation, is

$$\hat{\boldsymbol{\beta}} = \left( (\dot{\mathbf{X}}' \bar{\mathbf{X}}) (\bar{\mathbf{X}}' \bar{\mathbf{X}})^{-1} (\bar{\mathbf{X}}' \dot{\mathbf{X}}) \right)^{-1} \left( (\dot{\mathbf{X}}' \bar{\mathbf{X}}) (\bar{\mathbf{X}}' \bar{\mathbf{X}})^{-1} (\bar{\mathbf{X}}' \dot{\mathbf{y}}) \right).$$

Notice that

$$\begin{aligned} \bar{\mathbf{X}}' \bar{\mathbf{X}} &= \sum_{i=1}^n \begin{pmatrix} \dot{\mathbf{x}}_{i1} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \dot{\mathbf{x}}_{i2} & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \dot{\mathbf{x}}_{iT} \end{pmatrix} \begin{pmatrix} \dot{\mathbf{x}}'_{i1} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \dot{\mathbf{x}}'_{i2} & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \dot{\mathbf{x}}'_{iT} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^n \dot{\mathbf{x}}_{i1} \dot{\mathbf{x}}'_{i1} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \sum_{i=1}^n \dot{\mathbf{x}}_{i2} \dot{\mathbf{x}}'_{i2} & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \sum_{i=1}^n \dot{\mathbf{x}}_{iT} \dot{\mathbf{x}}'_{iT} \end{pmatrix}, \\ \bar{\mathbf{X}}' \dot{\mathbf{X}} &= \begin{pmatrix} \sum_{i=1}^n \dot{\mathbf{x}}_{i1} \dot{\mathbf{x}}'_{i1} \\ \vdots \\ \sum_{i=1}^n \dot{\mathbf{x}}_{iT} \dot{\mathbf{x}}'_{iT} \end{pmatrix}, \end{aligned}$$

and

$$\bar{\mathbf{X}}' \dot{\mathbf{y}} = \begin{pmatrix} \sum_{i=1}^n \dot{\mathbf{x}}_{i1} \dot{y}_{i1} \\ \vdots \\ \sum_{i=1}^n \dot{\mathbf{x}}_{iT} \dot{y}_{iT} \end{pmatrix}.$$

Thus the 2SLS estimator simplifies to

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{2sls}} &= \left( \sum_{t=1}^T \left( \sum_{i=1}^n \dot{\mathbf{x}}_{it} \dot{\mathbf{x}}'_{it} \right) \left( \sum_{i=1}^n \dot{\mathbf{x}}_{it} \dot{\mathbf{x}}'_{it} \right)^{-1} \left( \sum_{i=1}^n \dot{\mathbf{x}}_{it} \dot{\mathbf{x}}'_{it} \right) \right)^{-1} \\ &\quad \cdot \left( \sum_{t=1}^T \left( \sum_{i=1}^n \dot{\mathbf{x}}_{it} \dot{\mathbf{x}}'_{it} \right) \left( \sum_{i=1}^n \dot{\mathbf{x}}_{it} \dot{\mathbf{x}}'_{it} \right)^{-1} \left( \sum_{i=1}^n \dot{\mathbf{x}}_{it} \dot{y}_{it} \right) \right) \\ &= \left( \sum_{t=1}^T \sum_{i=1}^n \dot{\mathbf{x}}_{it} \dot{\mathbf{x}}'_{it} \right)^{-1} \left( \sum_{t=1}^T \sum_{i=1}^n \dot{\mathbf{x}}_{it} \dot{y}_{it} \right) \\ &= \hat{\boldsymbol{\beta}}_{\text{fe}} \end{aligned}$$

the fixed effects estimator!

This shows that if we treat each time period as a separate equation with its separate moment equation so that the system is over-identified, and then estimate by GMM using the 2SLS weight matrix, the resulting GMM estimator equals the simple fixed effects estimator. There is no change by adding the additional moment conditions.

The 2SLS estimator is the appropriate GMM estimator when the equation error is serially uncorrelated and homoskedastic. If we use a two-step efficient weight matrix which allows for heteroskedasticity and serial correlation the GMM estimator is

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{gmm}} &= \left( \sum_{t=1}^T \left( \sum_{i=1}^n \dot{\mathbf{x}}_{it} \dot{\mathbf{x}}'_{it} \right) \left( \sum_{i=1}^n \dot{\mathbf{x}}_{it} \dot{\mathbf{x}}'_{it} \hat{\sigma}_{it}^2 \right)^{-1} \left( \sum_{i=1}^n \dot{\mathbf{x}}_{it} \dot{\mathbf{x}}'_{it} \right) \right)^{-1} \\ &\quad \cdot \left( \sum_{t=1}^T \left( \sum_{i=1}^n \dot{\mathbf{x}}_{it} \dot{\mathbf{x}}'_{it} \right) \left( \sum_{i=1}^n \dot{\mathbf{x}}_{it} \dot{\mathbf{x}}'_{it} \hat{\sigma}_{it}^2 \right)^{-1} \left( \sum_{i=1}^n \dot{\mathbf{x}}_{it} \dot{y}_{it} \right) \right) \end{aligned}$$

where  $\hat{\sigma}_{it}$  are the fixed effects residuals.

Notationally, this GMM estimator has been written for a balanced panel. For an unbalanced panel the sums over  $i$  need to be replaced by sums over individuals observed during time period  $t$ . Otherwise no changes need to be made.

## 17.19 Identification in the Fixed Effects Model

The identification of the slope coefficient  $\beta$  in fixed effects regression is similar to that in conventional regression but somewhat more nuanced.

It is most useful to consider the within-transformed equation, which can be written as

$$\dot{y}_{it} = \dot{x}'_{it} \beta + \dot{\varepsilon}_{it}$$

or

$$\dot{y}_i = \dot{X}'_i \beta + \dot{\varepsilon}_i.$$

From regression theory we know that the coefficient  $\beta$  is the linear effect of  $\dot{x}_{it}$  on  $\dot{y}_{it}$ . The variable  $\dot{x}_{it}$  is the deviation of the regressor from its individual-specific mean, and similarly for  $\dot{y}_{it}$ . Thus the fixed effects model does not identify the effect of the average level of  $x_{it}$  on the average level of  $y_{it}$ , but rather the effect of the deviations in  $x_{it}$  on  $y_{it}$ .

In any given sample the fixed effects estimator is only defined if  $\sum_{i=1}^N \dot{X}'_i \dot{X}_i$  is full rank. The population analog (when individuals are i.i.d.) is

$$\mathbb{E}(\dot{X}'_i \dot{X}_i) > 0. \quad (17.54)$$

In the case of a balanced panel we can write this as

$$\sum_{t=1}^T \mathbb{E}(\dot{x}_{it} \dot{x}'_{it}) > 0.$$

Equation (17.54) is the identification condition for the fixed effects estimator. It requires that the regressor matrix is full-rank in expectation after application of the within transformation. Thus the regressors cannot contain any variable which does not have time-variation at the individual level, nor a set of regressors whose time-variation at the individual level is collinear.

## 17.20 Asymptotic Distribution of Fixed Effects Estimator

In this section we present an asymptotic distribution theory for the fixed effects estimator in the case of a balanced panel. The case of unbalanced panels is considered in the following section.

We use the following assumptions.

### Assumption 17.2

1.  $y_{it} = x'_{it} \beta + u_i + \varepsilon_{it}$  for  $i = 1, \dots, N$  and  $t = 1, \dots, T$  with  $T \geq 2$ .
2. The variables  $(\varepsilon_i, X_i)$ ,  $i = 1, \dots, N$ , are independent and identically distributed.
3.  $\mathbb{E}(x_{is} \varepsilon_{it}) = 0$  for all  $s = 1, \dots, T$ .
4.  $\mathbf{Q}_T = \mathbb{E}(\dot{X}'_i \dot{X}_i) > 0$ .
5.  $\mathbb{E}(\varepsilon_{it}^4) < \infty$ .
6.  $\mathbb{E}\|x_{it}\|^4 < \infty$ .

Given Assumption 17.2 we can establish asymptotic normality for  $\hat{\beta}_{fe}$ .

**Theorem 17.2** Under Assumption 17.2, as  $N \rightarrow \infty$ ,

$$\sqrt{N}(\hat{\beta}_{\text{fe}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, V_{\boldsymbol{\beta}})$$

where

$$V_{\boldsymbol{\beta}} = \mathbf{Q}_T^{-1} \boldsymbol{\Omega}_T \mathbf{Q}_T^{-1}$$

$$\boldsymbol{\Omega}_T = \mathbb{E}(\dot{\mathbf{X}}_i' \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i' \dot{\mathbf{X}}_i).$$

This asymptotic distribution is derived as the number of individuals  $N$  diverges to infinity while the time number of time periods  $T$  is held fixed. Therefore the normalization is  $\sqrt{N}$  rather than  $\sqrt{n}$  (though either could be used since  $T$  is fixed). This approximation is appropriate for the context of a large number of individuals  $i$ . We could alternatively derive an approximation for the case where both  $N$  and  $T$  diverge to infinity, but this would not be a stronger result. One way of thinking about this is that Theorem 17.2 does not require  $T$  to be large.

Theorem 17.2 may appear quite standard given our arsenal of asymptotic theory but in a fundamental sense it is quite different from any other result we have introduced. Fixed effects regression is effectively estimating  $N + k$  coefficients – the  $k$  slope coefficients  $\boldsymbol{\beta}$  plus the  $N$  fixed effects  $\mathbf{u}$  – and the theory specifies that  $N \rightarrow \infty$ . Thus the number of estimated parameters is diverging to infinity at the same rate as sample size, yet the estimator obtains a conventional mean-zero sandwich-form asymptotic distribution. In this sense Theorem 17.2 is quite new and special.

We now discuss the assumptions.

Assumption 17.2.2 states that the observations are independent across individuals  $i$ . This is commonly used for panel data asymptotic theory. An important implied restriction is that it means that we exclude from the regressors any serially correlated aggregate time series variation.

Assumption 17.2.3 imposes that  $\mathbf{x}_{it}$  is strictly exogenous for  $\varepsilon_{it}$ . This is stronger than simple projection, but is weaker than strict mean independence (17.18). It does not impose any condition on the individual-specific effects  $u_i$ .

Assumption 17.2.4 is the identification condition discussed in the previous section.

Assumptions 17.2.5 and 17.2.6 are needed for the central limit theorem.

We now prove Theorem 17.2. The assumptions imply that the variables  $(\dot{\mathbf{X}}_i, \boldsymbol{\epsilon}_i)$  are i.i.d. across  $i$  and have finite fourth moments. Thus by the WLLN

$$\frac{1}{N} \sum_{i=1}^N \dot{\mathbf{X}}_i' \dot{\mathbf{X}}_i \xrightarrow{p} \mathbb{E}(\dot{\mathbf{X}}_i' \dot{\mathbf{X}}_i) = \mathbf{Q}_T.$$

The random vectors  $\dot{\mathbf{X}}_i' \boldsymbol{\epsilon}_i$  are i.i.d. Assumption 17.2.3 implies

$$\mathbb{E}(\dot{\mathbf{X}}_i' \boldsymbol{\epsilon}_i) = \sum_{t=1}^T \mathbb{E}(\dot{\mathbf{x}}_{it} \varepsilon_{it}) = \sum_{t=1}^T \mathbb{E}(\mathbf{x}_{it} \varepsilon_{it}) - \sum_{t=1}^T \sum_{j=1}^T \mathbb{E}(\mathbf{x}_{ij} \varepsilon_{it}) = 0$$

so they are mean zero. Assumptions 17.2.5 and 17.2.6 imply that  $\dot{\mathbf{X}}_i' \boldsymbol{\epsilon}_i$  has a finite covariance matrix, which is  $\boldsymbol{\Omega}_T$ . The assumptions for the CLT (Theorem 6.11) hold, thus

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N \dot{\mathbf{X}}_i' \boldsymbol{\epsilon}_i \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega}_T).$$

Together we find

$$\sqrt{N}(\hat{\beta}_{\text{fe}} - \boldsymbol{\beta}) = \left( \frac{1}{N} \sum_{i=1}^N \dot{\mathbf{X}}_i' \dot{\mathbf{X}}_i \right)^{-1} \left( \frac{1}{N} \sum_{i=1}^N \dot{\mathbf{X}}_i' \boldsymbol{\epsilon}_i \right) \xrightarrow{d} \mathbf{Q}_T^{-1} N(\mathbf{0}, \boldsymbol{\Omega}_T) = N(\mathbf{0}, V_{\boldsymbol{\beta}})$$

as stated.

## 17.21 Asymptotic Distribution for Unbalanced Panels

In this section we extend the theory of the previous section to cover the case of unbalanced panels under random selection. Our presentation is built on Section 17.1 of Wooldridge (2010).

The key is to think of an unbalanced panel as a shortened version of an idealized balanced panel, where the shortening is due to “missing” observations due to random selection. Thus suppose that the underlying (potentially latent) variables are  $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})'$  and  $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{X}_{iT})'$ . Let  $\mathbf{s}_i = (s_{i1}, \dots, s_{iT})'$  be a vector of selection indicators, meaning that  $s_{it} = 1$  if the time period  $t$  is observed for individual  $i$ , and  $s_{it} = 0$  otherwise. Then, algebraically, we can describe the estimators on the observed sample as follows.

Let  $\mathbf{S}_i = \text{diag}(\mathbf{s}_i)$  and  $\mathbf{M}_i = \mathbf{S}_i - \mathbf{s}_i(\mathbf{s}'_i \mathbf{s}_i)^{-1} \mathbf{s}'_i$ , which is an idempotent matrix. The within transformations can be written as  $\hat{\mathbf{y}}_i = \mathbf{M}_i \mathbf{y}_i$  and  $\hat{\mathbf{X}}_i = \mathbf{M}_i \mathbf{X}_i$ . They have the property that if the  $s_{it} = 0$  (so that time period  $t$  is missing) then the  $t^{\text{th}}$  element of  $\hat{\mathbf{y}}_i$  and the  $t^{\text{th}}$  row of  $\hat{\mathbf{X}}_i$  are all zeros. Thus the missing observations have been replaced by zeros. Consequently, they do not appear in matrix products and sums.

The fixed effects estimator of  $\boldsymbol{\beta}$  based on the observed sample is

$$\hat{\boldsymbol{\beta}}_{\text{fe}} = \left( \sum_{i=1}^N \hat{\mathbf{X}}'_i \hat{\mathbf{X}}_i \right)^{-1} \left( \sum_{i=1}^N \hat{\mathbf{X}}'_i \hat{\mathbf{y}}_i \right).$$

Centered and normalized,

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\text{fe}} - \boldsymbol{\beta}) = \left( \frac{1}{N} \sum_{i=1}^N \hat{\mathbf{X}}'_i \hat{\mathbf{X}}_i \right)^{-1} \left( \frac{1}{N} \sum_{i=1}^N \hat{\mathbf{X}}'_i \boldsymbol{\epsilon}_i \right).$$

Notationally this appears to be identical to the case of a balanced panel, but the difference is that the within operator  $\mathbf{M}_i$  incorporates the sample selection induced by the unbalanced panel structure.

To derive a distribution theory for  $\hat{\boldsymbol{\beta}}_{\text{fe}}$  we need to be explicit about the stochastic nature of  $\mathbf{s}_i$ . That is, why are some time periods observed and some not? We can take several approaches:

1. We could treat  $\mathbf{s}_i$  as fixed (non-random). This is the easiest approach but the most unsatisfactory.
2. We could treat  $\mathbf{s}_i$  as random but independent of  $(\mathbf{y}_i, \mathbf{X}_i)$ . This is known as “missing at random” and is a common assumption used to justify methods with missing observations. It is justified when the reason why observations are not observed is independent of the observations. This is appropriate, for example, in panel data sets where individuals enter and exit in “waves”. The statistical treatment is not substantially different from the case of fixed  $\mathbf{s}_i$ .
3. We could treat  $(\mathbf{y}_i, \mathbf{X}_i, \mathbf{s}_i)$  as jointly random but impose a condition sufficient for consistent estimation of  $\boldsymbol{\beta}$ . This is the approach we take below. The condition turns out to be a form of mean independence. The advantage of this approach is that it is less restrictive than full independence. The disadvantage is that we must use a conditional mean restriction rather than uncorrelatedness to identify the coefficients.

The specific assumptions we impose are as follows.

**Assumption 17.3**

1.  $y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + u_i + \varepsilon_{it}$  for  $i = 1, \dots, N$  with  $T_i \geq 2$
2. The variables  $(\boldsymbol{\varepsilon}_i, \mathbf{X}_i, \mathbf{s}_i)$ ,  $i = 1, \dots, N$ , are independent and identically distributed.
3.  $\mathbb{E}(\varepsilon_{it} | \mathbf{X}_i, \mathbf{s}_i) = 0$ .
4.  $\mathbf{Q}_T = \mathbb{E}(\dot{\mathbf{X}}'_i \dot{\mathbf{X}}_i) > 0$ .
5.  $\mathbb{E}(\varepsilon_{it}^4) < \infty$ .
6.  $\mathbb{E}\|\mathbf{x}_{it}\|^4 < \infty$ .

The primary difference with Assumption 17.2 is that we have strengthened strict exogeneity to strict mean independence. This imposes that the regression model is properly specified, and that selection ( $\mathbf{s}_i$ ) does not affect the mean of  $\varepsilon_{it}$ . It is less restrictive than assuming full independence since  $\mathbf{s}_i$  can affect other moments of  $\varepsilon_{it}$ , and more importantly does not restrict the joint dependence between  $\mathbf{s}_i$  and  $\mathbf{X}_i$ .

Given the above development it is straightforward to establish asymptotic normality.

**Theorem 17.3** Under Assumption 17.3, as  $N \rightarrow \infty$ ,

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\text{fe}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}})$$

where

$$\begin{aligned} \mathbf{V}_{\boldsymbol{\beta}} &= \mathbf{Q}_T^{-1} \boldsymbol{\Omega}_T \mathbf{Q}_T^{-1} \\ \boldsymbol{\Omega}_T &= \mathbb{E}(\dot{\mathbf{X}}'_i \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}'_i \dot{\mathbf{X}}_i). \end{aligned}$$

We now prove Theorem 17.3. The assumptions imply that the variables  $(\dot{\mathbf{X}}_i, \boldsymbol{\varepsilon}_i)$  are i.i.d. across  $i$  and have finite fourth moments. By the WLLN

$$\frac{1}{N} \sum_{i=1}^N \dot{\mathbf{X}}'_i \dot{\mathbf{X}}_i \xrightarrow{p} \mathbb{E}(\dot{\mathbf{X}}'_i \dot{\mathbf{X}}_i) = \mathbf{Q}_T.$$

The random vectors  $\dot{\mathbf{X}}'_i \boldsymbol{\varepsilon}_i$  are i.i.d. The matrix  $\dot{\mathbf{X}}_i$  is a function of  $(\mathbf{X}_i, \mathbf{s}_i)$  only. Assumption 17.3.3 and the law of iterated expectations implies

$$\mathbb{E}(\dot{\mathbf{X}}'_i \boldsymbol{\varepsilon}_i) = \mathbb{E}(\dot{\mathbf{X}}'_i \mathbb{E}(\boldsymbol{\varepsilon}_i | \mathbf{X}_i, \mathbf{s}_i)) = 0.$$

so that  $\dot{\mathbf{X}}'_i \boldsymbol{\varepsilon}_i$  is mean zero. Assumptions 17.3.5 and 17.3.6 and the fact that  $\mathbf{s}_i$  is bounded implies that  $\dot{\mathbf{X}}'_i \boldsymbol{\varepsilon}_i$  has a finite covariance matrix, which is  $\boldsymbol{\Omega}_T$ . The assumptions for the CLT hold, thus

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N \dot{\mathbf{X}}'_i \boldsymbol{\varepsilon}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_T).$$

Together we obtain the stated result.

## 17.22 Heteroskedasticity-Robust Covariance Matrix Estimation

We have introduced two covariance matrix estimators for the fixed effects estimator. The classical estimator (17.36) is appropriate for the case where the idiosyncratic errors  $\varepsilon_{it}$  are homoskedastic and serially uncorrelated. The cluster-robust estimator (17.38) allows for heteroskedasticity and arbitrary serial correlation. In this and the following section we consider the intermediate case where  $\varepsilon_{it}$  is heteroskedastic but serially uncorrelated.

That is, assume that (17.18) and (17.26) hold but not necessarily (17.25). Define the conditional variances

$$\mathbb{E}(\varepsilon_{it}^2 | \mathbf{X}_i) = \sigma_{it}^2. \quad (17.55)$$

Then  $\Sigma_i = \mathbb{E}(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i' | \mathbf{X}_i) = \text{diag}(\sigma_{it}^2)$ . The covariance matrix (17.24) can be written as

$$\mathbf{V}_{\text{fe}} = (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1} \left( \sum_{i=1}^N \sum_{t \in S_i} \dot{\mathbf{x}}_{it} \dot{\mathbf{x}}_{it}' \sigma_{it}^2 \right) (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1}. \quad (17.56)$$

A natural estimator of  $\sigma_{it}^2$  is  $\hat{\varepsilon}_{it}^2$ . Replacing  $\sigma_{it}^2$  with  $\hat{\varepsilon}_{it}^2$  in (17.56) and making a degree-of-freedom adjustment we obtain a White-type covariance matrix estimator

$$\hat{\mathbf{V}}_{\text{fe}} = \frac{n}{n - N - k} (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1} \left( \sum_{i=1}^N \sum_{t \in S_i} \dot{\mathbf{x}}_{it} \dot{\mathbf{x}}_{it}' \hat{\varepsilon}_{it}^2 \right) (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1}.$$

Following the insight of White (1980) it may seem appropriate to expect  $\hat{\mathbf{V}}_{\text{fe}}$  to be a reasonable estimator of  $\mathbf{V}_{\text{fe}}$ . Unfortunately this is not the case, as discovered by Stock and Watson (2008). The problem is that  $\hat{\mathbf{V}}_{\text{fe}}$  is a function of the individual-specific means  $\bar{\varepsilon}_i$  which are negligible only if the number of time series observations  $T_i$  are large.

We can see this by a simple bias calculation. Assume that the sample is balanced and that the residuals are constructed with the true  $\boldsymbol{\beta}$ . Then

$$\hat{\varepsilon}_{it} = \dot{\varepsilon}_{it} = \varepsilon_{it} - \frac{1}{T} \sum_{t=1}^T \varepsilon_{ij}.$$

Using (17.26) and (17.55)

$$\mathbb{E}(\hat{\varepsilon}_{it}^2 | \mathbf{X}_i) = \left( \frac{T-2}{T} \right) \sigma_{it}^2 + \frac{\bar{\sigma}_i^2}{T} \quad (17.57)$$

where  $\bar{\sigma}_i^2 = T^{-1} \sum_{t=1}^T \sigma_{it}^2$ . (See Exercise 17.10.) Using (17.57) and setting  $k = 0$  we obtain

$$\begin{aligned} \mathbb{E}(\hat{\mathbf{V}}_{\text{fe}} | \mathbf{X}) &= \frac{T}{T-1} (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1} \left( \sum_{i=1}^N \sum_{t \in S_i} \dot{\mathbf{x}}_{it} \dot{\mathbf{x}}_{it}' \mathbb{E}(\hat{\varepsilon}_{it}^2 | \mathbf{X}_i) \right) (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1} \\ &= \left( \frac{T-2}{T-1} \right) \mathbf{V}_{\text{fe}} + \frac{1}{T-1} (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1} \left( \sum_{i=1}^N \dot{\mathbf{x}}_i' \dot{\mathbf{x}}_i \bar{\sigma}_i^2 \right) (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1}. \end{aligned}$$

Thus  $\hat{\mathbf{V}}_{\text{fe}}$  is a biased estimator for  $\mathbf{V}_{\text{fe}}$ , with a bias of order  $O(T^{-1})$ . Unless  $T \rightarrow \infty$ , this bias will persist as  $N \rightarrow \infty$ .

The estimator  $\hat{\mathbf{V}}_{\text{fe}}$  is unbiased in two contexts. The first is when the errors  $\varepsilon_{it}$  are homoskedastic. The second is when  $T = 2$ . (To show the latter requires some algebra.)

To correct the bias for the case  $T > 2$ , Stock and Watson (2008) proposed the estimator

$$\tilde{\mathbf{V}}_{\text{fe}} = \left( \frac{T-1}{T-2} \right) \hat{\mathbf{V}}_{\text{fe}} - \frac{1}{T-1} \hat{\mathbf{B}}_{\text{fe}} \quad (17.58)$$

$$\hat{\mathbf{B}}_{\text{fe}} = (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1} \left( \sum_{i=1}^N \dot{\mathbf{x}}_i' \dot{\mathbf{x}}_i \hat{\sigma}_i^2 \right) (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1}$$

$$\hat{\sigma}_i^2 = \frac{1}{T-1} \sum_{t=1}^T \hat{\varepsilon}_{it}^2. \quad (17.59)$$

You can check that  $\mathbb{E}(\hat{\sigma}_i^2 | \mathbf{X}_i) = \bar{\sigma}_i^2$  and  $\mathbb{E}(\tilde{\mathbf{V}}_{\text{fe}} | \mathbf{X}_i) = \mathbf{V}_{\text{fe}}$  so  $\tilde{\mathbf{V}}_{\text{fe}}$  is unbiased for  $\mathbf{V}_{\text{fe}}$ . (See Exercise 17.11.)

Stock and Watson (2008) show that  $\tilde{\mathbf{V}}_{\text{fe}}$  is consistent with  $T$  fixed and  $N \rightarrow \infty$ . In simulations they show that  $\tilde{\mathbf{V}}_{\text{fe}}$  has excellent performance.

Because of the Stock-Watson analysis, Stata no longer calculates the heteroskedasticity-robust covariance matrix estimator  $\hat{\mathbf{V}}_{\text{fe}}$  when the fixed effects estimator is calculated using the `xtreg` command. Instead, the cluster-robust estimator  $\hat{\mathbf{V}}_{\text{fe}}^{\text{cluster}}$  is reported when robust standard errors are requested. However, fixed effects is often implemented using the `areg` command, which will report the biased estimator  $\hat{\mathbf{V}}_{\text{fe}}$  if robust standard errors are requested. These leads to the practical recommendation that `areg` should typically be used with the `cluster(id)` option.

At present, the corrected estimator (17.58) has not been programmed as a Stata option.

## 17.23 Heteroskedasticity-Robust Estimation – Unbalanced Case

A limitation with the bias-corrected robust covariance matrix estimator of Stock and Watson (2008) is that it was only derived for balanced panels. In this section we generalize their estimator to cover the case of unbalanced panels.

The proposed estimator is

$$\begin{aligned}\tilde{\mathbf{V}}_{\text{fe}} &= (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1} \tilde{\boldsymbol{\Omega}}_{\text{fe}} (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1} \\ \tilde{\boldsymbol{\Omega}}_{\text{fe}} &= \sum_{i=1}^N \sum_{t \in S_i} \dot{\mathbf{x}}_i \dot{\mathbf{x}}_i' \left[ \left( \frac{T_i \hat{\varepsilon}_{it}^2 - \hat{\sigma}_i^2}{T_i - 2} \right) \mathbf{1}(T_i > 2) + \left( \frac{T_i \hat{\varepsilon}_{it}^2}{T_i - 1} \right) \mathbf{1}(T_i = 2) \right]\end{aligned}\quad (17.60)$$

where

$$\hat{\sigma}_i^2 = \frac{1}{T_i - 1} \sum_{t \in S_i} \hat{\varepsilon}_{it}^2.$$

To justify this estimator, as in the previous section make the simplifying assumption that the residuals are constructed with the true  $\beta$ . We calculate that

$$\mathbb{E}(\hat{\varepsilon}_{it}^2 | \mathbf{X}_i) = \left( \frac{T_i - 2}{T_i} \right) \sigma_{it}^2 + \frac{\bar{\sigma}_i^2}{T_i} \quad (17.61)$$

$$\mathbb{E}(\hat{\sigma}_i^2 | \mathbf{X}_i) = \bar{\sigma}_i^2. \quad (17.62)$$

You can show that under these assumptions,  $\mathbb{E}(\tilde{\mathbf{V}}_{\text{fe}} | \mathbf{X}) = \mathbf{V}_{\text{fe}}$  and thus  $\tilde{\mathbf{V}}_{\text{fe}}$  is unbiased for  $\mathbf{V}_{\text{fe}}$ . (See Exercise 17.12.)

The estimator  $\tilde{\mathbf{V}}_{\text{fe}}$  simplifies to the Stock-Watson estimator in the context of balanced panels and  $k = 0$ .

## 17.24 Hausman Test for Random vs Fixed Effects

The random effects model is a special case of the fixed effects model. Thus we can test the null hypothesis of random effects against the alternative of fixed effects. The Hausman test is typically used for this purpose. The statistic is a quadratic in the difference between the fixed effects and random effects estimators. The statistic is

$$\begin{aligned}H &= (\hat{\boldsymbol{\beta}}_{\text{fe}} - \hat{\boldsymbol{\beta}}_{\text{re}})' \widehat{\text{var}}(\hat{\boldsymbol{\beta}}_{\text{fe}} - \hat{\boldsymbol{\beta}}_{\text{re}})^{-1} (\hat{\boldsymbol{\beta}}_{\text{fe}} - \hat{\boldsymbol{\beta}}_{\text{re}}) \\ &= (\hat{\boldsymbol{\beta}}_{\text{fe}} - \hat{\boldsymbol{\beta}}_{\text{re}})' (\hat{\mathbf{V}}_{\text{fe}} - \hat{\mathbf{V}}_{\text{re}})^{-1} (\hat{\boldsymbol{\beta}}_{\text{fe}} - \hat{\boldsymbol{\beta}}_{\text{re}})\end{aligned}$$

where both  $\hat{\mathbf{V}}_{\text{fe}}$  and  $\hat{\mathbf{V}}_{\text{re}}$  take the classical (non-robust) form.

The test can be implemented on a subset of the coefficients  $\beta$ . In particular this needs to be done if the regressors  $\mathbf{x}_{it}$  contain time-invariant elements so that the random effects estimator contains more

coefficients than the fixed effects estimator. In this case the test should be implemented only on the coefficients on the time-varying regressors (and are thus estimated by both random and fixed effects).

An asymptotic  $100\alpha\%$  test rejects if  $H$  exceeds the  $1 - \alpha^{th}$  quantile of the  $\chi_k^2$  distribution, where  $k = \dim(\boldsymbol{\beta})$ . If the test rejects, this is evidence that the individual effect  $u_i$  is correlated with the regressors, so the random effects model is not appropriate. On the other hand if the test fails to reject, this evidence says that the random effects hypothesis cannot be rejected.

It is tempting to use the Hausman test to select whether to use the fixed effects or random effects estimator. One could imagine using the random effects estimator if the Hausman test fails to reject the random effects hypothesis, and using the fixed effects estimator if the Hausman test rejects random effects. This is not, however, a wise approach. This procedure – selecting an estimator based on a test – is known as a **pretest estimator** and is quite biased. The bias arises because the result of the test is random and correlated with the estimators.

Instead, the Hausman test can be used as a specification test. If you are planning to use the random effects estimator (and believe that the random effects assumptions are appropriate in your context), the Hausman test can be used to check this assumption and provide evidence to support your approach.

## 17.25 Random Effects or Fixed Effects?

We have presented the random effects and fixed effects estimators of the regression coefficients. Which should be used in practice? How should we view the difference?

The basic distinction is that the random effects estimator requires the individual error  $u_i$  to satisfy the conditional mean assumption (17.8). The fixed effect estimator does not require (17.8), and is robust to its violation. In particular, the individual effect  $u_i$  can be arbitrarily correlated with the regressors. On the other hand the random effect estimator is efficient under the random effects assumption (Assumption 17.1).

Current econometric practice is to prefer robustness over efficiency. Consequently current practice is (nearly uniformly) to use the fixed effects estimator for linear panel data models. Random effects estimators are only used in contexts where fixed effects estimation is unknown or challenging (which turns out to be the case in many nonlinear models).

The labels “random effects” and “fixed effects” are misleading. These are labels which arose in the early literature and we are stuck with them today. The term “fixed effects” was applied to  $u_i$  when it was viewed as an unobserved missing regressor in the era where regressors were viewed as “fixed”. Calling  $u_i$  “fixed” was equivalent to calling it a regressor. Today, we rarely refer to regressors as “fixed” when dealing with observational data. We view all variables as random. Consequently describing  $u_i$  as “fixed” does not make much sense, and it is hardly a contrast with the “random effect” label since under either assumption  $u_i$  is treated as random. Once again, the labels are unfortunate, but the key difference is whether  $u_i$  is correlated with the regressors.

## 17.26 Time Trends

In general we expect that economic agents will experience common shocks during the same time period. For example, business cycle fluctuations, inflation, and interest rates affect all agents in the economy. Therefore it is often desirable to include time effects in a panel regression model.

The simplest specification is a linear time trend

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \gamma t + u_i + \varepsilon_{it}.$$

For a introduction to time trends see Section 14.41. More flexible specifications (such as a quadratic) can also be used. For estimation, it is appropriate to include the time trend  $t$  as an element of the regressor vector  $\mathbf{x}_{it}$  and then apply fixed effects.

In some cases the time trends may be individual-specific. Series may be growing (or declining) at different rates. A linear time trend specification only extracts a common time trend. To allow for individual-specific time trends we need to include an interaction effect. This can be written as

$$y_{it} = \mathbf{x}'_{it} \boldsymbol{\beta} + \gamma_i t + u_i + \varepsilon_{it}.$$

In a fixed effects specification, the coefficients  $(\gamma_i, u_i)$  are treated as possibly correlated with the regressors. To eliminate them from the model we treat them as unknown parameters and estimate all by least squares. By the FWL theorem the estimator for  $\boldsymbol{\beta}$  equals

$$\hat{\boldsymbol{\beta}}_{FE} = (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1} (\dot{\mathbf{X}}' \dot{\mathbf{y}})$$

where the elements of  $\dot{\mathbf{X}}$  and  $\dot{\mathbf{Y}}$  are individual-level detrended observations. These are the residuals from the least-squares regressions

$$x_{it} = \hat{\alpha}_i + \hat{\gamma}_i t + \hat{x}_{it},$$

fit individual-by-individual, variable-by-variable.

## 17.27 Two-Way Error Components

In the previous section we discussed inclusion of time trends and individual-specific time trends. The functional forms imposed by linear time trends are very limiting. There is no economic reason to expect the “trend” of a series to be linear. If the “trend” is the business cycle it can alternate even in sign. This suggests that it is desirable to be more flexible than a simple linear (or quadratic) specification. In this section we consider the most flexible specification where the trend is allowed to take any arbitrary shape, but will require that it is common rather than individual-specific (otherwise it cannot be identified).

The model we consider is the **two-way error component model**

$$y_{it} = \mathbf{x}'_{it} \boldsymbol{\beta} + v_t + u_i + \varepsilon_{it}. \quad (17.63)$$

In this model,  $u_i$  is an unobserved individual-specific effect,  $v_t$  is an unobserved time-specific effect, and  $\varepsilon_{it}$  is an idiosyncratic error.

The two-way model (17.63) can be handled either using random effects or fixed effects. In a random effects framework, the errors  $v_t$  and  $u_i$  are modeled as in Assumption 17.1. When the panel is balanced and using matrix notation, the covariance matrix of the error vector  $\mathbf{e} = \mathbf{v} \otimes \mathbf{1}_N + \mathbf{1}_T \otimes \mathbf{u} + \boldsymbol{\varepsilon}$  is

$$\text{var}(\mathbf{e}) = \boldsymbol{\Omega} = (\mathbf{I}_T \otimes \mathbf{1}_N \mathbf{1}'_N) \sigma_v^2 + (\mathbf{1}_T \mathbf{1}'_T \otimes \mathbf{I}_N) \sigma_u^2 + \mathbf{I}_n \sigma_\varepsilon^2. \quad (17.64)$$

When the panel is unbalanced a similar but cumbersome expression for (17.64) can be derived. This variance (17.64) can be used for GLS estimation of  $\boldsymbol{\beta}$ .

More typically (17.63) is handled using fixed effects. The two-way within transformation subtracts both individual-specific means and time-specific means to eliminate both  $v_t$  and  $u_i$  from the two-way model (17.63). For a variable  $y_{it}$  we define the time-specific mean as follows. Let  $S_t$  be the set of individuals  $i$  for which the observation  $t$  is included in the sample, and let  $N_t$  be the number of these individuals. Then the time-specific mean at time  $t$  is

$$\tilde{y}_t = \frac{1}{N_t} \sum_{i \in S_t} y_{it}.$$

This is the average across all values of  $y_{it}$  observed at time  $t$ .

For the case of balanced panels the **two-way within transformation** is

$$\ddot{y}_{it} = y_{it} - \bar{y}_i - \tilde{y}_t + \bar{y} \quad (17.65)$$

where  $\bar{y} = n^{-1} \sum_{i=1}^N \sum_{t=1}^T y_{it}$  is the full-sample mean. If  $y_{it}$  satisfies the two-way component model

$$y_{it} = v_t + u_i + \varepsilon_{it}$$

then  $\bar{y}_i = \bar{v} + u_i + \bar{\varepsilon}_i$ ,  $\bar{y}_t = v_t + \bar{u} + \bar{\varepsilon}_t$  and  $\bar{y} = \bar{v} + \bar{u} + \bar{\varepsilon}$ . Hence

$$\begin{aligned}\bar{y}_{it} &= v_t + u_i + \varepsilon_{it} - (\bar{v} + u_i + \bar{\varepsilon}_i) - (v_t + \bar{u} + \bar{\varepsilon}_t) + \bar{v} + \bar{u} + \bar{\varepsilon} \\ &= \varepsilon_{it} - \bar{\varepsilon}_i - \bar{\varepsilon}_t + \bar{\varepsilon} \\ &= \ddot{\varepsilon}_{it}\end{aligned}$$

so the individual and time effects are eliminated.

The two-way within transformation applied to (17.63) yields

$$\ddot{y}_{it} = \ddot{\mathbf{x}}'_{it} \boldsymbol{\beta} + \ddot{\varepsilon}_{it} \quad (17.66)$$

which is invariant to both  $v_t$  and  $u_i$ . The **two-way within estimator** of  $\boldsymbol{\beta}$  is least-squares applied to (17.66).

For the unbalanced case there is a similar but algebraically more cumbersome two-way within transformation due to Wansbeek and Kapteyn (1989) and is described in Baltagi (2013, Section 9.4). We do not describe the algebra here as it is easier to implement using the technique described below.

If the two-way within estimator is used, then the regressors  $\mathbf{x}_{it}$  cannot include any time-invariant variables  $x_i$  or common time series variables  $x_t$ . Both are eliminated by the two-way within transformation. Thus coefficients are only identified for regressors which have variation both across individuals and across time.

Similarly to the one-way estimator, the two-way within estimator is equivalent to least squares estimation after including dummy variables for all individuals and for all time periods. Let  $\boldsymbol{\tau}_t$  be a set of  $T$  dummy variables where the  $t^{th}$  indicates the  $t^{th}$  time period. Thus the  $t^{th}$  element of  $\boldsymbol{\tau}_t$  is 1 and the remaining elements are zero. Set  $\boldsymbol{v} = (v_1, \dots, v_T)'$  as the vector of time fixed effects. Notice that  $v_t = \boldsymbol{\tau}'_t \boldsymbol{v}$ . Then we can write the two-way model as

$$y_{it} = \mathbf{x}'_{it} \boldsymbol{\beta} + \boldsymbol{\tau}'_t \boldsymbol{v} + \mathbf{d}'_i \mathbf{u} + \varepsilon_{it}. \quad (17.67)$$

This is the dummy variable representation of the two-way error components model.

The two-way dummy variable model (17.67) is collinear as written, for both the individual-specific dummies  $\mathbf{d}_i$  and the time-specific dummies  $\boldsymbol{\tau}_t$  span the intercept. Hence if (17.67) is to be estimated one dummy variable must be removed or otherwise normalized. The individual effects cannot be separately identified from the time effects.

Another way of thinking about (and estimating) the two-way fixed effects model is to write it as

$$y_{it} = \mathbf{x}'_{it} \boldsymbol{\beta} + \boldsymbol{\tau}'_t \boldsymbol{v} + u_i + \varepsilon_{it}. \quad (17.68)$$

This is a one-way fixed effects model with regressors  $\mathbf{x}_{it}$  and  $\boldsymbol{\tau}_t$ , with coefficient vectors  $\boldsymbol{\beta}$  and  $\boldsymbol{v}$ . This can be estimated by standard one-way fixed effects methods, including `xtreg` or `areg` in Stata. This produces estimates of the slopes  $\boldsymbol{\beta}$  as well as the time effects  $\boldsymbol{v}$ . Again to prevent singularity and achieve identification one time dummy variable is omitted from  $\boldsymbol{\tau}_t$  so the estimated time effects are all relative to this baseline time period. This is the most common method in practice to estimate a two-way fixed effects model.

If desired the relevance of the time effects can be tested by an exclusion test on the coefficients  $\boldsymbol{v}$ . If the test rejects the hypothesis of zero coefficients then this indicates that the time effects are relevant in the regression model.

The fixed effect estimator of (17.63) is invariant to the values of  $v_t$  and  $u_i$ , thus no assumptions need to be made concerning their stochastic properties.

To illustrate, the fourth column of Table 17.2 presents fixed effects estimates of the investment equation, augmented to included year dummy indicators, and is thus a two-way fixed effects model. In this example the coefficient estimates and standard errors are not greatly affected by the inclusion of the year dummy variables.

## 17.28 Instrumental Variables

Take the fixed effects model

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + u_i + \varepsilon_{it}. \quad (17.69)$$

We say  $\mathbf{x}_{it}$  is exogenous for  $\varepsilon_{it}$  if  $\mathbb{E}(\mathbf{x}_{it}\varepsilon_{it}) = 0$ , and we say  $\mathbf{x}_{it}$  is endogenous for  $\varepsilon_{it}$  if  $\mathbb{E}(\mathbf{x}_{it}\varepsilon_{it}) \neq 0$ . In Chapter 12 we discussed several economic examples of endogeneity, and the same issues apply in the panel data context. The primary difference is that in the fixed effects model, we only need to be concerned if the regressors are correlated with the idiosyncratic error  $\varepsilon_{it}$ , as correlation between  $\mathbf{x}_{it}$  and  $u_i$  is allowed.

As in Chapter 12 if the regressors are endogenous then the fixed effects estimator will be biased and inconsistent for the structural coefficient  $\boldsymbol{\beta}$ . The standard approach to handling endogeneity is to specify instrumental variables  $\mathbf{z}_{it}$  which are both relevant (correlated with  $\mathbf{x}_{it}$ ) yet exogenous (uncorrelated with  $\varepsilon_{it}$ ).

Let  $\mathbf{z}_{it}$  be an  $\ell \times 1$  instrumental variable where  $\ell \geq k$ . As in the cross-section case,  $\mathbf{z}_{it}$  may contain both included exogenous variables (variables in  $\mathbf{x}_{it}$  that are exogenous) and excluded exogenous variables (variables not in  $\mathbf{x}_{it}$ ). Let  $\mathbf{Z}_i$  be the stacked instruments for the individual  $i$ , and  $\mathbf{Z}$  be the stacked instruments for the full sample.

The dummy variable formulation of the fixed effects model is

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{d}'_i\mathbf{u} + \varepsilon_{it}$$

where  $\mathbf{d}_i$  is an  $N \times 1$  vector of dummy variables, one for each individual in the sample. The model in matrix notation for the full sample is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\mathbf{u} + \boldsymbol{\varepsilon}. \quad (17.70)$$

Theorem 17.1 shows that the fixed effects estimator for  $\boldsymbol{\beta}$  can be calculated by least squares estimation of (17.70). Thus the dummies  $\mathbf{D}$  should be viewed as an included exogenous variable.

Now consider 2SLS estimation of  $\boldsymbol{\beta}$  using the instruments  $\mathbf{Z}$  for  $\mathbf{X}$ . Since  $\mathbf{D}$  is an included exogenous variable it should also be used as an instrument. Thus 2SLS estimation of the fixed effects model (17.69) is algebraically 2SLS of the regression (17.70) of  $\mathbf{y}$  on  $(\mathbf{X}, \mathbf{D})$ , using the pair  $(\mathbf{Z}, \mathbf{D})$  as instruments.

Since the dimension of  $\mathbf{D}$  can be excessively large, as discussed in Section 17.11, it is advisable to use residual regression to compute the 2SLS estimator, as we now describe.

In Section 12.12, we described several alternative representations for the 2SLS estimator. The fifth (equation (12.34)) shows that the 2SLS estimator for  $\boldsymbol{\beta}$  can be written as

$$\hat{\boldsymbol{\beta}}_{2\text{sls}} = \left( \mathbf{X}' \mathbf{M}_D \mathbf{Z} (\mathbf{Z}' \mathbf{M}_D \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{M}_D \mathbf{X} \right)^{-1} \left( \mathbf{X}' \mathbf{M}_D \mathbf{Z} (\mathbf{Z}' \mathbf{M}_D \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{M}_D \mathbf{y} \right)$$

where  $\mathbf{M}_D = \mathbf{I}_n - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'$ . The latter is the matrix within operator, thus  $\mathbf{M}_D\mathbf{y} = \dot{\mathbf{y}}$ ,  $\mathbf{M}_D\mathbf{X} = \dot{\mathbf{X}}$ , and  $\mathbf{M}_D\mathbf{Z} = \dot{\mathbf{Z}}$ . It follows that the 2SLS estimator is

$$\hat{\boldsymbol{\beta}}_{2\text{sls}} = \left( \dot{\mathbf{X}}' \dot{\mathbf{Z}} (\dot{\mathbf{Z}}' \dot{\mathbf{Z}})^{-1} \dot{\mathbf{Z}}' \dot{\mathbf{X}} \right)^{-1} \left( \dot{\mathbf{X}}' \dot{\mathbf{Z}} (\dot{\mathbf{Z}}' \dot{\mathbf{Z}})^{-1} \dot{\mathbf{Z}}' \dot{\mathbf{y}} \right).$$

This is quite convenient. It shows that the 2SLS estimator for the fixed effects model can be calculated by applying the standard 2SLS formula to the within-transformed  $y_{it}$ ,  $\mathbf{x}_{it}$ , and  $\mathbf{z}_{it}$ . The 2SLS residuals are  $\hat{\mathbf{e}} = \dot{\mathbf{y}} - \dot{\mathbf{X}}\hat{\boldsymbol{\beta}}_{2\text{sls}}$ .

This estimator can be obtained using the Stata command `xtivreg fe`. It can also be obtained using the Stata command `ivregress` after making the within transformations.

The presentation above focused for clarity on the one-way fixed effects model. There is no substantial change in the two-way fixed effects model

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + u_i + v_t + \varepsilon_{it}.$$

The easiest way to estimate the two-way model is to add  $T - 1$  time-period dummies to the regression model, and include these dummy variables as both regressors and instruments.

## 17.29 Identification with Instrumental Variables

To understand the identification of the structural slope coefficient  $\beta$  in the fixed effects model it is necessary to examine the reduced form equation for the endogenous regressors  $x_{it}$ . This is

$$\mathbf{x}_{it} = \boldsymbol{\Gamma} \mathbf{z}_{it} + \mathbf{w}_i + \boldsymbol{\zeta}_{it}$$

where  $\mathbf{w}_i$  is a  $k \times 1$  vector of fixed effects for the  $k$  regressors and  $\boldsymbol{\zeta}_{it}$  is an idiosyncratic error.

The coefficient matrix  $\boldsymbol{\Gamma}$  is the linear effect of  $\mathbf{z}_{it}$  on  $\mathbf{x}_{it}$ , holding the fixed effects  $\mathbf{w}_i$  constant. Thus  $\boldsymbol{\Gamma}$  has a similar interpretation as the coefficient  $\beta$  in the fixed effects regression model. It is the effect of the variation in  $\mathbf{z}_{it}$  about its individual-specific mean on  $\mathbf{x}_{it}$ .

The 2SLS estimator is a function of the within transformed variables. Applying the within transformation to the reduced form we find

$$\dot{\mathbf{x}}_{it} = \boldsymbol{\Gamma} \dot{\mathbf{z}}_{it} + \dot{\boldsymbol{\zeta}}_{it}.$$

Again we see that  $\boldsymbol{\Gamma}$  is the effect of the within-transformed instruments on the within-transformed regressors. If there is no time-variation in the within-transformed instruments, or there is no correlation between the instruments and the regressors after removing the individual-specific means, then the coefficient  $\boldsymbol{\Gamma}$  will be either not identified or singular. In either case the structural coefficient  $\beta$  will not be identified.

Thus for identification of the fixed effects instrumental variables model we need

$$\mathbb{E}(\dot{\mathbf{Z}}_i' \dot{\mathbf{Z}}_i) > 0 \quad (17.71)$$

and

$$\text{rank}(\mathbb{E}(\dot{\mathbf{Z}}_i' \dot{\mathbf{X}}_i)) = k. \quad (17.72)$$

Condition (17.71) is the same as the condition for identification in fixed effects regression – the instruments must have full variation after the within transformation. Condition (17.72) is analogous to the relevance condition for identification of instrumental variable regression in the cross-section context, but applies to the within-transformed instruments and regressors.

Condition (17.72) shows that to examine instrument validity in the context of fixed effects 2SLS, it is important to estimate the reduced form equation using fixed effects (within) regression. Standard tests for instrument validity ( $F$  tests on the excluded instruments) can be applied. However, since the correlation structure of the reduced form equation is in general unknown, it is appropriate to use a cluster-robust covariance matrix, clustered at the level of the individual.

## 17.30 Asymptotic Distribution of Fixed Effects 2SLS Estimator

In this section we present an asymptotic distribution theory for the fixed effects estimator. We provide a formal theory for the case of balanced panels, and discuss an extension to the case of unbalanced panels.

We use the following assumptions for balanced panels.

**Assumption 17.4**

1.  $y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + u_i + \varepsilon_{it}$  for  $i = 1, \dots, N$  and  $t = 1, \dots, T$  with  $T \geq 2$ .
2. The variables  $(\boldsymbol{\varepsilon}_i, \mathbf{X}_i, \mathbf{Z}_i)$ ,  $i = 1, \dots, N$ , are independent and identically distributed.
3.  $\mathbb{E}(\mathbf{z}_{is}\varepsilon_{it}) = 0$  for all  $s = 1, \dots, T$ .
4.  $\mathbf{Q}_{zz} = \mathbb{E}(\dot{\mathbf{Z}}'_i \dot{\mathbf{Z}}_i) > 0$ .
5.  $\text{rank}(\mathbf{Q}_{zx}) = k$  where  $\mathbf{Q}_{zx} = \mathbb{E}(\dot{\mathbf{Z}}'_i \dot{\mathbf{X}}_i)$ .
6.  $\mathbb{E}(\varepsilon_{it}^4) < \infty$ .
7.  $\mathbb{E}\|\mathbf{x}_{it}\|^2 < \infty$ .
8.  $\mathbb{E}\|\mathbf{z}_{it}\|^4 < \infty$ .

Given Assumption 17.4 we can establish asymptotic normality for  $\hat{\boldsymbol{\beta}}_{2\text{sls}}$ .

**Theorem 17.4** Under Assumption 17.4, as  $N \rightarrow \infty$ ,

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{2\text{sls}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, V_{\boldsymbol{\beta}})$$

where

$$V_{\boldsymbol{\beta}} = (\mathbf{Q}'_{zx} \boldsymbol{\Omega}_{zz}^{-1} \mathbf{Q}_{zx})^{-1} (\mathbf{Q}'_{zx} \boldsymbol{\Omega}_{zz}^{-1} \boldsymbol{\Omega}_{ze} \boldsymbol{\Omega}_{zz}^{-1} \mathbf{Q}_{zx}) (\mathbf{Q}'_{zx} \boldsymbol{\Omega}_{zz}^{-1} \mathbf{Q}_{zx})^{-1}$$

$$\boldsymbol{\Omega}_{ze} = \mathbb{E}(\dot{\mathbf{Z}}'_i \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}'_i \dot{\mathbf{Z}}_i).$$

The proof of the result is similar to Theorem 17.2 so is omitted. The key orthogonality condition is Assumption 17.4.3, which states that the instruments are strictly exogenous for the idiosyncratic errors. The identification conditions are Assumptions 17.4.4 and 17.4.5, which were discussed in the previous section.

The theorem is stated for balanced panels. For unbalanced panels we can modify the theorem as in Theorem 17.3 by adding the selection indicators  $\mathbf{s}_i$ , and replacing Assumption 17.4.3 with  $\mathbb{E}(\varepsilon_{it} | \mathbf{Z}_i, \mathbf{s}_i) = 0$ , which states that the idiosyncratic errors are mean independent of the instruments and selection.

If the idiosyncratic errors  $\varepsilon_{it}$  are homoskedastic and serially uncorrelated, then the covariance matrix simplifies to

$$V_{\boldsymbol{\beta}} = (\mathbf{Q}'_{zx} \boldsymbol{\Omega}_{zz}^{-1} \mathbf{Q}_{zx})^{-1} \sigma_{\varepsilon}^2.$$

In this case a classical homoskedastic covariance matrix estimator can be used. Otherwise a cluster-robust covariance matrix estimator can be used, which takes the form

$$\widehat{V}_{\widehat{\boldsymbol{\beta}}} = \left( \dot{\mathbf{X}}' \dot{\mathbf{Z}} (\dot{\mathbf{Z}}' \dot{\mathbf{Z}})^{-1} \dot{\mathbf{Z}}' \dot{\mathbf{X}} \right)^{-1} \left( \dot{\mathbf{X}}' \dot{\mathbf{Z}} \right) \left( \dot{\mathbf{Z}}' \dot{\mathbf{Z}} \right)^{-1} \left( \sum_{i=1}^N \dot{\mathbf{Z}}'_i \widehat{\boldsymbol{\varepsilon}}_i \widehat{\boldsymbol{\varepsilon}}'_i \dot{\mathbf{Z}}_i \right)$$

$$\cdot \left( \dot{\mathbf{Z}}' \dot{\mathbf{Z}} \right)^{-1} \left( \dot{\mathbf{Z}}' \dot{\mathbf{X}} \right) \left( \dot{\mathbf{X}}' \dot{\mathbf{Z}} (\dot{\mathbf{Z}}' \dot{\mathbf{Z}})^{-1} \dot{\mathbf{Z}}' \dot{\mathbf{X}} \right)^{-1}.$$

As for the case of fixed effects regression, the heteroskedasticity-robust covariance matrix estimator is not recommended due to bias when  $T$  is small, and a bias-corrected version has not been developed.

The Stata command `xtivreg`, `f`e by default reports the classical homoskedastic covariance matrix estimator. To obtain a cluster-robust covariance matrix, use the option `vce(robust)` or `vce(cluster id)`.

## 17.31 Linear GMM

Consider the just-identified 2SLS estimator. It solves the equation

$$\dot{\mathbf{Z}}' (\dot{\mathbf{y}} - \dot{\mathbf{X}}\boldsymbol{\beta}) = 0$$

or equivalently

$$\dot{\mathbf{Z}}' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0.$$

These are sample analogs of the population moment condition

$$\mathbb{E}(\dot{\mathbf{Z}}'_i (\dot{\mathbf{y}}_i - \dot{\mathbf{X}}_i \boldsymbol{\beta})) = 0.$$

or equivalently

$$\mathbb{E}(\dot{\mathbf{Z}}'_i (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})) = 0.$$

These population conditions hold at the true  $\boldsymbol{\beta}$  since  $\dot{\mathbf{Z}}' \mathbf{u} = \mathbf{Z}' \mathbf{M} \mathbf{D} \mathbf{u} = 0$  since  $\mathbf{u}$  lies in the null space of  $\mathbf{D}$ , and  $\mathbb{E}(\dot{\mathbf{Z}}'_i \boldsymbol{\epsilon}) = 0$  is implied by Assumption 17.4.3.

The population orthogonality conditions hold in the overidentified case as well. In this case an alternative to 2SLS is GMM. Let  $\widehat{\Omega}_i$  be an estimate of

$$\mathbf{W} = \mathbb{E}(\dot{\mathbf{Z}}'_i \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}'_i \dot{\mathbf{Z}}_i),$$

for example

$$\widehat{\mathbf{W}} = \frac{1}{N} \sum_{i=1}^N \dot{\mathbf{Z}}'_i \widehat{\boldsymbol{\epsilon}}_i \widehat{\boldsymbol{\epsilon}}'_i \dot{\mathbf{Z}}_i \quad (17.73)$$

where  $\widehat{\boldsymbol{\epsilon}}_i$  are the 2SLS fixed effects residuals. The GMM fixed effects estimator is

$$\widehat{\boldsymbol{\beta}}_{\text{gmm}} = \left( \dot{\mathbf{X}}' \dot{\mathbf{Z}} \widehat{\mathbf{W}}^{-1} \dot{\mathbf{Z}}' \dot{\mathbf{X}} \right)^{-1} \left( \dot{\mathbf{X}}' \dot{\mathbf{Z}} \widehat{\mathbf{W}}^{-1} \dot{\mathbf{Z}}' \dot{\mathbf{y}} \right). \quad (17.74)$$

The estimator (17.74)-(17.73) does not have a Stata command, but can be obtained by generating the within transformed variables  $\dot{\mathbf{X}}$ ,  $\dot{\mathbf{Z}}$  and  $\dot{\mathbf{y}}$ , and then estimating by GMM a regression of  $\dot{\mathbf{y}}$  on  $\dot{\mathbf{X}}$  using  $\dot{\mathbf{Z}}$  as instruments, using a weight matrix clustered by individual.

## 17.32 Estimation with Time-Invariant Regressors

One of the disappointments with the fixed effects estimator is that it cannot estimate the effect of regressors which are time-invariant. They are not identified separately from the fixed effect, and are eliminated by the within transformation. In contrast, the random effects estimator allows for time-invariant regressors, but does so only by assuming strict exogeneity, which is stronger than typically desired in economic applications.

It turns out that we can consider an intermediate case which maintains the fixed effects assumptions for the time-varying regressors, but uses stronger assumptions on the time-invariant regressors. For our exposition we will denote the time-varying regressors by the  $k \times 1$  vector  $\mathbf{x}_{it}$ , and the time-invariant regressors by the  $\ell \times 1$  vector  $\mathbf{z}_i$ .

Consider the linear regression model

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_i\boldsymbol{\gamma} + u_i + \varepsilon_{it}.$$

At the level of the individual this can be written as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma} + \mathbf{u}_i + \boldsymbol{\varepsilon}_i$$

where  $\mathbf{Z}_i = \mathbf{I}_i \mathbf{z}'_i$ . For the full sample in matrix notation we can write this as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u} + \boldsymbol{\varepsilon}. \quad (17.75)$$

We will maintain the assumption that the idiosyncratic errors  $\varepsilon_{it}$  are uncorrelated with both  $\mathbf{x}_{it}$  and  $\mathbf{z}_i$  at all time horizons:

$$\mathbb{E}(\mathbf{x}_{is}\varepsilon_{it}) = 0 \quad (17.76)$$

$$\mathbb{E}(\mathbf{z}_i\varepsilon_{it}) = 0. \quad (17.77)$$

In this section we consider the case where  $\mathbf{z}_i$  is uncorrelated with the individual-level error  $u_i$ , thus

$$\mathbb{E}(\mathbf{z}_i u_i) = 0, \quad (17.78)$$

but the correlation of  $\mathbf{x}_{it}$  and  $u_i$  is left unrestricted. In this context we say that  $\mathbf{z}_i$  is exogenous with respect to the fixed effect  $u_i$ , while  $\mathbf{x}_{it}$  is endogenous with respect to  $u_i$ . Note that this is a different type of endogeneity than considered in the sections on instrumental variables: there we were concerned with correlation with the idiosyncratic error  $\varepsilon_{it}$ . Here we are concerned with correlation with the fixed effect  $u_i$ .

We consider estimation of (17.75) by instrumental variables, and thus need instruments which are uncorrelated with the error  $\mathbf{u} + \boldsymbol{\varepsilon}$ . The time-invariant regressors  $\mathbf{Z}$  satisfy this condition due to (17.77) and (17.78), thus

$$\mathbb{E}(\mathbf{Z}'_i (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\boldsymbol{\gamma})) = 0.$$

While the time-varying regressors  $\mathbf{X}$  are correlated with  $\mathbf{u}$ , the within transformed variables  $\dot{\mathbf{X}}$  are uncorrelated with  $\mathbf{u} + \boldsymbol{\varepsilon}$  under (17.76), thus

$$\mathbb{E}(\dot{\mathbf{X}}'_i (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\boldsymbol{\gamma})) = 0.$$

Therefore we can estimate  $(\boldsymbol{\beta}, \boldsymbol{\gamma})$  by instrumental variable regression, using the instrument set  $(\dot{\mathbf{X}}, \mathbf{Z})$ . That is, regression of  $\mathbf{y}$  on  $\mathbf{X}$  and  $\mathbf{Z}$ , treating  $\mathbf{X}$  as endogenous,  $\mathbf{Z}$  as exogenous, and using the instrument  $\dot{\mathbf{X}}$ . Write this estimator as  $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ . This can be implemented using the Stata ivregress command after constructing the within transformed  $\dot{\mathbf{X}}$ .

This instrumental variables estimator is algebraically equal to a simple two-step estimator. The first step  $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{fe}$  is the fixed effects estimator. The second step sets  $\hat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\hat{\mathbf{u}})$ , the least-squares coefficient from the regression of the estimated fixed effect  $\hat{\mathbf{u}}$  on  $\mathbf{Z}$ . To see this equivalence, observe that the instrumental variables estimator solves the sample moment equations

$$\dot{\mathbf{X}}' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma}) = 0 \quad (17.79)$$

$$\mathbf{Z}' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma}) = 0. \quad (17.80)$$

Notice that  $\dot{\mathbf{X}}'_i \mathbf{Z}_i = \dot{\mathbf{X}}'_i \mathbf{I}_i \mathbf{z}'_i = 0$  so  $\dot{\mathbf{X}}' \mathbf{Z} = 0$ . Thus (17.79) is the same as  $\dot{\mathbf{X}}' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$  whose solution is  $\hat{\boldsymbol{\beta}}_{fe}$ . Plugging this into the left-side of (17.80) we obtain

$$\begin{aligned} \mathbf{Z}' (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{fe} - \mathbf{Z}\boldsymbol{\gamma}) &= \mathbf{Z}' (\bar{\mathbf{y}} - \bar{\mathbf{X}}\hat{\boldsymbol{\beta}}_{fe} - \mathbf{Z}\boldsymbol{\gamma}) \\ &= \mathbf{Z}' (\hat{\mathbf{u}} - \mathbf{Z}\boldsymbol{\gamma}) \end{aligned}$$

where  $\bar{\mathbf{y}}$  and  $\bar{\mathbf{X}}$  are the stacked individual means  $\iota_i \bar{\mathbf{y}}_i$  and  $\iota_i \bar{\mathbf{x}}'_i$ . Set equal to 0 and solving we obtain the least-squares estimator  $\hat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\hat{\mathbf{u}})$  as claimed. This equivalence was first observed by Hausman and Taylor (1981).

For standard error calculation it is recommended to estimate  $(\boldsymbol{\beta}, \boldsymbol{\gamma})$  jointly by instrumental variable regression, using a cluster-robust covariance matrix estimator, clustered at the individual level. Classical and heteroskedasticity-robust estimators are misspecified due to the individual-specific effect  $u_i$ .

The estimator  $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$  is a special case of the Hausman-Taylor estimator described in the next section. (However, for an unknown reason the above estimator cannot be estimated using the Stata `xhtaylor` command.)

### 17.33 Hausman-Taylor Model

Hausman and Taylor (1981) consider a generalization of the model of the previous section. The model is

$$y_{it} = \mathbf{x}'_{1it}\boldsymbol{\beta}_1 + \mathbf{x}'_{2it}\boldsymbol{\beta}_2 + \mathbf{z}'_{1i}\boldsymbol{\gamma}_1 + \mathbf{z}'_{2i}\boldsymbol{\gamma}_2 + u_i + \varepsilon_{it}$$

where  $\mathbf{x}_{1it}$  and  $\mathbf{x}_{2it}$  are time-varying and  $\mathbf{z}_{1i}$  and  $\mathbf{z}_{2i}$  are time-invariant. Let the dimensions of  $\mathbf{x}_{1it}$ ,  $\mathbf{x}_{2it}$ ,  $\mathbf{z}_{1i}$ , and  $\mathbf{z}_{2i}$  be  $k_1$ ,  $k_2$ ,  $\ell_1$ , and  $\ell_2$ , respectively.

Write the model in matrix notation as

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{Z}_1\boldsymbol{\gamma}_1 + \mathbf{Z}_2\boldsymbol{\gamma}_2 + \mathbf{u} + \boldsymbol{\varepsilon}. \quad (17.81)$$

Let  $\bar{\mathbf{X}}_1$  and  $\bar{\mathbf{X}}_2$  denote conformable matrices of individual-specific means, and let  $\dot{\mathbf{X}}_1 = \mathbf{X}_1 - \bar{\mathbf{X}}_1$  and  $\dot{\mathbf{X}}_2 = \mathbf{X}_2 - \bar{\mathbf{X}}_2$  denote the within-transformed variables.

The Hausman-Taylor model assumes that all regressors are uncorrelated with the idiosyncratic error  $\varepsilon_{it}$  at all time horizons, and also that  $\mathbf{x}_{1it}$  and  $\mathbf{z}_{1i}$  are exogenous with respect to the fixed effect  $u_i$ , so that

$$\begin{aligned} \mathbb{E}(\mathbf{x}_{1it}u_i) &= 0 \\ \mathbb{E}(\mathbf{z}_{1i}u_i) &= 0. \end{aligned}$$

The regressors  $\mathbf{x}_{2it}$  and  $\mathbf{z}_{2i}$ , however, are allowed to be correlated with  $u_i$ .

Set  $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{Z}_1, \mathbf{Z}_2)$  and  $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ . The assumptions imply the following population moment conditions

$$\begin{aligned} \mathbb{E}(\dot{\mathbf{X}}'_1(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})) &= 0 \\ \mathbb{E}(\dot{\mathbf{X}}'_2(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})) &= 0 \\ \mathbb{E}(\bar{\mathbf{X}}'_1(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})) &= 0 \\ \mathbb{E}(\mathbf{Z}'_1(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})) &= 0. \end{aligned}$$

There are  $2k_1 + k_2 + \ell_1$  moment conditions and  $k_1 + k_2 + \ell_1 + \ell_2$  coefficients. Thus identification requires  $k_1 \geq \ell_2$ ; that there are at least as many exogenous time-varying regressors as endogenous time-invariant regressors. (This includes the model of the previous section, where  $k_1 = \ell_2 = 0$ .)

Given the moment conditions, the coefficients  $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$  can be estimated by 2SLS regression of (17.81) using the instruments  $\mathbf{Z} = (\dot{\mathbf{X}}_1, \dot{\mathbf{X}}_2, \bar{\mathbf{X}}_1, \mathbf{Z}_1)$ , or equivalently  $\mathbf{Z} = (\mathbf{X}_1, \dot{\mathbf{X}}_2, \bar{\mathbf{X}}_1, \mathbf{Z}_1)$ . This is 2SLS regression treating  $\mathbf{X}_1$  and  $\mathbf{Z}_1$  as exogenous and  $\mathbf{X}_2$  and  $\mathbf{Z}_2$  as endogenous, using the excluded instruments  $\dot{\mathbf{X}}_2$  and  $\bar{\mathbf{X}}_1$ . Setting  $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{Z}_1, \mathbf{Z}_2)$ , this is

$$\hat{\boldsymbol{\beta}}_{2sls} = \left( \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} \right)^{-1} \left( \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} \right).$$

It is recommended to use cluster-robust covariance matrix estimation, clustered at the individual level. Neither conventional nor heteroskedasticity-robust covariance matrix estimators should be used, as they are misspecified due to the individual-specific effect  $u_i$ .

When the model is just-identified, the estimators simplify as follows.  $\hat{\beta}_1$  and  $\hat{\beta}_2$  are the fixed effects estimator.  $\hat{\gamma}_1$  and  $\hat{\gamma}_2$  equal the 2SLS estimator from a regression of  $\hat{u}$  on  $Z_1$  and  $Z_2$ , using  $\bar{X}_1$  as an instrument for  $Z_2$ . (See Exercise 17.14.)

When the model is over-identified the equation can also be estimated by GMM with a cluster-robust weight matrix, using the same equations and instruments.

This estimator with cluster-robust standard errors can be calculated using the Stata `ivregress cluster(id)` command after constructing the transformed variables  $\tilde{X}_2$  and  $\bar{X}_1$ .

The 2SLS estimator described above corresponds with the Hausman and Taylor (1981) estimator in the just-identified case with a balanced panel.

Hausman and Taylor derived their estimator under the stronger assumption that the errors  $\varepsilon_{it}$  and  $u_i$  are strictly mean independent and homoskedastic, and consequently proposed a GLS-type estimator which is more efficient when these assumptions are correct. Define  $\Omega = \text{diag}(\Omega_i)$  where  $\Omega_i = I_i + \mathbf{1}_i \mathbf{1}'_i \sigma_u^2 / \sigma_\varepsilon^2$  and  $\sigma_\varepsilon^2$  and  $\sigma_u^2$  are the variances of the error components  $\varepsilon_{it}$  and  $u_i$ . Define as well the transformed variables  $\tilde{y} = \Omega^{-1/2} y$ ,  $\tilde{X} = \Omega^{-1/2} X$  and  $\tilde{Z} = \Omega^{-1/2} Z$ . The Hausman-Taylor estimator is

$$\begin{aligned}\hat{\beta}_{ht} &= \left( X' \Omega^{-1} Z (Z' \Omega^{-1} Z)^{-1} Z' \Omega^{-1} X \right)^{-1} \left( X' \Omega^{-1} Z (Z' \Omega^{-1} Z)^{-1} Z' \Omega^{-1} y \right) \\ &= \left( \tilde{X}' \tilde{Z} (\tilde{Z}' \tilde{Z})^{-1} \tilde{Z}' \tilde{X} \right)^{-1} \left( \tilde{X}' \tilde{Z} (\tilde{Z}' \tilde{Z})^{-1} \tilde{Z}' \tilde{y} \right).\end{aligned}$$

Recall from (17.47) that  $\Omega_i^{-1/2} = M_i + \rho_i P_i$  where  $\rho_i$  is defined in (17.46). Thus

$$\begin{aligned}\tilde{y}_i &= y_i - (1 - \rho_i) \bar{y}_i \\ \tilde{X}_{1i} &= X_{1i} - (1 - \rho_i) \bar{X}_{1i} \\ \tilde{X}_{2i} &= X_{2i} - (1 - \rho_i) \bar{X}_{2i} \\ \tilde{Z}_{1i} &= \rho_i Z_{1i} \\ \tilde{Z}_{2i} &= \rho_i Z_{2i} \\ \tilde{X}_{1i} &= \dot{X}_{1i} \\ \tilde{X}_{2i} &= \dot{X}_{2i} \\ \tilde{X}_{1i} &= \rho_i \bar{X}_{1i}.\end{aligned}$$

It follows that the Hausman-Taylor estimator can be calculated by 2SLS regression of  $\tilde{y}_i$  on  $(\tilde{X}_{1i}, \tilde{X}_{2i}, \rho_i Z_{1i}, \rho_i Z_{2i})$  using the instruments  $(\dot{X}_{1i}, \dot{X}_{2i}, \rho_i \bar{X}_{1i}, \rho_i Z_{2i})$ .

When the panel is balanced the coefficients  $\rho_i$  all equal and scale out from the instruments. Thus the estimator can be calculated by 2SLS regression of  $\tilde{y}_i$  on  $(\tilde{X}_{1i}, \tilde{X}_{2i}, Z_{1i}, Z_{2i})$  using the instruments  $(\dot{X}_{1i}, \dot{X}_{2i}, \bar{X}_{1i}, Z_{2i})$ .

In practice  $\rho_i$  is unknown. It can be estimated as in (17.48), with the modification that the variance of the combined error can be estimated from the untransformed 2SLS regression. Under the homoskedasticity assumptions used by Hausman and Taylor, the estimator  $\hat{\beta}_{ht}$  has a classical asymptotic covariance matrix. When these assumptions are relaxed the covariance matrix can be estimated using cluster-robust methods.

The Hausman-Taylor estimator with cluster-robust standard errors can be implemented in Stata by the command `xthtaylor vce(robust)`. This Stata command, for an unknown reason, requires that there is at least one exogenous time-invariant variable ( $\ell_1 \geq 1$ ), and at least one exogenous time-varying variable ( $k_1 \geq 1$ ), even when the model is identified. Otherwise, the estimator can be implemented using the instrumental variable method described above.

The Hausman-Taylor estimator was refined by Amemiya and MacCurdy (1986) and Breusch, Mizon and Schmidt (1989), who proposed more efficient versions using additional instruments which are valid under stronger orthogonality conditions. The observation that in the unbalanced case the instruments should be weighted by  $\rho_i$  was made by Gardner (1998).

In the over-identified case it is unclear if it is preferred to use the simpler 2SLS estimator  $\hat{\beta}_{2\text{sls}}$  or the GLS-type Hausman-Taylor estimator  $\hat{\beta}_{\text{ht}}$ . The advantages of  $\hat{\beta}_{\text{ht}}$  are that it is asymptotically efficient under their stated homoskedasticity and serial correlation conditions, and that there is an available program in Stata. The advantage of  $\hat{\beta}_{2\text{sls}}$  is that it is much simpler to program (if doing so yourself), may have better finite sample properties (since it avoids variance-component estimation), and is the natural estimator from the the modern GMM viewpoint.

To illustrate, the final column of Table 17.2 contains Hausman-Taylor estimates of the investment model, treating  $Q_{it-1}$ ,  $D_{it-1}$ , and  $T_i$  as endogenous for  $u_i$ , and  $CF_{it-1}$  and the industry dummies as exogenous. Relative to the fixed effects models, this allows estimation of the coefficients on the trading indicator  $T_i$ . The most interesting change relative to the previous estimates is that the coefficient on the trading indicator  $T_i$  doubles in magnitude (relative to the random effects estimate). This is consistent with the hypothesis that  $T_i$  is correlated with the fixed effect, and hence the random effects estimate is biased.

### 17.34 Jackknife Covariance Matrix Estimation

As an alternative to asymptotic inference, the delete-cluster jackknife can be used for covariance matrix calculation. In the context of fixed effects estimaion the delete-cluster estimators take the form

$$\begin{aligned}\hat{\beta}_{(-i)} &= \left( \sum_{j \neq i} \dot{X}'_j \dot{X}_j \right)^{-1} \left( \sum_{j \neq i} \dot{X}'_j \dot{y}_j \right) \\ &= \hat{\beta}_{\text{fe}} - \left( \sum_{i=1}^N \dot{X}'_i \dot{X}_i \right)^{-1} \dot{X}'_i \tilde{\mathbf{e}}_g.\end{aligned}$$

where

$$\begin{aligned}\tilde{\mathbf{e}}_g &= \left( \mathbf{I}_i - \dot{X}_i \left( \dot{X}'_i \dot{X}_i \right)^{-1} \dot{X}'_i \right)^{-1} \hat{\mathbf{e}}_i \\ \hat{\mathbf{e}}_i &= \dot{y}_i - \dot{X}_i \hat{\beta}_{\text{fe}}.\end{aligned}$$

The delete-cluster jackknife estimator of the variance of  $\hat{\beta}_{\text{fe}}$  is

$$\begin{aligned}\hat{V}_{\hat{\beta}}^{\text{jack}} &= \frac{N-1}{N} \sum_{i=1}^N \left( \hat{\beta}_{(-i)} - \bar{\beta} \right) \left( \hat{\beta}_{(-i)} - \bar{\beta} \right)' \\ \bar{\beta} &= \frac{1}{N} \sum_{i=1}^N \hat{\beta}_{(-i)}.\end{aligned}$$

The delete-cluster jackknife estimator  $\hat{V}_{\hat{\beta}}^{\text{jack}}$  is similar to the cluster-robust covariance matrix estimator.

For parameters which are functions  $\hat{\theta}_{\text{fe}} = \mathbf{r}(\hat{\beta}_{\text{fe}})$  of the fixed effects estimator, the delete-cluster jackknife estimator of the variance of  $\hat{\theta}_{\text{fe}}$  is

$$\begin{aligned}\hat{V}_{\hat{\theta}}^{\text{jack}} &= \frac{N-1}{N} \sum_{i=1}^N \left( \hat{\theta}_{(-i)} - \bar{\theta} \right) \left( \hat{\theta}_{(-i)} - \bar{\theta} \right)' \\ \hat{\theta}_{(-i)} &= \mathbf{r}(\hat{\beta}_{(-i)}) \\ \bar{\theta} &= \frac{1}{N} \sum_{i=1}^N \hat{\theta}_{(-i)}.\end{aligned}$$

The estimator  $\hat{V}_{\hat{\theta}}^{\text{jack}}$  is similar to the delta-method cluster-robust covariance matrix estimator for  $\hat{\theta}$ .

As in the context of i.i.d. samples, one advantage of the jackknife covariance matrix estimators is that they do not require the user to make a technical calculation of the asymptotic distribution. A downside is an increase in computation cost, as  $N$  separate regressions are effectively estimated. This can be particularly costly in micro panels which have a large number  $N$  of individuals.

In Stata, jackknife standard errors for fixed effects estimators are obtained by using either `xtreg fe vce(jackknife)` or `areg absorb(id) cluster(id) vce(jackknife)` where `id` is the cluster variable. For the fixed effects 2SLS estimator, use `xtivreg fe vce(jackknife)`.

## 17.35 Panel Bootstrap

Bootstrap methods can also be applied to panel data by a straightforward application of the pairs cluster bootstrap, which samples entire individuals rather than single observations. In the context of panel data we call this the panel nonparametric bootstrap.

The **panel nonparametric bootstrap** samples  $N$  individual histories  $(\mathbf{y}_i, \mathbf{X}_i)$  to create the bootstrap sample. Fixed effects (or any other estimation method) is applied to the bootstrap sample to obtain the coefficient estimates. By repeating  $B$  times, bootstrap standard errors for coefficients estimates, or functions of the coefficient estimates, can be calculated. Percentile-type and percentile-t confidence intervals can be calculated. The  $BC_a$  interval requires an estimator of the acceleration coefficient  $a$  which is a scaled jackknife estimate of the third moment of the estimator. In panel data the delete-cluster jackknife should be used for estimation of  $a$ .

In Stata, to obtain bootstrap standard errors and confidence intervals use either `xtreg, vce(bootstrap, reps(#))` or `areg, absorb(id) cluster(id) vce(bootstrap, reps(#))`, where `id` is the cluster variable and `#` is the number of bootstrap replications. For the fixed effects 2SLS estimator, use `xtivreg, fe vce(bootstrap, reps(#))`.

## 17.36 Dynamic Panel Models

The models so far considered in this chapter have been static, with no dynamic relationships. In many economic contexts it is natural to expect that behavior and decisions are dynamic, explicitly depending on past behavior. In our investment equation, for example, economic models predict that a firm's investment in any given year will depend on investment decisions from previous years. These considerations lead us to consider explicitly dynamic models.

The workhorse dynamic model in a panel framework is the  $p^{th}$ -order autoregression (or AR( $p$ )) with regressors and a one-way error component structure. This is

$$y_{it} = \alpha_1 y_{it-1} + \cdots + \alpha_p y_{it-p} + \mathbf{x}'_{it} \boldsymbol{\beta} + u_i + \varepsilon_{it}. \quad (17.82)$$

where  $\alpha_j$  are the autoregressive coefficients,  $\mathbf{x}_{it}$  is a  $k$  vector of regressors,  $u_i$  is an individual-effect, and  $\varepsilon_{it}$  is an idiosyncratic error. It is conventional to assume that the errors  $u_i$  and  $\varepsilon_{it}$  are mutually independent, and the  $\varepsilon_{it}$  are serially uncorrelated and mean zero. For the present we will assume that the regressors  $\mathbf{x}_{it}$  are strictly exogenous (17.17). In Section 17.41 we discuss the case of predetermined regressors.

For many illustrations we will focus on the AR(1) model

$$y_{it} = \alpha y_{it-1} + u_i + \varepsilon_{it} \quad (17.83)$$

The dynamics should be interpreted individual-by-individual. The coefficient  $\alpha$  in (17.83) equals the first-order autocorrelation. When  $\alpha = 0$  the series is serially uncorrelated (conditional on  $u_i$ ).  $\alpha > 0$  means  $y_{it}$  is positively serially correlated.  $\alpha < 0$  means  $y_{it}$  is negatively serially correlated. An autoregressive unit root holds when  $\alpha = 1$ , which means that  $y_{it}$  follows a random walk with possible drift. Since  $u_i$  is constant for a given individual, it should be treated as an individual-specific intercept. The idiosyncratic error  $\varepsilon_{it}$  plays the role of the error in a standard time series autoregression.

If  $|\alpha| < 1$  then the model (17.83) is stationary. By standard autoregressive backwards recursion we can calculate that

$$y_{it} = \sum_{j=0}^{\infty} \alpha^j (u_i + \varepsilon_{it}) = (1 - \alpha)^{-1} u_i + \sum_{j=0}^{\infty} \alpha^j \varepsilon_{it-j}. \quad (17.84)$$

Thus if we condition on  $u_i$  the conditional mean and variance of  $y_{it}$  is  $(1 - \alpha)^{-1} u_i$  and  $(1 - \alpha^2)^{-1} \sigma_\varepsilon^2$ , respectively. The  $k^{th}$  autocorrelation (conditional on  $u_i$ ) is  $\alpha^k$ . Notice that the effect of cross-section variation in  $u_i$  is to shift the (conditional) mean, but not the variance or serial correlation. This implies that if we view time series plots of  $y_{it}$  against time for a set of individuals  $i$ , the series  $y_{it}$  will appear to have different means, but have similar variances and time series serial correlation.

As with the case with time series data, serial correlation (large  $\alpha$ ) can proxy for other factors such as time trends. Thus in applications it will often be useful to include time effects to eliminate spurious serial correlation.

### 17.37 The Bias of Fixed Effects Estimation

To estimate the panel autoregression (17.82) it may appear natural to use the fixed effects (within) estimator. Indeed, the within transformation eliminates the individual effect  $u_i$ . The trouble is that the within operator induces correlation between the AR(1) lag and the error. The result is that the within estimator is inconsistent for the coefficients when  $T$  is fixed. A thorough explanation appears in Nickell (1981). We describe the basic problem in this section focusing on the AR(1) model (17.83).

Applying the within operator to (17.83) we obtain

$$\dot{y}_{it} = \alpha \dot{y}_{it-1} + \dot{\varepsilon}_{it}$$

for  $t \geq 2$ . As expected the individual effect is eliminated. The difficulty is that  $\mathbb{E}(\dot{y}_{it-1} \dot{\varepsilon}_{it}) \neq 0$ , since both  $\dot{y}_{it-1}$  and  $\dot{\varepsilon}_{it}$  are functions of the entire time series.

To see this clearly in a simple example, suppose we have a balanced panel with  $T = 3$ . There are two observed pairs  $(y_{it}, y_{it-1})$  per individual so the within estimator equals the differenced estimator. Applying the differencing operator to (17.83) for  $t = 3$  we find

$$\Delta y_{i3} = \alpha \Delta y_{i2} + \Delta \varepsilon_{i3}. \quad (17.85)$$

Because of the lagged dependent variable and differencing there is effectively one observation per individual. Notice that the individual effect has been eliminated.

The fixed effects estimator of  $\alpha$  is equal to the least-squares estimator applied to (17.85), which is

$$\begin{aligned} \hat{\alpha}_{\text{fe}} &= \left( \sum_{i=1}^N \Delta y_{i2}^2 \right)^{-1} \left( \sum_{i=1}^N \Delta y_{i2} \Delta y_{i3} \right) \\ &= \alpha + \left( \sum_{i=1}^N \Delta y_{i2}^2 \right)^{-1} \left( \sum_{i=1}^N \Delta y_{i2} \Delta \varepsilon_{i3} \right). \end{aligned}$$

This estimator is inconsistent for  $\alpha$  since the differenced regressor and error are negatively correlated. Indeed

$$\begin{aligned} \mathbb{E}(\Delta y_{i2} \Delta \varepsilon_{i3}) &= \mathbb{E}((y_{i2} - y_{i1})(\varepsilon_{i3} - \varepsilon_{i2})) \\ &= \mathbb{E}(y_{i2} \varepsilon_{i3}) - \mathbb{E}(y_{i1} \varepsilon_{i3}) - \mathbb{E}(y_{i2} \varepsilon_{i2}) + \mathbb{E}(y_{i1} \varepsilon_{i2}) \\ &= 0 - 0 - \sigma_\varepsilon^2 + 0 \\ &= -\sigma_\varepsilon^2. \end{aligned}$$

Using the variance formula for AR(1) models (assuming  $|\alpha| < 1$ ) we can calculate that  $\mathbb{E}(\Delta y_{i2})^2 = 2\sigma_\varepsilon^2/(1 + \alpha)$ . It follows that the probability limit of the fixed effects estimator  $\hat{\alpha}_{\text{fe}}$  of  $\alpha$  in (17.85) is

$$\text{plim}_{N \rightarrow \infty} (\hat{\alpha}_{\text{fe}} - \alpha) = \frac{\mathbb{E}(\Delta y_{i2} \Delta \varepsilon_{i3})}{\mathbb{E}(\Delta y_{i2})^2} = -\frac{1 + \alpha}{2}. \quad (17.86)$$

It is typical to call (17.86) the “bias” of  $\hat{\alpha}_{\text{fe}}$ , though it is technically the probability limit.

The bias found in (17.86) is large. For  $\alpha = 0$  the bias is  $-1/2$  and increases towards 1 as  $\alpha \rightarrow 1$ . Thus for any  $\alpha < 1$  the probability limit of  $\hat{\alpha}_{\text{fe}}$  is negative! This is extreme bias.

From Nickell's (1981) expressions and some algebra, we can calculate that the probability limit of the fixed effects estimator for  $|\alpha| < 1$  and general  $T$  is

$$\underset{N \rightarrow \infty}{\text{plim}} (\hat{\alpha}_{\text{fe}} - \alpha) = \frac{1 + \alpha}{\frac{2\alpha}{1 - \alpha} - \frac{T - 1}{1 - \alpha^{T-1}}}. \quad (17.87)$$

It follows that the bias is of order  $O(1/T)$ .

One might guess (and is often asserted) that it is okay to use fixed effects if  $T$  is large, say  $T \geq 30$  or perhaps  $T \geq 60$ . However, from (17.87) we can calculate that for  $T = 30$  the bias of the fixed effects estimator is  $-0.056$  when  $\alpha = 0.5$  and the bias is  $-0.15$  when  $\alpha = 0.9$ . For  $T = 60$  and  $\alpha = 0.9$  the bias is  $-0.05$ . These magnitudes are unacceptably large. This includes the longer time series encountered in macro panels. Thus the Nickell bias problem applies to both micro and macro panel applications.

The conclusion from this analysis is that the fixed effects estimator should not be used for models with lagged dependent variables, even if the time series dimension  $T$  is large.

### 17.38 Anderson-Hsiao Estimator

Anderson and Hsiao (1982) made an important breakthrough by showing that a simple instrumental variables estimator is consistent for the parameters of (17.82).

The method first eliminates the individual effect  $u_i$  by first-differencing (17.82) for  $t \geq p + 1$

$$\Delta y_{it} = \alpha_1 \Delta y_{it-1} + \alpha_2 \Delta y_{it-2} + \cdots + \alpha_p \Delta y_{it-p} + \Delta \mathbf{x}'_{it} \boldsymbol{\beta} + \Delta \varepsilon_{it}. \quad (17.88)$$

This eliminates the individual effect  $u_i$ . The challenge is that first-differencing induces correlation between  $\Delta y_{it-1}$  and  $\Delta \varepsilon_{it}$ :

$$\mathbb{E}(\Delta y_{it-1} \Delta \varepsilon_{it}) = \mathbb{E}((y_{it} - y_{it-1})(\varepsilon_{it} - \varepsilon_{it-1})) = -\sigma_\varepsilon^2.$$

The other regressors are not correlated with  $\Delta \varepsilon_{it}$ . For  $s > 1$

$$\mathbb{E}(\Delta y_{it-s} \Delta \varepsilon_{it}) = 0$$

and when  $\mathbf{x}_{it}$  is strictly exogenous

$$\mathbb{E}(\Delta \mathbf{x}'_{it} \Delta \varepsilon_{it}) = 0.$$

The correlation between  $\Delta y_{it-1}$  and  $\Delta \varepsilon_{it}$  is an endogeneity problem. One solution to endogeneity is to use an instrument. Anderson-Hsiao pointed out that  $y_{it-2}$  is a valid instrument since it is correlated with  $\Delta y_{it-1}$  yet uncorrelated with  $\Delta \varepsilon_{it}$ .

$$\mathbb{E}(y_{it-2} \Delta \varepsilon_{it}) = \mathbb{E}(y_{it-2} \varepsilon_{it}) - \mathbb{E}(y_{it-2} \varepsilon_{it-1}) = 0. \quad (17.89)$$

The Anderson-Hsiao estimator is IV using  $y_{it-2}$  as an instrument for  $\Delta y_{it-1}$ . Equivalently, this is IV using the instruments  $(y_{it-2}, \dots, y_{it-p-1})$  for  $(\Delta y_{it-1}, \dots, \Delta y_{it-p})$ . The estimator requires  $T \geq p + 2$ .

To show that this estimator is consistent, for simplicity assume we have a balanced panel with  $T = 3$ ,  $p = 1$ , and no regressors. In this case the Anderson-Hsiao IV estimator is

$$\begin{aligned} \hat{\alpha}_{\text{iv}} &= \left( \sum_{i=1}^N y_{i1} \Delta y_{i2} \right)^{-1} \left( \sum_{i=1}^N y_{i1} \Delta y_{i3} \right) \\ &= \alpha + \left( \sum_{i=1}^N y_{i1} \Delta y_{i2} \right)^{-1} \left( \sum_{i=1}^N y_{i1} \Delta \varepsilon_{i3} \right). \end{aligned}$$

Under the assumption that  $\varepsilon_{it}$  is serially uncorrelated, (17.89) shows that  $\mathbb{E}(y_{i1}\Delta\varepsilon_{i3}) = 0$ . In general,  $\mathbb{E}(y_{i1}\Delta y_{i2}) \neq 0$ . As  $N \rightarrow \infty$

$$\hat{\alpha}_{\text{iv}} \xrightarrow{p} \alpha - \frac{\mathbb{E}(y_{i1}\Delta\varepsilon_{i3})}{\mathbb{E}(y_{i1}\Delta y_{i2})} = \alpha.$$

Thus the IV estimator is consistent for  $\alpha$ .

The Anderson-Hsiao IV estimator relies on two critical assumptions. First, the validity of the instrument (uncorrelatedness with the equation error) relies on the assumption that the dynamics are correctly specified so that  $\varepsilon_{it}$  is serially uncorrelated. For example, many applications use an AR(1). If instead the true model is an AR(2) then  $y_{it-2}$  is not a valid instrument and the IV estimates will be biased. Second, the relevance of the instrument (correlatedness with the endogenous regressor) requires  $\mathbb{E}(y_{i1}\Delta y_{i2}) \neq 0$ . This turns out to be problematic and is explored further in Section 17.40. These considerations suggest that the validity and accuracy of the estimator are likely to be sensitive to these unknown features.

### 17.39 Arellano-Bond Estimator

The orthogonality condition (17.89) is one of many implied by the dynamic panel model. Indeed, all lags  $y_{it-2}, y_{it-3}, \dots$  are valid instruments. If  $T > p+2$  these can be used to potentially improve estimation efficiency. This was first pointed out by Holtz-Eakin, Newey and Rosen (1988) and further developed by Arellano and Bond (1991).

Using these extra instruments has a complication that there are a different number of instruments for each time period. The solution is to view the model as a system of  $T$  equations as in Section 17.18.

It will be useful to first write the model in vector notation. Stacking the differenced regressors  $(\Delta y_{it-1}, \dots, \Delta y_{it-p}, \Delta \mathbf{x}'_{it})$  into a matrix  $\Delta \mathbf{X}_i$  and the coefficients into a vector  $\boldsymbol{\theta}$  we can write (17.88) as

$$\Delta \mathbf{y}_i = \Delta \mathbf{X}_i \boldsymbol{\theta} + \Delta \boldsymbol{\varepsilon}_i.$$

Stacking all  $N$  individuals this can be written as

$$\Delta \mathbf{y} = \Delta \mathbf{X} \boldsymbol{\theta} + \Delta \boldsymbol{\varepsilon}.$$

For period  $t = p+2$  we have the  $p+k$  valid instruments  $[y_{i1}, \dots, y_{ip}, \Delta \mathbf{x}'_{i,p+2}]$ . For period  $t = p+3$  there are  $p+1+k$  valid instruments  $[y_{i1}, \dots, y_{ip+1}, \Delta \mathbf{x}'_{i,p+3}]$ . For period  $t = p+4$  there are  $p+2+k$  instruments. In general, for any  $t \geq p+2$  there are  $t-2$  instruments  $[y_{i1}, \dots, y_{i,t-2}, \Delta \mathbf{x}'_{it}]$ . Similarly to (17.53) we can define the instrument matrix for individual  $i$  as

$$\mathbf{Z}_i = \begin{bmatrix} [y_{i1}, \dots, y_{ip}, \Delta \mathbf{x}'_{i,p+2}] & 0 & 0 \\ 0 & [y_{i1}, \dots, y_{ip+1}, \Delta \mathbf{x}'_{i,p+3}] & 0 \\ & & \ddots \\ 0 & 0 & [y_{i1}, y_{i2}, \dots, y_{i,T-2}, \Delta \mathbf{x}'_{i,T}] \end{bmatrix}. \quad (17.90)$$

This is  $(T-p-1) \times \ell$  where  $\ell = k(T-p-1) + ((T-2)(T-1) - (p-2)(p-1))/2$ . This instrument matrix consists of all lagged values  $y_{it-2}, y_{it-3}, \dots$  which are available in the data set, plus the differenced strictly exogenous regressors.

The  $\ell$  moment conditions are

$$\mathbb{E}(\mathbf{Z}'_i (\Delta \mathbf{y}_i - \Delta \mathbf{X}_i \boldsymbol{\alpha})) = 0. \quad (17.91)$$

If  $T > p+2$  then  $\ell > p$  and the model is overidentified. Define the  $\ell \times \ell$  covariance matrix for the moment conditions

$$\boldsymbol{\Omega} = \mathbb{E}(\mathbf{Z}'_i \Delta \boldsymbol{\varepsilon}_i \Delta \boldsymbol{\varepsilon}'_i \mathbf{Z}_i).$$

Let  $\mathbf{Z}$  denote  $\mathbf{Z}_i$  stacked into a  $(T-p-1)N \times \ell$  matrix. The efficient GMM estimator of  $\boldsymbol{\alpha}$  is

$$\hat{\boldsymbol{\alpha}}_{\text{gmm}} = (\Delta \mathbf{X}' \mathbf{Z} \boldsymbol{\Omega}^{-1} \mathbf{Z}' \Delta \mathbf{X})^{-1} (\Delta \mathbf{X}' \mathbf{Z} \boldsymbol{\Omega}^{-1} \mathbf{Z}' \Delta \mathbf{y}).$$

If the errors  $\varepsilon_{it}$  are conditionally homoskedastic then

$$\boldsymbol{\Omega} = \mathbb{E}(\mathbf{Z}'_i \mathbf{H} \mathbf{Z}_i) \sigma_\varepsilon^2$$

where  $\mathbf{H}$  is given in (17.31). In this case set

$$\widehat{\boldsymbol{\Omega}}_1 = \sum_{i=1}^N \mathbf{Z}'_i \mathbf{H} \mathbf{Z}_i$$

as a (scaled) estimate of  $\boldsymbol{\Omega}$ . Under these assumptions an asymptotically efficient GMM estimator is

$$\widehat{\alpha}_1 = \left( \Delta \mathbf{X}' \mathbf{Z} \widehat{\boldsymbol{\Omega}}_1^{-1} \mathbf{Z}' \Delta \mathbf{X} \right)^{-1} \left( \Delta \mathbf{X}' \mathbf{Z} \widehat{\boldsymbol{\Omega}}_1^{-1} \mathbf{Z}' \Delta \mathbf{y} \right). \quad (17.92)$$

Estimator (17.92) is known as the **one-step Arellano-Bond GMM estimator**.

Under the assumption that the error  $\varepsilon_{it}$  is homoskedastic and serially uncorrelated, a classical covariance matrix estimator for  $\widehat{\alpha}_1$  is

$$\widehat{V}_1^0 = \left( \Delta \mathbf{X}' \mathbf{Z} \widehat{\boldsymbol{\Omega}}_1^{-1} \mathbf{Z}' \Delta \mathbf{X} \right)^{-1} \widehat{\sigma}_\varepsilon^2 \quad (17.93)$$

where  $\widehat{\sigma}_\varepsilon^2$  is the sample variance of the one-step residuals  $\widehat{\varepsilon}_i = \Delta \mathbf{y}_i - \Delta \mathbf{X}_i \widehat{\alpha}$ . A covariance matrix estimator which is robust to violation of these assumptions is

$$\widehat{V}_1 = \left( \Delta \mathbf{X}' \mathbf{Z} \widehat{\boldsymbol{\Omega}}_1^{-1} \mathbf{Z}' \Delta \mathbf{X} \right)^{-1} \left( \Delta \mathbf{X}' \mathbf{Z} \widehat{\boldsymbol{\Omega}}_1^{-1} \mathbf{Z}' \widehat{\boldsymbol{\Omega}}_2 \mathbf{Z} \widehat{\boldsymbol{\Omega}}_1^{-1} \mathbf{Z}' \Delta \mathbf{X} \right) \left( \Delta \mathbf{X}' \mathbf{Z} \widehat{\boldsymbol{\Omega}}_1^{-1} \mathbf{Z}' \Delta \mathbf{X} \right)^{-1} \quad (17.94)$$

where

$$\widehat{\boldsymbol{\Omega}}_2 = \sum_{i=1}^N \mathbf{Z}'_i \widehat{\varepsilon}_i \widehat{\varepsilon}'_i \mathbf{Z}_i$$

is a (scaled) cluster-robust estimator of  $\boldsymbol{\Omega}$  using the one-step residuals.

An asymptotically efficient two-step GMM estimator which relaxes the assumption of homoskedasticity is

$$\widehat{\alpha}_2 = \left( \Delta \mathbf{X}' \mathbf{Z} \widehat{\boldsymbol{\Omega}}_2^{-1} \mathbf{Z}' \Delta \mathbf{X} \right)^{-1} \left( \Delta \mathbf{X}' \mathbf{Z} \widehat{\boldsymbol{\Omega}}_2^{-1} \mathbf{Z}' \Delta \mathbf{y} \right). \quad (17.95)$$

Estimator (17.95) is known as the **two-step Arellano-Bond GMM estimator**. An appropriate robust covariance matrix estimator for  $\widehat{\alpha}_2$  is

$$\widehat{V}_2 = \left( \Delta \mathbf{X}' \mathbf{Z} \widehat{\boldsymbol{\Omega}}_2^{-1} \mathbf{Z}' \Delta \mathbf{X} \right)^{-1} \left( \Delta \mathbf{X}' \mathbf{Z} \widehat{\boldsymbol{\Omega}}_2^{-1} \mathbf{Z}' \widehat{\boldsymbol{\Omega}}_3 \mathbf{Z} \widehat{\boldsymbol{\Omega}}_2^{-1} \mathbf{Z}' \Delta \mathbf{X} \right) \left( \Delta \mathbf{X}' \mathbf{Z} \widehat{\boldsymbol{\Omega}}_2^{-1} \mathbf{Z}' \Delta \mathbf{X} \right)^{-1} \quad (17.96)$$

where

$$\widehat{\boldsymbol{\Omega}}_3 = \sum_{i=1}^N \mathbf{Z}'_i \widehat{\varepsilon}_i \widehat{\varepsilon}'_i \mathbf{Z}_i$$

is a (scaled) cluster-robust estimator of  $\boldsymbol{\Omega}$  using the two-step residuals  $\widehat{\varepsilon}_i = \Delta \mathbf{y}_i - \Delta \mathbf{X}_i \widehat{\alpha}_2$ . Asymptotically,  $\widehat{V}_2$  is equivalent to

$$\tilde{V}_2 = \left( \Delta \mathbf{X}' \mathbf{Z} \widehat{\boldsymbol{\Omega}}_2^{-1} \mathbf{Z}' \Delta \mathbf{X} \right)^{-1}. \quad (17.97)$$

The GMM estimator can be iterated until convergence to produce an iterated GMM estimator.

The advantage of the Arellano-Bond estimator over the Anderson-Hsiao estimator is that when  $T > p+2$  the additional (overidentified) moment conditions reduce the asymptotic variance of the estimator and stabilize its performance. The disadvantage is that when  $T$  is large using the full set of lags as instruments may cause a “many weak instruments” problem. The advised compromise is to limit the number of lags used as instruments.

The advantage of the one-step Arellano-Bond estimator is that the weight matrix  $\widehat{\boldsymbol{\Omega}}_1$  does not depend on residuals and is therefore less random than the two-step weight matrix  $\widehat{\boldsymbol{\Omega}}_2$ . This can result in better performance by the one-step estimator in small to moderate samples, especially when the errors are

approximately homoskedastic. The advantage of the two-step estimator is that it achieves asymptotic efficiency allowing for heteroskedasticity, and is thus expected to perform better in large samples with non-homoskedastic errors.

To summarize, the Arellano-Bond estimator applies GMM to the first-differenced equation (17.88) using a set of available lags  $y_{it-2}, y_{it-3}, \dots$  as instruments for  $\Delta y_{it-1}, \dots, \Delta y_{it-p}$ .

The Arellano-Bond estimator may be obtained in Stata using either the `xtabond` or `xtdpd` command. The default setting is the one-step estimator (17.92) and non-robust standard errors (17.93). For the two-step estimator and robust standard errors use the `twostep vce(robust)` options. Reported standard errors in Stata are based on Windmeijer's (2005) finite-sample correction to the asymptotic estimate (17.97). The robust covariance matrix (17.96) nor the iterated GMM estimator are implemented.

## 17.40 Weak Instruments

Blundell and Bond (1998) pointed out that the Anderson-Hsiao and Arellano-Bond class of estimators suffer from the problem of weak instruments. This can be seen easiest in the AR(1) model with the Anderson-Hsiao estimator which uses  $y_{it-2}$  as an instrument for  $\Delta y_{it-1}$ . The reduced form equation for  $\Delta y_{it-1}$  is

$$\Delta y_{it-1} = y_{it-2}\gamma + \nu_{it}.$$

The reduced form coefficient  $\gamma$  is defined by projection. Using  $\Delta y_{it-1} = (\alpha - 1)y_{it-2} + u_i + \varepsilon_{it-1}$  and  $\mathbb{E}(y_{it-2}\varepsilon_{it-1}) = 0$  we can calculate that

$$\begin{aligned}\gamma &= \frac{\mathbb{E}(y_{it-2}\Delta y_{it-1})}{\mathbb{E}(y_{it-2}^2)} \\ &= (\alpha - 1) + \frac{\mathbb{E}(y_{it-2}u_i)}{\mathbb{E}(y_{it-2}^2)}.\end{aligned}$$

Assuming stationarity so that (17.84) holds,

$$\begin{aligned}\mathbb{E}(y_{it-2}u_i) &= \mathbb{E}\left[\left(\frac{u_i}{1-\alpha} + \sum_{j=0}^{\infty} \alpha^j \varepsilon_{it-2-j}\right)u_i\right] \\ &= \frac{\sigma_u^2}{1-\alpha}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}(y_{it-2}^2) &= \mathbb{E}\left(\frac{u_i}{1-\alpha} + \sum_{j=0}^{\infty} \alpha^j \varepsilon_{it-2-j}\right)^2 \\ &= \frac{\sigma_u^2}{(1-\alpha)^2} + \frac{\sigma_\varepsilon^2}{(1-\alpha^2)}\end{aligned}$$

where  $\sigma_u^2 = \mathbb{E}(u_i^2)$  and  $\sigma_\varepsilon^2 = \mathbb{E}(\varepsilon_{it}^2)$ . Using these expressions and a fair amount of algebra, Blundell and Bond (1998) found that the reduced form coefficient equals

$$\gamma = (\alpha - 1) \left( \frac{k}{k + \sigma_u^2/\sigma_\varepsilon^2} \right) \quad (17.98)$$

where  $k = (1 - \alpha) / (1 + \alpha)$ .

The Anderson-Hsiao instrument  $y_{it-2}$  is weak if  $\gamma$  is close to zero. From (17.98) we see that  $\gamma = 0$  when either  $\alpha = 1$  (a unit root) or  $\sigma_u^2/\sigma_\varepsilon^2 = \infty$  (the idiosyncratic effect is small relative to the individual-specific effect). In either case the coefficient  $\alpha$  is not identified. We know from our earlier study of the weak instruments problem (Section 12.36) that when  $\gamma$  is close to zero then  $\alpha$  is weakly identified and

the estimators will perform poorly. This means that when the autoregressive coefficient  $\alpha$  is large or the individual-specific effect dominates the idiosyncratic effect, these estimators will be weakly identified, have poor performance, and conventional inference methods will be misleading. Since the value of  $\alpha$  and the relative variances are unknown *a priori*, this means that we should generically treat this class of estimators as weakly identified.

An alternative estimator which has improved performance under weak identification is discussed in Section 17.42.

## 17.41 Dynamic Panels with Predetermined Regressors

The assumption that regressors are strictly exogenous is restrictive. A less restrictive assumption is that the regressors are predetermined. Dynamic panel methods can be modified to handle predetermined regressors by using their lags as instruments

**Definition 17.2** The regressor  $\mathbf{x}_{it}$  is **predetermined** for the error  $\varepsilon_{it}$  if

$$\mathbb{E}(\mathbf{x}_{it-s}\varepsilon_{it}) = 0 \quad (17.99)$$

for all  $s \geq 0$ .

The difference between strictly exogenous and predetermined regressors is that for the former (17.99) holds for all  $s$ , not just  $s \geq 0$ . One way of interpreting a regression model with predetermined regressors is that the model is a projection on the complete past history of the regressors.

Under (17.99), leads of  $\mathbf{x}_{it}$  can be correlated with  $\varepsilon_{it}$ , that is  $\mathbb{E}(\mathbf{x}_{it+s}\varepsilon_{it}) \neq 0$  for  $s \geq 1$ , or equivalently  $\mathbf{x}_{it}$  can be correlated with lags of  $\varepsilon_{ij}$ , that is  $\mathbb{E}(\mathbf{x}_{it}\varepsilon_{it-s}) \neq 0$  for  $s \geq 1$ . This means that  $\mathbf{x}_{it}$  can respond dynamically to past values of  $y_{it}$ , as in, for example, an unrestricted vector autoregression.

Consider the differenced equation (17.88)

$$\Delta y_{it} = \alpha_1 \Delta y_{it-1} + \alpha_2 \Delta y_{it-2} + \cdots + \alpha_p \Delta y_{it-p} + \Delta \mathbf{x}'_{it} \boldsymbol{\beta} + \Delta \varepsilon_{it}.$$

When the regressors are predetermined but not strictly exogenous,  $\mathbf{x}_{it}$  and  $\varepsilon_{it}$  are uncorrelated, but  $\Delta \mathbf{x}_{it}$  and  $\Delta \varepsilon_{it}$  are correlated. To see this,

$$\begin{aligned} \mathbb{E}(\Delta \mathbf{x}_{it} \Delta \varepsilon_{it}) &= \mathbb{E}(\mathbf{x}_{it} \varepsilon_{it}) - \mathbb{E}(\mathbf{x}_{it-1} \varepsilon_{it}) - \mathbb{E}(\mathbf{x}_{it} \varepsilon_{it-1}) + \mathbb{E}(\mathbf{x}_{it-1} \varepsilon_{it-1}) \\ &= -\mathbb{E}(\mathbf{x}_{it} \varepsilon_{it-1}) \\ &\neq 0. \end{aligned}$$

This means that if we treat  $\Delta \mathbf{x}_{it}$  as exogenous, the coefficient estimates will be biased.

To solve the correlation problem we can use instruments for  $\Delta \mathbf{x}_{it}$ . A valid instrument is  $\mathbf{x}_{it-1}$ , since it is generally correlated with  $\Delta \mathbf{x}_{it}$  yet uncorrelated with  $\Delta \varepsilon_{it}$ . Indeed, for any  $s \geq 1$

$$\mathbb{E}(\mathbf{x}_{it-s} \Delta \varepsilon_{it}) = \mathbb{E}(\mathbf{x}_{it-s} \varepsilon_{it}) - \mathbb{E}(\mathbf{x}_{it-s} \varepsilon_{it-1}) = 0.$$

Consequently, Arellano and Bond (1991) recommend using the instrument set  $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{it-1})$ . When the number of time periods is large it is advised to limit the number of instrument lags to avoid the many weak instruments problem.

Algebraically, GMM estimation is the same as the estimators described in Section 17.39, except that the instrument matrix (17.90) is modified to

$$\mathbf{Z}_i = \begin{bmatrix} [y_{i1}, \dots, y_{ip}, \mathbf{x}'_{i1}, \dots, \mathbf{x}'_{ip+1}] & 0 & 0 \\ 0 & [y_{i1}, \dots, y_{ip+1}, \mathbf{x}'_{i1}, \dots, \mathbf{x}'_{ip+2}] & 0 \\ & & \ddots \\ 0 & 0 & [y_{i1}, \dots, y_{iT-2}, \mathbf{x}'_{i1}, \dots, \mathbf{x}'_{iT-1}] \end{bmatrix}. \quad (17.100)$$

To understand how the model is identified we examine the reduced form equation for the regressor. For  $t = p + 2$  and using the first lag as an instrument the reduced form is

$$\Delta \mathbf{x}_{it} = \gamma_1 y_{it-2} + \Gamma_2 \mathbf{x}_{it-1} + \zeta_{it}.$$

The model is identified if  $\Gamma_2$  is full rank. This is valid (in general) when  $\mathbf{x}_{it}$  is stationary. Identification fails, however, when  $\mathbf{x}_{it}$  has a unit root. This indicates that the model will be weakly identified when the predetermined regressors are highly persistent.

The method generalizes to handle multiple lags of the predetermined regressors. To see this, write the model explicitly as

$$y_{it} = \alpha_1 y_{it-1} + \dots + \alpha_p y_{it-p} + \mathbf{x}'_{it} \boldsymbol{\beta}_1 + \dots + \mathbf{x}'_{it-q} \boldsymbol{\beta}_q + u_i + \varepsilon_{it}.$$

In first differences the model is

$$\Delta y_{it} = \alpha_1 \Delta y_{it-1} + \dots + \alpha_p \Delta y_{it-p} + \Delta \mathbf{x}'_{it} \boldsymbol{\beta}_1 + \dots + \Delta \mathbf{x}'_{it-q} \boldsymbol{\beta}_q + \Delta \varepsilon_{it}.$$

A sufficient set of instruments for the regressors are  $(\mathbf{x}_{it-1}, \Delta \mathbf{x}_{it-1}, \dots, \Delta \mathbf{x}_{it-q})$ , or equivalently  $(\mathbf{x}_{it-1}, \mathbf{x}_{it-2}, \dots, \mathbf{x}_{it-q-1})$ .

In many cases it is more reasonable to assume that  $\mathbf{x}_{it-1}$  is predetermined but not  $\mathbf{x}_{it}$ , since  $\mathbf{x}_{it}$  and  $\varepsilon_{it}$  may be endogenous. This, for example, is the standard assumption in vector autoregressions. In this case the estimation method is modified to use the instruments  $(\mathbf{x}_{it-2}, \mathbf{x}_{it-3}, \dots, \mathbf{x}_{it-q-1})$ . While this weakens the exogeneity assumption it also weakens the instrument set, as now the reduced form uses the second lag  $\mathbf{x}_{it-2}$  to predict  $\Delta \mathbf{x}_{it}$ .

The advantage obtained by treating a regressor as predetermined (rather than strictly exogenous) is that it is a substantial relaxation of the dynamic assumptions. Otherwise the parameter estimates will be inconsistent due to endogeneity.

The major disadvantage of treating a regressor as predetermined is that it substantially reduces the strength of identification, especially when the predetermined regressors are highly persistent.

In Stata, the `xtabond` command by default treats independent regressors as strictly exogenous. To treat the regressors as predetermined, use the option `pre`. By default all regressor lags are used as instruments, but the number can be limited if specified.

## 17.42 Blundell-Bond Estimator

Arellano and Bover (1995) and Blundell and Bond (1998) introduced a set of orthogonality conditions which reduce the weak instrument problem discussed in the Section 17.40 and improve performance in finite samples.

Consider the levels AR(1) model with no regressors (17.83)

$$y_{it} = \alpha y_{it-1} + u_i + \varepsilon_{it}.$$

Recall, least squares (pooled) regression is inconsistent because the regressor  $y_{it-1}$  is correlated with the error  $u_i$ . This raises the question: Is there an instrument  $z_{it}$  which solves this problem, in the sense that  $z_{it}$  is correlated with  $y_{it-1}$  yet uncorrelated with  $u_i + \varepsilon_{it}$ ? Blundell-Bond propose the instrument

$\Delta y_{it-1}$ . Clearly,  $\Delta y_{it-1}$  and  $y_{it-1}$  are correlated, so  $\Delta y_{it-1}$  satisfies the relevance condition. Also,  $\Delta y_{it-1}$  is uncorrelated with the idiosyncratic error  $\varepsilon_{it}$  when the latter is serially uncorrelated. Thus the key to the Blundell-Bond instrument is whether or not

$$\mathbb{E}(\Delta y_{it-1} u_i) = 0. \quad (17.101)$$

Blundell and Bond (1998) show that a sufficient condition for (17.101) is

$$\mathbb{E}\left(\left(y_{i1} - \frac{u_i}{1-\alpha}\right) u_i\right) = 0. \quad (17.102)$$

Recall that  $u_i/(1-\alpha)$  is the conditional mean of  $y_{it}$  under stationarity. Condition (17.102) states that the deviation of the initial condition  $y_{i1}$  from this conditional mean is uncorrelated with the individual effect  $u_i$ . Condition (17.102) is implied by stationarity, but is somewhat weaker.

To see that (17.102) implies (17.101), by applying recursion to (17.88) we find that

$$\Delta y_{it-1} = \alpha^{t-3} \Delta y_{i2} + \sum_{j=0}^{t-3} \Delta \varepsilon_{it-1-j}.$$

Hence

$$\begin{aligned} \mathbb{E}(\Delta y_{it-1} u_i) &= \alpha^{t-3} \mathbb{E}(\Delta y_{i2} u_i) \\ &= \alpha^{t-3} \mathbb{E}((\alpha - 1) y_{i1} + u_i + \varepsilon_{it}) u_i \\ &= \alpha^{t-3} (\alpha - 1) \mathbb{E}\left(\left(y_{i1} - \frac{u_i}{1-\alpha}\right) u_i\right) \\ &= 0 \end{aligned}$$

under (17.102), as claimed.

Now consider the full model (17.82) with predetermined regressors. Consider the assumption that the regressors have constant correlation with the individual effect

$$\mathbb{E}(\mathbf{x}_{it} u_i) = \mathbb{E}(\mathbf{x}_{is} u_i)$$

for all  $s$ . This implies

$$\mathbb{E}(\Delta \mathbf{x}_{it} u_i) = 0 \quad (17.103)$$

which means that the differenced predetermined regressors  $\Delta \mathbf{x}_{it}$  can also be used as instruments for the level equation.

Using (17.101) and (17.103), Blundell and Bond propose the following moment conditions for GMM estimation

$$\mathbb{E}(\Delta y_{it-1} (y_{it} - \alpha_1 y_{it-1} - \cdots - \alpha_p y_{it-p} - \mathbf{x}'_{it} \boldsymbol{\beta})) = 0 \quad (17.104)$$

$$\mathbb{E}(\Delta \mathbf{x}_{it} (y_{it} - \alpha_1 y_{it-1} - \cdots - \alpha_p y_{it-p} - \mathbf{x}'_{it} \boldsymbol{\beta})) = 0 \quad (17.105)$$

for  $t = p+2, \dots, T$ . Notice that these are for the levels (undifferenced) equation, while the Arellano-Bond (17.91) moments are in the differenced equation (17.88). We can write (17.104)-(17.105) in vector notation if we set  $\mathbf{Z}_{2i} = \text{diag}(\Delta y_{i2}, \dots, \Delta y_{iT-1}, \Delta \mathbf{x}_{i3}, \dots, \Delta \mathbf{x}_{iT})$ . Then (17.104)-(17.105) equals

$$\mathbb{E}(\mathbf{Z}_{2i} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\theta})) = 0. \quad (17.106)$$

Blundell and Bond proposed combining the  $\ell$  Arellano-Bond moments with the levels moments. This can be done by stacking the moment conditions (17.91) and (17.106). Recall from Section 17.39 the variables  $\Delta \mathbf{y}_i$ ,  $\Delta \mathbf{X}_i$ , and  $\mathbf{Z}_i$ . Now, define the stacked variables  $\bar{\mathbf{y}}_i = (\Delta \mathbf{y}'_i, \mathbf{y}'_i)'$ ,  $\bar{\mathbf{X}}_i = (\Delta \mathbf{X}'_i, \mathbf{X}'_i)'$  and  $\bar{\mathbf{Z}}_i = \text{diag}(\mathbf{Z}_i, \mathbf{Z}_{2i})$ . The stacked moment conditions are

$$\mathbb{E}(\bar{\mathbf{Z}}_i (\bar{\mathbf{y}}_i - \bar{\mathbf{X}}_i \boldsymbol{\theta})) = 0.$$

The Blundell-Bond estimator is found by applying GMM to this equation. They call this a systems GMM estimator. Let  $\bar{\mathbf{y}}$ ,  $\bar{\mathbf{X}}$  and  $\bar{\mathbf{Z}}$  denote  $\bar{\mathbf{y}}_i$ ,  $\bar{\mathbf{X}}_i$ , and  $\bar{\mathbf{Z}}_i$  stacked into matrices. Define  $\bar{\mathbf{H}} = \text{diag}(\mathbf{H}, \mathbf{I}_{T-2})$  where  $\mathbf{H}$  is from (17.31) and set

$$\hat{\boldsymbol{\Omega}}_1 = \sum_{i=1}^N \bar{\mathbf{Z}}_i' \bar{\mathbf{H}} \bar{\mathbf{Z}}_i.$$

The **Blundell-Bond one-step GMM estimator** is

$$\hat{\boldsymbol{\theta}}_1 = \left( \bar{\mathbf{X}}' \bar{\mathbf{Z}} \hat{\boldsymbol{\Omega}}_1^{-1} \bar{\mathbf{Z}}' \bar{\mathbf{X}} \right)^{-1} \left( \bar{\mathbf{X}}' \bar{\mathbf{Z}} \hat{\boldsymbol{\Omega}}_1^{-1} \bar{\mathbf{Z}}' \bar{\mathbf{y}} \right). \quad (17.107)$$

The systems residuals are  $\hat{\boldsymbol{\varepsilon}}_i = \bar{\mathbf{y}}_i - \bar{\mathbf{X}}_i \hat{\boldsymbol{\theta}}_1$ . A robust covariance matrix estimator is

$$\hat{\mathbf{V}}_1 = \left( \bar{\mathbf{X}}' \bar{\mathbf{Z}} \hat{\boldsymbol{\Omega}}_1^{-1} \bar{\mathbf{Z}}' \bar{\mathbf{X}} \right)^{-1} \left( \bar{\mathbf{X}}' \bar{\mathbf{Z}} \hat{\boldsymbol{\Omega}}_1^{-1} \bar{\mathbf{Z}}' \hat{\boldsymbol{\Omega}}_2 \bar{\mathbf{Z}} \hat{\boldsymbol{\Omega}}_1^{-1} \bar{\mathbf{Z}}' \bar{\mathbf{X}} \right) \left( \bar{\mathbf{X}}' \bar{\mathbf{Z}} \hat{\boldsymbol{\Omega}}_1^{-1} \bar{\mathbf{Z}}' \bar{\mathbf{X}} \right)^{-1} \quad (17.108)$$

where

$$\hat{\boldsymbol{\Omega}}_2 = \sum_{i=1}^N \bar{\mathbf{Z}}_i' \hat{\boldsymbol{\varepsilon}}_i \hat{\boldsymbol{\varepsilon}}_i' \bar{\mathbf{Z}}_i.$$

The **Blundell-Bond two-step GMM estimator** is

$$\hat{\boldsymbol{\theta}}_2 = \left( \bar{\mathbf{X}}' \bar{\mathbf{Z}} \hat{\boldsymbol{\Omega}}_2^{-1} \bar{\mathbf{Z}}' \bar{\mathbf{X}} \right)^{-1} \left( \bar{\mathbf{X}}' \bar{\mathbf{Z}} \hat{\boldsymbol{\Omega}}_2^{-1} \bar{\mathbf{Z}}' \bar{\mathbf{y}} \right). \quad (17.109)$$

The two-step systems residuals are  $\hat{\boldsymbol{\varepsilon}}_i = \bar{\mathbf{y}}_i - \bar{\mathbf{X}}_i \hat{\boldsymbol{\theta}}_2$ . A robust covariance matrix estimator is

$$\hat{\mathbf{V}}_2 = \left( \bar{\mathbf{X}}' \bar{\mathbf{Z}} \hat{\boldsymbol{\Omega}}_2^{-1} \bar{\mathbf{Z}}' \bar{\mathbf{X}} \right)^{-1} \left( \bar{\mathbf{X}}' \bar{\mathbf{Z}} \hat{\boldsymbol{\Omega}}_2^{-1} \bar{\mathbf{Z}}' \hat{\boldsymbol{\Omega}}_3 \bar{\mathbf{Z}} \hat{\boldsymbol{\Omega}}_2^{-1} \bar{\mathbf{Z}}' \bar{\mathbf{X}} \right) \left( \bar{\mathbf{X}}' \bar{\mathbf{Z}} \hat{\boldsymbol{\Omega}}_2^{-1} \bar{\mathbf{Z}}' \bar{\mathbf{X}} \right)^{-1} \quad (17.110)$$

where

$$\hat{\boldsymbol{\Omega}}_3 = \sum_{i=1}^N \bar{\mathbf{Z}}_i' \hat{\boldsymbol{\varepsilon}}_i \hat{\boldsymbol{\varepsilon}}_i' \bar{\mathbf{Z}}_i.$$

Asymptotically,  $\hat{\mathbf{V}}_2$  is equivalent to

$$\tilde{\mathbf{V}}_2 = \left( \bar{\mathbf{X}}' \bar{\mathbf{Z}} \hat{\boldsymbol{\Omega}}_2^{-1} \bar{\mathbf{Z}}' \bar{\mathbf{X}} \right)^{-1}. \quad (17.111)$$

The GMM estimator can be iterated until convergence to produce an iterated GMM estimator.

Simulation experiments reported in Blundell and Bond (1998) indicate that their systems GMM estimator performs substantially better than the Arellano-Bond estimator, especially when  $\alpha$  is close to one or the variance ratio  $\sigma_u^2/\sigma_\varepsilon^2$  is large. The explanation is that the orthogonality condition (17.104) does not suffer the weak instrument problem in these cases.

The advantage of the Blundell-Bond estimator is that the added orthogonality condition (17.104) greatly improves performance relative to the Arellano-Bond estimator when the latter is weakly identified. A disadvantage of the Blundell-Bond estimator is that their orthogonality condition is justified by a stationarity condition (17.102), and violation of the latter may induce estimation bias.

The advantages and disadvantages of the one-step versus two-step Blundell-Bond estimators are the same as described for the Arellano-Bond estimator as described in Section 17.39. Also as described there, when  $T$  is large it may be desired to limit the number of lags to use as instruments in order to avoid the many weak instruments problem.

The Blundell-Bond estimator may be obtained in Stata using either the `xtdpdsys` or `xtdpd` command. The default setting is the one-step estimator (17.107) and non-robust standard errors. For the two-step estimator and robust standard errors use the `twostep vce(robust)` options. Reported standard errors in Stata are based on Windmeijer's (2005) finite-sample correction to the asymptotic estimate (17.111). The robust covariance matrix estimator (17.110) nor the iterated GMM estimator are implemented.

### 17.43 Forward Orthogonal Transformation

Arellano and Bover (1995) proposed an alternative transformation to first differencing which eliminates the individual-specific effect and may have advantages in dynamic panel models. The **forward orthogonal transformation** is

$$y_{it}^* = c_{it} \left( y_{it} - \frac{1}{T_i - t} (y_{i,t+1} + \dots + y_{iT_i}) \right) \quad (17.112)$$

where  $c_{it}^2 = (T_i - t) / (T_i - t + 1)$ . This can be applied to all but the final observation (which is lost). Essentially,  $y_{it}^*$  subtracts from  $y_{it}$  the average of the remaining values, and then rescales so that the variance is constant under the assumption of homoskedastic errors.

At the level of the individual this can be written as

$$\mathbf{y}_i^* = \mathbf{A}_i \mathbf{y}_i$$

where  $\mathbf{A}_i$  is the  $(T_i - 1) \times T_i$  orthogonal deviation operator

$$\mathbf{A}_i = \text{diag}\left(\frac{T_i - 1}{T_i}, \dots, \frac{1}{2}\right) \begin{bmatrix} 1 & -\frac{1}{T_i-1} & -\frac{1}{T_i-1} & \cdots & -\frac{1}{T_i-1} & -\frac{1}{T_i-1} & -\frac{1}{T_i-1} \\ 0 & 1 & -\frac{1}{T_i-2} & \cdots & -\frac{1}{T_i-2} & -\frac{1}{T_i-2} & -\frac{1}{T_i-2} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{bmatrix}.$$

Important properties of the matrix  $\mathbf{A}_i$  are that  $\mathbf{A}_i \mathbf{1}_i = 0$  (so it eliminates individual effects),  $\mathbf{A}_i' \mathbf{A}_i = \mathbf{M}_i$ , and  $\mathbf{A}_i \mathbf{A}_i' = \mathbf{I}_{T_i-1}$ . These can be verified by direct multiplication.

Applying the transformation  $\mathbf{A}_i$  to (17.82) we obtain

$$y_{it}^* = \alpha_1 y_{it-1}^* + \dots + \alpha_p y_{it-p}^* + \mathbf{x}_{it}' \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i^*. \quad (17.113)$$

for  $t = p + 1, \dots, T - 1$ . This is equivalent to first differencing (17.88) when  $T = 3$  but differs for  $T > 3$ .

What is special about the transformed equation (17.113) is that under the assumption that  $\varepsilon_{it}$  are serially correlated and homoskedastic, the error  $\boldsymbol{\varepsilon}_i^*$  has variance  $\sigma_\varepsilon^2 \mathbf{A}_i \mathbf{A}_i' = \sigma_\varepsilon^2 \mathbf{I}_{T_i-1}$ . This means that  $\boldsymbol{\varepsilon}_i^*$  has the same covariance structure as  $\boldsymbol{\varepsilon}_i$ . Thus the orthogonal transformation operator eliminates the fixed effect while preserving the covariance structure. This is in contrast to (17.88) which has serially correlated errors  $\Delta \boldsymbol{\varepsilon}_i$ .

The transformed error  $\boldsymbol{\varepsilon}_i^*$  is a function of  $\varepsilon_{it}, \varepsilon_{it+1}, \dots, \varepsilon_{iT}$ . Thus valid instruments are  $y_{it-1}, y_{it-2}, \dots$ . Using the instrument matrix  $\mathbf{Z}_i$  from (17.90) in the case of strictly exogenous regressors or (17.100) with predetermined regressors, the  $\ell$  moment conditions can be written using matrix notation as

$$\mathbb{E}(\mathbf{Z}_i' (\mathbf{y}_i^* - \mathbf{X}_i^* \boldsymbol{\theta})) = 0. \quad (17.114)$$

Define the  $\ell \times \ell$  covariance matrix

$$\boldsymbol{\Omega} = \mathbb{E}(\mathbf{Z}_i' \boldsymbol{\varepsilon}_i^* \boldsymbol{\varepsilon}_i^{*\prime} \mathbf{Z}_i).$$

If the errors  $\varepsilon_{it}$  are conditionally homoskedastic then  $\boldsymbol{\Omega} = \mathbb{E}(\mathbf{Z}_i' \mathbf{Z}_i) \sigma_\varepsilon^2$ . Thus an asymptotically efficient GMM estimator is 2SLS applied to the orthogonalized equation using  $\mathbf{Z}_i$  as an instrument. In matrix notation,

$$\hat{\boldsymbol{\theta}}_1 = \left( \mathbf{X}^{*\prime} \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X}^* \right)^{-1} \left( \mathbf{X}^{*\prime} \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{y}^* \right).$$

This is the one-step GMM estimator.

Given the residuals  $\hat{\boldsymbol{\varepsilon}}_i = \mathbf{y}_i^* - \mathbf{X}_i^* \hat{\boldsymbol{\theta}}_1$  the two-step GMM estimator which is robust to heteroskedasticity and arbitrary serial correlation is

$$\hat{\boldsymbol{\theta}}_2 = \left( \mathbf{X}^{*\prime} \mathbf{Z} \hat{\boldsymbol{\Omega}}_2^{-1} \mathbf{Z}' \mathbf{X}^* \right)^{-1} \left( \mathbf{X}^{*\prime} \mathbf{Z} \hat{\boldsymbol{\Omega}}_2^{-1} \mathbf{Z}' \mathbf{y}^* \right)$$

where

$$\widehat{\Omega}_2 = \sum_{i=1}^N \mathbf{Z}'_i \widehat{\boldsymbol{\epsilon}}_i \widehat{\boldsymbol{\epsilon}}'_i \mathbf{Z}_i.$$

Standard errors for  $\widehat{\boldsymbol{\theta}}_1$  and  $\widehat{\boldsymbol{\theta}}_2$  can be obtained using cluster-robust methods.

Forward orthogonalization may have advantages over first differencing. First, the equation errors in (17.113) have a scalar covariance structure under i.i.d. idiosyncratic errors, which is expected to improve estimation precision. It also implies that the one-step estimator is 2SLS rather than GMM. Second, while there has not been a formal analysis of the weak instrument properties of the estimators after forward orthogonalization, it appears that if  $T > p + 2$  the method is less affected by weak instruments than first differencing. The disadvantages of forward orthogonalization are that it treats early observations asymmetrically from late observations, it is less thoroughly studied than first differencing, and is not available with several popular estimation methods.

The Stata command `xtdpd` includes forward orthogonalization as an option, but not when levels (Blundell-Bond) instruments are included or if there are gaps in the data. An alternative is the downloadable Stata package `xtabond2`.

## 17.44 Empirical Illustration

We illustrate the dynamic panel methods with the investment model (17.3). Estimates from two models are presented in Table 17.3. Both are estimated by Blundell-Bond two-step GMM with lags 2 through 6 as instruments, a cluster-robust weight matrix, and clustered standard errors.

The first column presents estimates of an AR(2) model. The estimates show that the series has a moderate amount of positive serial correlation, but appears to be well modeled as an AR(1) as the AR(2) coefficient is close to zero. This pattern of serial correlation is consistent with the presence of investment projects which span two years.

The second column presents estimates of the dynamic version of the investment regression (17.3), excluding the trading indicator. Two lags are included of the dependent variable and each regressor. The regressors are treated as predetermined, in contrast to the fixed effects regressions which treated the regressors as strictly exogenous. The regressors not contemporaneous with the dependent variable, but lagged one and two periods. This is done so that they are valid predetermined variables. Contemporaneous variables are likely endogenous so should not be treated as predetermined.

The estimates from the second column of Table 17.3 complement the earlier results. The evidence shows that investment has a moderate degree of serial dependence, is positively related to the first lag of  $Q$ , and is negatively related to lagged debt. Investment appears to be positively related to change in cash flow, rather than the level. Thus an increase in cash flow in year  $t - 1$  leads to investment in year  $t$ .

Table 17.3: Estimates of Dynamic Investment Equation

	AR(2)	AR(2) with Regressors
$I_{it-1}$	0.3191 (0.0172)	0.2519 (0.0220)
$I_{it-2}$	0.0309 (0.0112)	0.0137 (0.0125)
$Q_{it-1}$		0.0018 (0.0007)
$Q_{it-2}$		-0.0000 (0.0003)
$D_{it-1}$		-0.0154 (0.0058)
$D_{it-2}$		-0.0043 (0.0054)
$CF_{it-1}$		0.0400 (0.0091)
$CF_{it-2}$		-0.0290 (0.0051)

Two-step GMM estimates. Cluster-robust standard errors in parenthesis.

All regressions include time effects. GMM instruments include lags 2 through 6.

## Exercises

### Exercise 17.1

- (a) Show (17.11) and (17.12).
- (b) Show (17.13).

**Exercise 17.2** Is  $\mathbb{E}(\varepsilon_{it} | \mathbf{x}_{it}) = 0$  sufficient for  $\hat{\boldsymbol{\beta}}_{\text{fe}}$  to be unbiased for  $\boldsymbol{\beta}$ ? Explain why or why not.

**Exercise 17.3** Show that  $\text{var}(\dot{x}_{it}) \leq \text{var}(x_{it})$ .

**Exercise 17.4** Show (17.24).

**Exercise 17.5** Show (17.28).

**Exercise 17.6** Show that when  $T = 2$  the differenced estimator equals the fixed effects estimator.

**Exercise 17.7** In Section 17.14 it is described how to estimate the individual-effect variance  $\sigma_u^2$  using the between residuals. Develop an alternative estimator of  $\sigma_u^2$  only using the fixed effects error variance  $\hat{\sigma}_\epsilon^2$  and the levels error variance  $\hat{\sigma}_e^2 = n^{-1} \sum_{i=1}^N \sum_{t \in S_i} \hat{\epsilon}_{it}^2$  where  $\hat{\epsilon}_{it} = y_{it} - \mathbf{x}'_{it} \hat{\boldsymbol{\beta}}_{\text{fe}}$  are computed from the levels variables.

**Exercise 17.8** Verify that  $\hat{\sigma}_\epsilon^2$  defined in (17.37) is unbiased for  $\sigma_\epsilon^2$  under (17.18), (17.25) and (17.26).

**Exercise 17.9** Develop a version of Theorem 17.2 for the differenced estimator  $\hat{\boldsymbol{\beta}}_\Delta$ . Can you weaken Assumption 17.2.3? State an appropriate version which is sufficient for asymptotic normality.

**Exercise 17.10** Show (17.57).

### Exercise 17.11

- (a) For  $\hat{\sigma}_i^2$  defined in (17.59) show  $\mathbb{E}(\hat{\sigma}_i^2 | \mathbf{X}_i) = \bar{\sigma}_i^2$ .
- (b) For  $\tilde{\mathbf{V}}_{\text{fe}}$  defined in (17.58) show  $\mathbb{E}(\tilde{\mathbf{V}}_{\text{fe}} | \mathbf{X}) = \mathbf{V}_{\text{fe}}$ .

### Exercise 17.12

- (a) Show (17.61).
- (b) Show (17.62).
- (c) For  $\tilde{\mathbf{V}}_{\text{fe}}$  defined in (17.60) show  $\mathbb{E}(\tilde{\mathbf{V}}_{\text{fe}} | \mathbf{X}) = \mathbf{V}_{\text{fe}}$ .

**Exercise 17.13** Take the fixed effects model  $y_{it} = x_{it}\beta_1 + x_{it}^2\beta_2 + u_i + \varepsilon_{it}$ . A researcher estimates the model by first obtaining the within transformed  $\dot{y}_{it}$  and  $\dot{x}_{it}$  and then regressing  $\dot{y}_{it}$  on  $\dot{x}_{it}$  and  $\dot{x}_{it}^2$ . Is the correct estimation method? If not, describe the correct fixed effects estimator.

**Exercise 17.14** In Section 17.33, verify that in the just-identified case, the 2SLS estimator  $\hat{\boldsymbol{\beta}}_{2\text{sls}}$  simplifies as claimed:  $\hat{\boldsymbol{\beta}}_1$  and  $\hat{\boldsymbol{\beta}}_2$  are the fixed effects estimator.  $\hat{\boldsymbol{\gamma}}_1$  and  $\hat{\boldsymbol{\gamma}}_2$  equal the 2SLS estimator from a regression of  $\hat{\mathbf{u}}$  on  $\mathbf{Z}_1$  and  $\mathbf{Z}_2$ , using  $\bar{\mathbf{X}}_1$  as an instrument for  $\mathbf{Z}_2$ .

**Exercise 17.15** In this exercise you will replicate and extend the empirical work reported in Arellano and Bond (1991) and Blundell and Bond (1998). Arellano-Bond gathered a dataset of 1031 observations from an unbalanced panel of 140 U.K. companies for 1976-1984, and is in the datafile AB1991 on the textbook webpage. The variables we will be using are log employment (n), log real wages (w), and log capital (k). See the description file for definitions.

- (a) Estimate the panel AR(1)

$$k_{it} = \alpha k_{it-1} + u_i + v_t + \varepsilon_{it}$$

using Arellano-Bond one-step GMM with clustered standard errors. Note that the model includes year fixed effects.

- (b) Re-estimate using Blundell-Bond one-step GMM with clustered standard errors.  
(c) Explain the difference in the estimates.

**Exercise 17.16** This exercise uses the same dataset as the previous question. Blundell and Bond (1998) estimated a dynamic panel regression of log employment  $n$  on log real wages  $w$  and log capital  $k$ . The following specification<sup>1</sup> used the Arellano-Bond one-step estimator, treating  $w_{it-1}$  and  $k_{it-1}$  as predetermined

$$\begin{aligned} n_{it} = & .7075 \quad n_{it-1} - .7088 \quad w_{it} + .5000 \quad w_{it-1} + .4660 \quad k_{it} - .2151 \quad k_{it-1}. \\ & (.0842) \quad \quad \quad (.1171) \quad \quad \quad (.1113) \quad \quad \quad (.1010) \quad \quad \quad (.0859) \end{aligned} \quad (17.115)$$

This equation also included year dummies, and the standard errors are clustered.

- (a) Estimate (17.115) using the Arellano-Bond one-step estimator treating  $w_{it}$  and  $k_{it}$  as strictly exogenous.  
(b) Estimate (17.115) treating  $w_{it-1}$  and  $k_{it-1}$  as predetermined to verify the results in (17.115). What is the difference between the estimates treating the regressors as strictly exogenous versus predetermined?  
(c) Estimate the equation using the Blundell-Bond one-step systems GMM estimator.  
(d) Interpret the coefficient estimates viewing (17.115) as a firm-level labor demand equation.  
(e) Describe the impact on the standard errors of the Blundell-Bond estimates in part (c) if you forget to use clustering. (You do not have to list all the standard errors, but describe the magnitude of the impact.)

**Exercise 17.17** Use the datafile Invest1993 on the textbook webpage. You will be estimating the panel AR(1)

$$D_{it} = \alpha D_{it-1} + u_i + \varepsilon_{it}$$

for  $D = \text{debt/assets}$  (this is `debta` in the datafile). See the description file for definitions.

- (a) Estimate the above autoregression using Arellano-Bond two-step GMM with clustered standard errors.  
(b) Re-estimate using Blundell-Bond two-step GMM.  
(c) Experiment with your results, trying two-step versus one-step, AR(1) versus AR(2), number of lags used as instruments, and classical versus robust standard errors. What makes the most difference for the coefficient estimates? For the standard errors?

**Exercise 17.18** Use the datafile Invest1993 on the textbook webpage. You will be estimating the model

$$D_{it} = \alpha D_{it-1} + \beta_1 I_{it-1} + \beta_2 Q_{it-1} + \beta_3 C_{it-1} + u_i + \varepsilon_{it}.$$

The variables are `debta`, `inva`, `vala`, and `cfa` in the datafile). See the description file for definitions.

---

<sup>1</sup> Blundell and Bond (1998), Table 4, column 3.

- (a) Estimate the above regression using Arellano-Bond twostep GMM with clustered standard errors, treating all regressors as predetermined.
- (b) Re-estimate using Blundell-Bond twostep GMM, treating all regressors as predetermined.
- (c) Experiment with your results, trying twostep versus onestep, number of lags used as instruments, and classical versus robust standard errors. What makes the most difference for the coefficient estimates? For the standard errors?

# Chapter 18

## Difference in Differences

### 18.1 Introduction

One of the most popular methods to estimate the effect of a policy change is by the method of difference in differences, often called “diff in diffs”. Estimation is typically a two-way panel data regression with a policy indicator as a regressor. Clustered variance estimation is generally recommended for inference.

In order to interpret a difference in difference estimate as a policy effect there are three key conditions. First, that the estimated regression is the correct conditional mean. In particular, this requires that all trends and interactions are properly included. Second, that the policy is exogenous – it satisfies conditional independence. Third, there are no other relevant unincluded factors coincident with the policy change. If these assumptions are satisfied then the difference in difference estimand is a valid causal effect.

### 18.2 Minimum Wage in New Jersey

The most well known application of the difference in difference methodology is Card and Krueger (1994) who investigated the impact of New Jersey's 1992 increase of the minimum hourly wage from \$4.25 to \$5.05. Classical economics teaches that an increase in the minimum wage will lead to decreases in employment and increases in prices. To investigate the magnitude of this impact the authors surveyed a panel of 331 fast food restaurants in New Jersey during the period 2/15/1992-3/4/1992 (before the enactment of the minimum wage increase) and then again during the period 11/5/1992-12/31/1992 (after the enactment). Fast food restaurants were selected for investigation as they are a major employer of minimum wage employees. (Before the change about 30% of the sampled workers were paid the minimum wage of \$4.25).

Table 18.1: Average Employment at Fast Food Restaurants

	New Jersey	Pennsylvania	Difference
Before Increase	20.43	23.38	2.95
After Increase	20.90	21.10	0.20
Difference	0.47	-2.28	<b>2.75</b>

The data file CK1994 is extracted from the original Card-Krueger data set and is posted on the textbook webpage.

Table 18.1 (first column) displays the mean number<sup>1</sup> of full-time equivalent employees<sup>2</sup> at New Jer-

<sup>1</sup>Our calculations drop restaurants if they were missing the number of full-type equivalent employees in either survey.

<sup>2</sup>Following Card and Krueger full-time equivalent employees is defined as the sum of the number of full-time employees, one-half of the number of part-time employees, and the number of managers and assistant managers.

sey fast food restaurants before and after the minimum wage increase. Before the increase the average number of employees was 20.4. After the increase the average number of employees was 20.9. Contrary to the predictions of conventional theory, employment slightly increased (by 0.5 employees per restaurant) rather than decreased.

This estimate – the change in employment – could be called a **difference estimator**. It is the change in employment coincident with the change in policy. A difficulty in interpretation is that all employment change is attributed to the policy. It does not provide direct evidence of the counterfactual – what would have happened if the minimum wage had not been increased.

A **difference in difference estimator** improves on a difference estimator by comparing the change in the treatment sample with a comparable change in a control sample.

Card and Krueger selected eastern Pennsylvania for their control sample. The minimum wage was constant at \$4.25 an hour in the state of Pennsylvania during 1992. At the beginning of the year starting wages at fast food restaurants in the two states were very similar. The two areas (New Jersey and eastern Pennsylvania) share further similarities. Any trends or economic shocks which affect one state are likely to affect both. Therefore Card and Krueger argued that it is appropriate to treat eastern Pennsylvania as a control. This means that in the absence of a minimum wage increase they expected the same changes in employment to occur in both New Jersey and eastern Pennsylvania.

Card and Krueger surveyed a panel of 79 fast food restaurants in eastern Pennsylvania, simultaneously while surveying the New Jersey restaurants. The average number of full-time equivalent employees is displayed in the second column of Table 18.1. Before the policy change the average number of employees was 23.4. After the policy change the average number was 21.1. Thus in Pennsylvania average employment decreased by 2.3 employees per restaurant.

Treating Pennsylvania as a control means comparing the change in New Jersey (0.5) with that in Pennsylvania (-2.3). The difference (2.75 employees per restaurant) is the difference-in-difference estimate of the impact of the minimum wage increase. In complete contradiction to conventional economic theory, the estimate indicates an increase in employment rather than a decrease. This surprising estimate has been widely discussed among economists<sup>3</sup> and the popular press.

It is constructive to re-write the estimates from Table 18.1 in regression format. Let  $y_{it}$  denote employment at restaurant  $i$  surveyed at time  $t$ . Let  $State_i$  be a dummy variable indicating the state, with  $State_i = 1$  for New Jersey and  $State_i = 0$  for Pennsylvania. Let  $Time_t$  be a dummy variable indicating the time period, with  $Time_t = 0$  for the period before the policy change and  $Time_t = 1$  for the period after the policy change. Let  $D_{it}$  denote a treatment dummy, with  $D_{it} = 1$  if the minimum wage equals \$5.05 and  $D_{it} = 0$  if the minimum wage equals \$4.25. In this application it equals the interaction dummy  $D_{it} = State_i Time_t$ .

Table 18.1 is a saturated regression in the two dummy variables and can therefore be written as the regression equation

$$y_{it} = \beta_0 + \beta_1 State_i + \beta_2 Time_t + \theta D_{it} + \varepsilon_{it}. \quad (18.1)$$

Indeed the coefficients can be written in terms of Table 18.1 by the following correspondence

	New Jersey	Pennsylvania	Difference
Before Increase	$\beta_0 + \beta_1$	$\beta_0$	$\beta_1$
After Increase	$\beta_0 + \beta_1 + \beta_2 + \theta$	$\beta_0 + \beta_2$	$\beta_1 + \theta$
Difference	$\beta_2 + \theta$	$\beta_2$	$\theta$

We see that the coefficients in the regression (18.1) correspond to interpretable difference and difference in difference estimands.  $\beta_1$  is the difference estimand of the effect of “New Jersey vs. Pennsylvania” in the period before the policy change.  $\beta_2$  is the difference estimand of the time effect in the control

<sup>3</sup>Most economists do not take the estimate literally – they do not believe that increasing the minimum wage will cause employment increases. Instead it has been interpreted as evidence that small changes in the minimum wage may have only minor impacts on employment levels.

state.  $\theta$  is the difference in difference estimand – the change in New Jersey relative to the change in Pennsylvania.

Our estimate of the regression (18.1) is

$$y_{it} = 23.4 - 2.9 \text{ } State_i - 2.3 \text{ } Time_t + 2.75 \text{ } D_{it} + \varepsilon_{it}. \quad (18.2)$$

(1.4)      (1.5)      (1.2)      (1.34)

The standard errors are calculated by clustering by restaurant. As expected the coefficient  $\hat{\theta}$  on the treatment dummy precisely corresponds to the difference in difference estimate from Table 18.1. The coefficient estimates can be interpreted as described. The estimated pre-change difference between New Jersey and Pennsylvania is  $-2.9$ , and the estimated time effect is  $-2.3$ , but neither is statistically significant. The estimated difference in difference estimate of  $2.75$ , on the other hand, is statistically significant.

Since the observations are divided into the groups  $State_i = 0$  and  $State_i = 1$ , and  $Time_t$  is equivalent to a time index, this regression is identical to a two-way fixed effects regression of  $y_{it}$  on  $D_{it}$  with state and time fixed effects. Furthermore, since the regressor  $D_{it}$  does not vary across individuals within the state, this fixed effects regression is unchanged if restaurant-level fixed effects are included instead of state fixed effects. (Restaurant fixed effects are orthogonal to any variable demeaned at the state level. See Exercise 18.1.) Thus the above regression is algebraically identical to the two-way fixed effects regression

$$y_{it} = \theta D_{it} + u_i + v_t + \varepsilon_{it} \quad (18.3)$$

where  $u_i$  is a restaurant fixed effect and  $v_t$  is a time fixed effect. The simplest method to implement this is by a one-way fixed effects regression with time dummies. The estimates are

$$y_{it} = 2.75 \text{ } D_{it} - 2.3 \text{ } Time_t + u_i + \varepsilon_{it} \quad (18.4)$$

(1.34)      (1.2)

which are identical to the previous regression.

Equation (18.3) is the basic difference-in-difference model. It is a two-way fixed effects regression of the response  $y_{it}$  on a binary policy  $D_{it}$ . The coefficient  $\theta$  corresponds to the double difference in sample means, and can be interpreted as the policy impact (also called the treatment effect) of  $D$  on  $y$ . (We discuss identification in the next section.) Our presentation (and the Card-Krueger example) focuses on the basic case of two aggregate units (states) and two time periods. The regression formulation (18.3) is convenient as it can be easily generalized to allow for multiple states and time periods. Doing so can provide more convincing evidence of an identified policy effect. The equation (18.3) can also be generalized by changing the trend specification, and by using a continuous treatment variable.

Another common generalization is to augment the regression with controls  $x_{it}$ . This model takes the form

$$y_{it} = \theta D_{it} + \mathbf{x}'_{it} \boldsymbol{\beta} + u_i + v_t + \varepsilon_{it}.$$

Many empirical studies report estimates both of the basic model and regressions with controls. For example we could augment the Card-Krueger regression to include the variable *hoursopen*, the number of hours a day the restaurant is open. A restaurant with longer hours will tend to have more employees.

$$y_{it} = 2.84 \text{ } D_{it} - 2.2 \text{ } Time_t + 1.2 \text{ } hoursopen_{it} + u_i + \varepsilon_{it}. \quad (1.31) \quad (1.2) \quad (0.4)$$

Indeed the estimated effect is that a restaurant employs an additional 1.2 employees for each hour open, and this effect is statistically significant. The estimated treatment effect is not meaningfully changed.

### 18.3 Identification

Consider the difference-in-difference equation

$$y_{it} = \theta D_{it} + \mathbf{x}'_{it} \boldsymbol{\beta} + u_i + v_t + \varepsilon_{it} \quad (18.5)$$

for  $i = 1, \dots, N$  and  $t = 1, \dots, T$ . We are interested in conditions under which the coefficient  $\theta$  is the causal impact of the treatment  $D_{it}$  on the outcome  $y_{it}$ . The answer can be found by applying Theorem 2.11 from Section 2.30.

In Section 2.30 we introduced the potential outcomes framework which writes the outcome as a function of the treatment, controls, and unobservables. Thus the outcome (e.g. employment at a restaurant) can be written as  $y = h(D, \mathbf{x}, \mathbf{e})$  where  $D$  is treatment (minimum wage policy),  $\mathbf{x}$  are controls, and  $\mathbf{e}$  is a vector of unobserved factors. Model (18.5) specifies that  $h(D, \mathbf{x}, \mathbf{e})$  is separable and linear in its arguments, and that the unobservables consist of individual-specific, time-specific, and idiosyncratic effects.

We now present sufficient conditions under which the coefficient  $\theta$  can be interpreted as a causal effect. Recall the two-way within transformation (17.65) and set  $\ddot{\mathbf{z}}_{it} = (\ddot{D}_{it}, \ddot{\mathbf{x}}'_{it})'$ .

**Theorem 18.1** Suppose the following conditions hold:

1.  $y_{it} = \theta D_{it} + \mathbf{x}'_{it} \boldsymbol{\beta} + u_i + v_t + \varepsilon_{it}$ .
2.  $\mathbb{E}(\ddot{\mathbf{z}}_{it} \ddot{\mathbf{z}}'_{it}) > 0$ .
3.  $\mathbb{E}(\mathbf{x}_{it} \varepsilon_{is}) = 0$  for all  $t$  and  $s$ .
4. Conditional on  $\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT}$  the random variables  $D_{it}$  and  $\varepsilon_{is}$  are statistically independent for all  $t$  and  $s$ .

Then the coefficient  $\theta$  in (18.5) equals the average causal effect for  $D$  on  $y$  conditional on  $\mathbf{x}$ .

Condition 1 states that the outcome equation equals the specified linear regression model, which is additively separable in the observables, individual effect, and time effect.

Condition 2 states that the two-way within transformed regressors have a non-singular design matrix. This requires that all elements of  $D_{it}$  and  $\mathbf{x}_{it}$  vary across time and individuals.

Condition 3 is the standard exogeneity assumption for regressors in a fixed-effects model.

Condition 4 states that the treatment variable is conditionally independent of the idiosyncratic error. This is the conditional independence assumption for fixed effects regression.

To show Theorem 18.1 apply the two-way within transformation (17.65) to (18.5). We obtain

$$\ddot{y}_{it} = \theta \ddot{D}_{it} + \ddot{\mathbf{x}}'_{it} \boldsymbol{\beta} + \ddot{\varepsilon}_{it}.$$

Under Condition 2 the projection coefficients  $(\theta, \boldsymbol{\beta})$  are uniquely defined and under Conditions 3 and 4 they equal the linear regression coefficients. Thus  $\theta$  is the regression derivative with respect to  $D$ . Condition 4 implies that conditional on  $\ddot{\mathbf{x}}_{it}$  the random variables  $\ddot{D}_{it}$  and  $\ddot{\varepsilon}_{is}$  are statistically independent. Theorem 2.11 shows that this implies that the regression derivative  $\theta$  equals the average causal effect as stated.

The assumption that  $D$  and  $\varepsilon$  are independent is the fundamental exogeneity assumption. To interpret  $\theta$  as a treatment effect it is important that  $D$  is defined as the treatment and not simply as an interaction (time and state) dummy. This is subtle. Examine equation (18.5) recalling that  $D$  is defined

as the treatment (an increase in the minimum wage). In this equation the error  $\varepsilon_{it}$  contains all variables and effects not included in the regression. Thus if there are other changes in New Jersey which are coincident with the minimum wage increase, the assumption that  $D$  and  $\varepsilon$  are independent means that those coincident changes are independent of  $\varepsilon$ , and thus do not affect employment. This is a strong assumption. Once again, Condition 4 states that all other effects which are coincident with the minimum wage increase have no effect on employment. Without this assumption it would not be possible to claim that the diff-in-diff regression identifies the causal effect of the treatment.

Furthermore, independence of  $D_{it}$  and  $\varepsilon_{is}$  means that neither can be affected by the other. This means that the policy (treatment) was not enacted in response to knowledge about the response variable in either period, and it means that the outcome (employment) did not change in the first period in anticipation of the upcoming policy change.

It is difficult to know if the exogeneity of  $D$  is a reasonable assumption. It is similar to instrument exogeneity in instrumental variable regression. Its validity hinges on a well-articulated structural argument. An empirical investigation based on a difference-in-difference specification needs to make an explicit case for exogeneity of  $D$  similar to that for IV regression.

In the case of the Card-Krueger application, the authors argue that the policy was exogenous because it was adopted two years before taking effect. At the time of the passage of the legislation the economy was in an expansion, but by the time of adoption the economy has slipped into recession. This suggests that it is credible to assume that the policy decision in 1990 was not affected by employment levels in 1992. Furthermore, concern about the impact of the increased minimum wage during a recession led to a serious discussion about reversing the policy, meaning that there was uncertainty about whether or not the policy would actually be enacted at the time of the first survey. It thus seems credible that employment decisions at that time were not determined in anticipation of the upcoming minimum wage increase.

The authors do not discuss, however, whether or not there were other coincident events in the New Jersey or Pennsylvania economies during 1992 which could have affected employment differentially in the two states. It seems plausible that there could have been many such coincident events. This seems to be the greatest weakness in their identification argument.

Identification (the conditions for Theorem 18.1) also requires that the regression model is correctly specified. These means that the true model is linear in the specified variables and all interactions are included. Since the basic  $2 \times 2$  specification is a saturated dummy variable model it is necessarily a conditional mean and thus correctly specified. This is not necessarily the case in applications with more than two states or time periods, and thus model specification needs to be carefully considered in such cases.

## 18.4 Multiple Units

The basic difference-in-difference model has two aggregate units (e.g. states) and two time periods. Additional information can be obtained if there are multiple units or multiple time periods. In this section we focus on the case of multiple units. There can be multiple treatment units, multiple control units, or both. In this section we suppose that the number of periods is  $T = 2$ . Let  $N_1 \geq 1$  be the number of untreated (control) units, and  $N_2 \geq 1$  be the number of treated units, with  $N = N_1 + N_2$ .

The basic regression model

$$y_{it} = \theta D_{it} + u_i + v_t + \varepsilon_{it}$$

imposes two strong restrictions. First, that all units are equally affected by time as  $v_t$  is common across  $i$ . Second, that the treatment effect  $\theta$  is common across all treated units.

The Card-Krueger data set only contains observations from two states, but the authors did record additional variables including the region of the state. They divided New Jersey into three regions (North, Central, and South) and eastern Pennsylvania into two regions (1 for northeast Philadelphia suburbs, and 2 for the remainder).

Table 18.2 displays the mean number of full-time equivalent employees by region, before and after the minimum wage increase. We observe that two of the three New Jersey regions had nearly identical increases in employment, and all three changes are small. We can also observe that both of the Pennsylvania regions had employment decreases, though with different magnitudes.

We can test the assumption of equal treatment effect  $\theta$  by a regression exclusion test. This can be done by adding interaction dummies to the regression and testing for the exclusion of the interactions. As there are three treated regions in New Jersey we include two of the three New Jersey region dummies interacted with the time index. In general we would include  $N_2 - 1$  such interactions. These coefficients measure the treatment effect difference across regions. Testing that these two coefficients are zero we obtain a p-value of 0.60 which as expected is far from significant. Thus we accept the hypothesis that the treatment effect  $\theta$  is common across the New Jersey regions.

In contrast, when the treatment effect  $\theta$  varies we call this a **heterogeneous treatment effect**. It is not a violation of the treatment effect framework, but it can be considerably more complicated to analyze. (A model which incorrectly imposes a homogeneous treatment effect is misspecified and produces inconsistent estimates.)

A more serious problem arises if the control effect is heterogeneous. The control effect is the change in the control group. Table 18.2 breaks down the estimated control effect across the two Pennsylvania regions. While both estimates are negative they are somewhat different from one another. If the effects are distinct there is not a homogeneous control effect. We can test the assumption of equal control effects by a regression exclusion test. As there are two Pennsylvania regions we include the interaction of one of the Pennsylvania regions with the time index. (In general we would include  $N_1 - 1$  interactions.) This coefficient measures the difference in the control effect across the regions. We test that this coefficient is zero, obtaining a t-statistic of 1.2 and a p-value of 0.23. It is also not statistically significant, meaning that we cannot reject the hypothesis that the control effect is homogeneous.

In contrast, if the control effect were heterogeneous then the difference-in-difference estimation strategy is misspecified. The method relies on the ability to identify a credible control sample. Therefore if a test for equal control effects rejects the hypothesis of homogeneous control effects, this should be taken as evidence against interpretation of the difference-in-difference parameter as a treatment effect.

Table 18.2: Average Employment at Fast Food Restaurants

	South NJ	Central NJ	North NJ	PA 1	PA 2
Before Increase	16.6	22.0	22.0	24.8	22.2
After Increase	17.3	21.4	22.7	21.0	21.2
Difference	0.7	-0.6	0.7	-3.8	-1.0

## 18.5 Do Police Reduce Crime?

DiTella and Schargrodsy (2004) use a difference-in-difference approach to study the question of whether the street presence of police officers reduces car theft. Rational crime models predict that the presence of an observable police force will reduce crime rates (at least locally) due to deterrence. The causal effect is difficult to measure, however, as police forces are not allocated exogenously, but rather are allocated in anticipation of need. A difference-in-difference estimator requires an exogenous event which changes police allocations. The innovation in DiTella-Schargrodsy was to use the police response to a terrorist attack as exogenous variation.

In July 1994 there was a horrific terrorist attack on the main Jewish center in Buenos Aires, Argentina. Within two weeks the federal government provided police protection to all Jewish and Muslim buildings in the country. DiTella and Schargrodsy (2004) hypothesized that their presence, while allocated to deter a terror or reprisal attack, would also deter other street crimes such as automobile theft locally to the deployed police. The authors collected detailed information on car thefts in selected neighborhoods

of Buenos Aires for April–December 1994, resulting in a panel for 876 city blocks. They hypothesized that the terrorist attack and the government’s response were exogenous to auto thievery and is thus a valid treatment. They postulated that the deterrence effect would be strongest for any city block which contained a Jewish institution (and thus police protection). Potential car thieves would be deterred from a burglary due to the threat of being caught. The deterrence effect was expected to weaken as the distance from the protected sites increased. The authors therefore proposed a difference-in-difference estimator based on the average number of car thefts per block, before and after the terrorist attack, and between city blocks with and without a Jewish institution. Their sample has 37 blocks with Jewish institutions (the treatment sample) and 839 blocks without an institution (the control sample).

The data file DS2004 is a slightly revised version of the author’s AER replication file and is posted on the textbook webpage.

Table 18.3: Number of Car Thefts by City Block

	Same Block	Not on Same Block	Difference
April–June	0.112	0.095	-0.017
August–December	0.035	0.105	0.070
Difference	-0.077	0.010	<b>-0.087</b>

Table 18.3 displays the average number of car thefts per block, separately for the months before the July attack and the months after the July attack, and separately for city blocks which have a Jewish institution (and therefore received police protection starting in late July) and for other city blocks. We can see that the average number of car thefts dramatically decreased in the protected city blocks, from 0.112 per month to 0.035, while the average number in non-protected blocks was near-contant, rising from 0.095 to 0.105. Taking the difference in difference we find that the effect of police presence decreased car thefts by 0.087, which is about 78%.

The general way to estimate a diff-in-diff model of this form is as a regression of the form (18.3) where  $y_{it}$  is the number of car thefts on block  $i$  during month  $t$ , and  $u_i$  and  $v_t$  are block and month fixed effects. This regression<sup>4</sup> yields the same estimate of 0.087 since the panel is balanced and there are no control variables.

Table 18.4: Number of Car Thefts by City Block

		Same Block	Not on Same Block	Difference
Pre-Attack	April	0.112	0.110	-0.012
	May	0.088	0.100	0.012
	June	0.128	0.076	-0.052
Post-Attack	August	0.047	0.111	0.064
	September	0.014	0.099	0.085
	October	0.061	0.108	0.047
	November	0.027	0.100	0.073
	December	0.027	0.106	0.079

The model (18.3) makes the strong assumption that the treatment effect is constant across the five treated months. We investigate this assumption in Table 18.4 which breaks down the car thefts by month. For the control sample the number of car thefts is near constant across the months. For seven of the eight months the average number per block ranges from .10 to .11, with only one month (June) a bit lower at 0.08. In the treatment sample the average number of thefts per block in the three months before the terrorist attack are similar to the averages in the control sample. But in the five months following the attack the number of car thefts is uniformly reduced. The averages range from 0.014 to 0.061. In each

<sup>4</sup>We omit the observations for July as the car theft data is only for the first half of the month.

month after the attack the control sample has lower thefts, with averages ranging from 0.047 to 0.085. Given the small sample size (37) of the treatment sample this is strikingly uniform evidence.

We can formally test the homogeneity of the treatment effect by including four dummy variables for the interactions of four post-attack months with the treatment sample, and then testing the exclusion of these variables. The p-value for this test is 0.81, exceedingly far from significant. Thus there is no reason in the data to be suspicious of the homogeneity assumption.

The goal was to estimate the causal effect of police presence as a deterrence for crime. Let us evaluate the case for identification. It seems reasonable to treat the terrorist attack as exogenous. The government response also appears exogenous. Neither is reasonably related to the auto theft rate. We also observe that the evidence in Tables 18.3 and 18.4 indicate that theft rates were similar in the pre-attack treatment and control samples. Thus the additional police protection seems credibly provided for the purpose of attack prevention rather than as an excuse for crime prevention. The general homogeneity of the theft rate across months, once allowing for the treatment effect, gives credibility to the claim that the police response was a causal effect. The terror attack itself did not reduce car theft rates as there seems to be no measurable effect outside of the treatment sample. Finally, while the paper does not explicitly address whether or not there was any other coincident event in July 1994 which may have effected these specific city blocks, it is difficult to conceive of an alternative explanation for such a large effect. Our conclusion is that this is a very strong identification argument. Police presence greatly reduces the incidence of car theft.

The authors asserted the inference that police presence deters crime more broadly. This is a more tenuous extension as the paper does not provide direct evidence of such a claim. While it seem reasonable, we should be cautious about making generalizations without supporting evidence.

Overall, DiTella and Schargrodsy (2004) is an excellent example of a well-articulated and credibly identified difference-in-difference estimate of an important policy effect.

## 18.6 Trend Specification

Some applications (including the two introduced earlier in this chapter) apply to a short period of time such as one year, in which case we may not expect the variables to be trended. Other applications cover many years or decades, in which case the variables are likely to be trended. These trend can reflect long-term growth, business cycle effects, changing tastes, or many other features. If trends are incorrectly specified then the model will be misspecified, and the estimated policy effect will be inconsistent due to omitted variable bias. Consider the difference-in-difference equation

$$y_{it} = \theta D_{it} + \mathbf{x}'_{it} \boldsymbol{\beta} + u_i + v_t + \varepsilon_{it}.$$

This model imposes the strong assumption that the trends in  $y_{it}$  are entirely explained by the included controls  $\mathbf{x}_{it}$  and the common unobserved time component  $v_t$ . This can be quite restrictive. It is reasonable to expect that trends may differ across units and are not fully captured by observed controls.

One way to think about this is in terms of overidentification. For simplicity suppose there are no controls and the panel is balanced. Then there are  $NT$  observations. The two-way model with a policy effect has  $N + T$  coefficients. Unless  $N = T = 2$  this model is overidentified. In addition to considering heterogeneous treatment effects it is reasonable to consider heterogeneous trends.

One generalization is to include interactions of a linear time trend with a control variable. This model takes the form

$$y_{it} = \theta D_{it} + \mathbf{x}'_{it} \boldsymbol{\beta} + \mathbf{z}'_i \boldsymbol{\delta} t + u_i + v_t + \varepsilon_{it}.$$

It specifies that the trend in  $y_{it}$  differs across units depending on the controls  $\mathbf{z}_i$ .

A broader generalization is to include unit-specific linear time trends. This model takes the form

$$y_{it} = \theta D_{it} + \mathbf{x}'_{it} \boldsymbol{\beta} + u_i + v_t + t w_i + \varepsilon_{it}. \quad (18.6)$$

In this model  $w_i$  is a time trend fixed effect which varies across units. If there are no controls this model has  $2N + T$  coefficients, and is identified as long as  $T \geq 4$ .

Estimation of model (18.6) can be done one of three ways. If  $N$  is small (for example, applications with state-level data) then the regression can be estimated using the explicit dummy variable approach. Let  $d_i$  and  $S_t$  be dummy variables indicating the  $i^{th}$  unit and  $t^{th}$  time period. Set  $d_{it} = d_i t$ , the interaction of the individual dummy with the time trend. The equation is estimated by regression of  $y_{it}$  on  $D_{it}$ ,  $\mathbf{x}_{it}$ ,  $d_i$ ,  $S_t$ , and  $d_{it}$ . Equivalently, one can apply one-way fixed effects with regressors  $D_{it}$ ,  $\mathbf{x}_{it}$ ,  $S_t$ , and  $d_{it}$ .

When  $N$  is large a computationally more efficient approach is to use residual regression. For each unit  $i$ , estimate a time trend model for each variable  $y_{it}$ ,  $D_{it}$ ,  $\mathbf{x}_{it}$  and  $S_t$ . That is, for each  $i$  estimate

$$y_{it} = \hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{y}_{it}.$$

This is a generalized within transformation. The residuals  $\hat{y}_{it}$  are then used in place of the original observations. Regress  $\hat{y}_{it}$  on  $D_{it}$ ,  $\mathbf{x}_{it}$ , and  $S_t$  to obtain the estimates of (18.6).

The relevance of the trend fixed effects  $v_t$  can be assessed by a significance test. Specifically the hypothesis that the coefficients on the period dummies can be tested using a standard exclusion test. Similarly trend interaction terms can be tested for significance using standard exclusion tests. If the tests are statistically significant this indicates that their inclusion is relevant for correct specification. Unfortunately the unit-specific linear time trends cannot be tested for significance when the covariance matrix is clustered at the unit level. This is similar to the problem of testing the significance of a dummy variable with a single observation. The unit-specific time trends can only be tested for significance if the covariance matrix is clustered at a finer level. Otherwise the covariance matrix estimate is singular and biased downwards. Naive tests will over-state significance.

## 18.7 Do Blue Laws Affect Liquor Sales?

Historically many U.S. states prohibited or limited the sale of alcoholic beverages on Sundays (and are known as “blue laws”). In recent years these laws have been relaxed. Have these changes led to increased consumption of alcoholic beverages? Bernheim, Meer and Novarro (2016) investigated this question using a detailed panel on alcohol consumption and sales hours. It is possible that observed changes coincident with changes in the law might reflect underlying trends. The fact that different states changed their laws during different years allows for a difference-in-difference methodology to identify the treatment effect.

The paper focuses on distilled liquor sales, but wine and beer sales are also included in their data. An abridged version of their data set BMN2016 is posted on the textbook webpage. Liquor is measured in per capita gallons of pure ethanol equivalent. The data are state-level for 47 U.S. states for the years 1970-2007, unbalanced.

The authors carefully gathered information on the allowable hours that alcohol can be sold on a Sunday. They make a distinction between off-premise sales (liquor stores, supermarkets) where consumption is off-premise, and on-premise sales (restaurants, bars) where consumption is on-premises. Let  $y_{it}$  denote the natural logarithm of per-capita liquor sales in state  $i$  in year  $t$ . A simplified version of their basic model is

$$\begin{aligned} y_{it} = & 0.011 \text{ } OnHours_{it} + 0.003 \text{ } OffHours_{it} - 0.013 \text{ } UR_{it} \\ & (0.003) \quad (0.003) \quad (0.004) \\ & + 0.029 \text{ } OnOutFlows_{it} - 0.000 \text{ } OffOutFlows_{it} + u_i + v_t + \varepsilon_{it}. \\ & (0.008) \quad (0.010) \end{aligned} \tag{18.7}$$

OnHours and OffHours are the number of allowable Sunday on-premises and off-premises sale hours. UR is the state unemployment rate. OnOutFlows (OffOutFlows) is the weighted number of on(off)-premises

sale hours less than neighbor states. These are added to adjust for possible cross-border transactions. The model includes both state and year fixed effects. The standard errors are clustered by state.

The estimates indicate that increased on-premise sale hours lead to a small increase in liquor sales. This is consistent with alcohol being a complementary good in social (restaurant and bar) settings. The small and insignificant coefficient on *OffHours* indicates that increased off-premise sale hours does not lead to an increase in liquor sales. This is consistent with rational consumers who adjust their purchases to known hours. The negative effect of the unemployment rate means that liquor sales are pro-cyclical.

The authors were concerned whether their dynamic and trend specifications were correctly specified, so tried some alternative specifications and interactions. To understand the trend issue, we plot in Figure 18.1 the time-series path of the log of per-capita liquor sales for three states: California, Iowa, and New York. You can see that all three exhibit a downward trend from 1970 until about 1995, and then an increasing trend. The slopes of the three trends, however, are not identical. This suggests that there is both a national common component as well as a localized component.

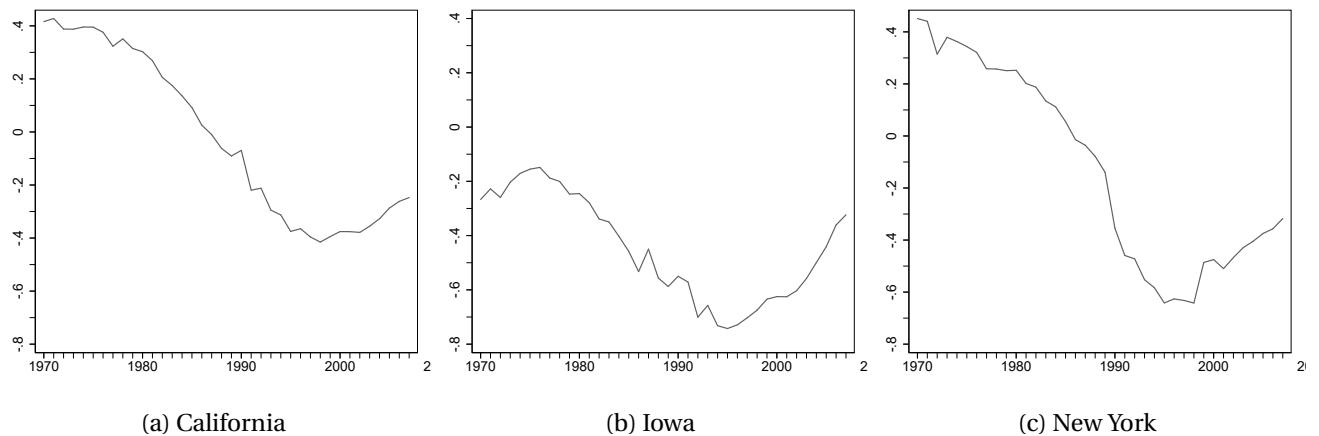


Figure 18.1: Log of Per-Capita Liquor Sales

If we augment the basic model to include state-specific linear trends, the estimates are as follows.

$$\begin{aligned}
 y_{it} = & 0.000 \text{ } OnHours_{it} + 0.002 \text{ } OffHours_{it} - 0.015 \text{ } UR_{it} \\
 & (0.002) \qquad \qquad \qquad (0.002) \qquad \qquad \qquad (0.004) \tag{18.8} \\
 & + 0.005 \text{ } OnOutFlows_{it} - 0.005 \text{ } OffOutFlows_{it} + tw_i + u_i + v_t + \varepsilon_{it} \\
 & (0.005) \qquad \qquad \qquad (0.005)
 \end{aligned}$$

The estimated coefficient for *OnHours* drops to zero and becomes insignificant. The other estimates do not change meaningfully. The authors only discuss this regression in a footnote, stating that adding state-specific trends “demands a great deal from the data and leaves too little variation to identify the effects of interest.” This is an unfortunate claim as actually the standard errors have decreased, not increased, indicating that the effects are better identified. The trouble is that *OnHours* and *OffHours* are trended, and the trends vary by state. This means that these variables are correlated with the state-trend interaction. Omitting the trend interaction induces omitted variable bias. That explains why the coefficient estimates change when the trend specification changes.

Bernheim, Meer and Novarro (2016) is an excellent example of meticulous empirical work with careful attention to detail and isolating a treatment strategy. It is also a good example of how attention to trend specification can affect results.

## 18.8 Check Your Code: Does Abortion Impact Crime?

In a highly-discussed paper, Donohue and Levitt (2001) used a difference-in-difference approach to develop an unusual theory. Crime rates fell dramatically throughout the United States in the 1990s. Donohue and Levitt postulated that one contributing explanation was the landmark 1973 legalization of abortion. The latter might affect the crime rate through two potential channels. First, it reduced the cohort size of young males. Second, it reduced the cohort size of young males at risk for criminal behavior. This suggests the substantial increase in abortions in the early 1970s will translate into a substantial reduction in crime 20 years later.

As you might imagine, this paper was controversial on several dimensions. The paper was also meticulous in its empirical analysis, investigating the potential links using a variety of tools and differing levels of granularity. The most detailed-oriented regressions were presented at the very end of the paper, where the authors exploited differences across age groups. These regressions took the form

$$\log(Arrests_{itb}) = \beta Abortion_{ib} + u_i + \lambda_{tb} + \theta_{it} + \varepsilon_{itb}$$

where  $i$ ,  $t$ , and  $b$  index state, year, and birth cohort.  $Arrests$  is the raw number of arrests for a given crime and  $Abortion$  is the ratio of abortions per live births. The regression includes state fixed effects, cohort-year interactions, and state-year interactions. By including all these interaction effects the regression is estimating a triple-difference, and is identifying the abortion impact on within-state cross-cohort variation, which is a much stronger identification argument than a simple cross-state diff-in-diff regression. Donohue and Levitt reported an estimate of  $\beta$  equalling  $-0.028$  with a small standard error. Based on these estimates Donohue and Levitt suggest that legalizing abortion reduced crime by about 15-25%.

Unfortunately their estimates contained an error. In an attempt to replicate Donohue-Levitt's work, Foote and Goetz (2008) discovered that Donohue-Levitt's computer code inadvertently omitted the state-year interactions  $\theta_{it}$ . This was an important omission as without  $\theta_{it}$  the estimates are based on a mix of cross-state and cross-cohort variation rather than just cross-cohort variation as claimed. Foote and Goetz re-estimated the regression and found an estimate of  $\beta$  equalling  $-0.010$ . While still statistically different from zero, the reduction in magnitude substantially decreased the estimated impact. Foote and Gootz include more extensive empirical analysis as well.

Regardless of the errors and political ramifications, the Donohue-Levitt paper is a very clever and creative use of the difference-in-difference method. It is unfortunate that this creative work was somewhat overshadowed by a debate over computer code.

I believe there are two important messages from this episode. First, include the appropriate controls! In the Donohue-Levitt regression they were correct to advocate for the regression which includes state-year interactions as this allows the most precise measurement of the desired causal impact. Second, check your code! Computation errors are pervasive in applied economic work. It is very easy to make errors; it is very difficult to clean them out of lengthy code. Errors in most papers are ignored as the details receive minor attention. Important and influential papers, however, are scrutinized. If you ever are so blessed as to write a paper which receives significant attention, you will find it most embarrassing if a coding error is found after publication. The solution is to be pro-active and vigilant.

## 18.9 Inference

Many difference-in-difference applications use highly aggregate (e.g. state level) data, because they are investigating the impact of policy changes which occur at an aggregate level. It has become customary in the recent literature to use clustering methods to calculate standard errors with clustering applied at a high level of aggregation.

To understand the motivation for this choice it is useful to review the traditional argument for clustered variance estimation. Suppose that the error  $e_{ig}$  for individual  $i$  in group  $g$  is independent of the regressors, has variance  $\sigma^2$ , and has correlation  $\rho$  across individuals within the group. If the number of

individuals in each group is  $N$  then the exact variance of least squares estimator (recall equation (4.48)) is

$$V_{\hat{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} \sigma^2 (1 + \rho(N - 1))$$

as originally derived by Moulton (1990). This inflates the “usual” variance by the factor  $(1 + \rho(N - 1))$ . Even if  $\rho$  is very small, if  $N$  is huge then this inflation factor can be large as well.

The clustered variance estimator imposes no structure on the conditional variances and correlations within each group. It allows for arbitrary relationships. The advantage is that the resulting variance estimators are robust to a broad range of correlation structures. The disadvantage is that the estimators can be much less precise. Effectively, clustered variance estimators should be viewed as constructed from the number of groups. If you are using U.S. states as your groups (as is commonly seen in applications) then the number of groups is (at most) 51. This means that you are estimating the covariance matrix using 51 observations, regardless of the number of “observations” in the sample. One implication is that if you are estimating more than 51 coefficients the sample covariance matrix estimator will not be full rank, which can invalidate potentially relevant inference methods.

The case for using clustered standard errors was made convincingly in an influential paper by Bertrand, Duflo, and Mullainathan (2004). These authors demonstrated their point by taking the well-known CPS dataset, and then adding randomly generated regressors. They found that if non-clustered variance estimators were used then standard errors would be much too small and a researcher would inappropriately conclude that the randomly generated “variable” has a significant effect in a regression. The false rejections could be eliminated by using clustered standard errors, clustered at the state level. Based on the recommendations from this paper, researchers in economics now routinely cluster similar estimators at the state level.

There are limitations, however. Take the Card-Krueger (1994) example introduced earlier. Their sample had only two states (New Jersey and Pennsylvania). If the standard errors are clustered at the state level then there are only two effective observations available for standard error calculation, which is much too few. For this application clustering at the state level is impossible. One implication might be that this casts doubts on applications involving just a handful of states. If we cannot rule out clustered dependence structures, and cannot use clustering methods due to the small number of states, then it may be that it is inappropriate to put too much trust in the reported standard errors.

Another challenge arises when treatment ( $D_{it} = 1$ ) applies to only a small number of units. The most extreme case is where there is only one treated unit. This could arise, for example, when you are interested in measuring the effect of a policy which only one state has adopted. This situation is particularly treacherous, and is algebraically identical to the problem of robust covariance matrix estimation with sparse dummy variables. (See Section 4.16.) As we learned from that analysis, in the extreme case of a single treated unit, the robust covariance matrix estimator is singular and highly biased towards zero. The problem is due to the fact that the variance of the sub-group is being estimated from a single observation.

The same analysis applies to cluster-variance estimators. If there is a single treated unit then the standard clustered covariance matrix estimator will be singular. If you calculate a standard error for the sub-group mean it will be algebraically zero despite being the most imprecisely estimated coefficient. The treatment effect will have a non-zero reported standard error, but it will be incorrect and highly biased towards zero. For a more detailed analysis and recommendations for inference see Conley and Taber (2011).

## Exercises

**Exercise 18.1** In the text it was claimed that in a balanced sample, individual-level fixed effects are orthogonal to any variable demeaned at the state level.

- (a) Show this claim.
- (b) Does this claim hold in unbalanced samples?
- (c) Explain why this claim implies that the regressions

$$y_{it} = \beta_0 + \beta_1 State_i + \beta_2 Time_t + \theta D_{it} + \varepsilon_{it}$$

and

$$y_{it} = \theta D_{it} + u_i + \delta_t + \varepsilon_{it}$$

yield identical estimates of  $\theta$ .

**Exercise 18.2** In regression (18.1) with  $T = 2$  and  $N = 2$  suppose the time variable is omitted. Thus the estimating equation is

$$y_{it} = \beta_0 + \beta_1 State_i + \theta D_{it} + \varepsilon_{it}.$$

where  $D_{it} = State_i Time_t$  is the treatment indicator.

- (a) Find an algebraic expression for the least squares estimator  $\hat{\theta}$ .
- (b) Show that  $\hat{\theta}$  is a function only of the treated sub-sample and is not a function of the untreated sub-sample.
- (c) Is  $\hat{\theta}$  a difference-in-difference estimator?
- (d) Under which assumptions might  $\hat{\theta}$  be an appropriate estimator of the treatment effect?

**Exercise 18.3** Take the basic difference-in-difference model

$$y_{it} = \theta D_{it} + u_i + \delta_t + \varepsilon_{it}.$$

Instead of assuming that  $D_{it}$  and  $\varepsilon_{it}$  are independent, assume we have an instrumental variable  $z_{it}$  which is independent of  $\varepsilon_{it}$  but is correlated with  $D_{it}$ . Describe how to estimate  $\theta$ .

Hint: Review Section 17.28.

**Exercise 18.4** For the specification tests of Section 18.4 explain why the regression test for homogeneous treatment effects includes only  $N_2 - 1$  interaction dummy variables rather than all  $N_2$  interaction dummies. Also explain why the regression test for equal control effects includes only  $N_1 - 1$  interaction dummy variables rather than all  $N_1$  interaction dummies.

**Exercise 18.5** Use the datafile CK1994 on the textbook webpage. Classical economics teaches that increasing the minimum wage will increase product prices. You can therefore use the Card-Krueger diff-in-diff methodology to estimate the effect of the 1992 New Jersey minimum wage increase on product prices. The data file contains the variables *priceentree*, *pricefry* and *pricesoda*. Create the variable *price* as the sum of these three, indicating the cost of a typical meal.

- (a) Some values of *price* are missing. Delete these observations. This will produce an unbalanced panel, as *price* may be missing for only one of the two surveys. Balance the panel by deleting the paired observation. This can be accomplished by the commands:

- `drop if price == .`

- `bys store: gen nperiods = [_N]`
- `keep if nperiods == 2`

- (b) Create an analog of Table 18.1 but with the price of a meal rather than the number of employees. Interpret the results.
- (c) Estimate an analog of regression (18.2), with price as the dependent variable.
- (d) Estimate an analog of regression (18.4) with state fixed effects, with price as the dependent variable.
- (e) Estimate an analog of regression (18.4) with restaurant fixed effects, with price as the dependent variable.
- (f) Are the results of these regressions the same?
- (g) Create an analog of Table 18.2 for the price of a meal. Interpret the results.
- (h) Test for homogeneous treatment effects across regions.
- (i) Test for equal control effects across regions.

**Exercise 18.6** Use the datafile DS2004 on the textbook webpage. The authors argued that an exogenous police presence would deter automobile theft. The evidence presented in the chapter showed that car theft was reduced for city blocks which received police protection. Does this deterrence effect extend beyond the same block? The dataset has the dummy variable *oneblock* which indicates if the city block is one block away from a protected institution.

- (a) Calculate an analog of Table 18.3 which shows the difference between city blocks which are one block away from a protected institution and those which are more than one block away from a protected institution.
- (b) Estimate a regression model with block and month fixed effects which includes two treatment variables: for city blocks which are on the same block as a protected institution, and for city blocks which are one block away, both interacted with a post-July dummy. Exclude the observations for July.
- (c) Comment on your findings. Does the deterrence effect extend beyond the same city block?

**Exercise 18.7** Use the datafile BMN2016 on the textbook webpage. The authors report results for liquor sales. The data file contains the same information for beer and wine sales. For either beer or wine sales, estimate diff-in-diff models similar to (18.7) and (18.8) and interpret your results. Some relevant variables are *id* (state identification), *year*, *unempw* (unemployment rate). For beer the relevant variables are *logbeer* (log of beer sales), *beeronsun* (number of hours of allowed on-premise sales), *beeroffsun* (number of hours of allowed off-premise sales), *beerOnOutflows*, *beerOffOutflows*. For wine the variables have similar names.

## **Part V**

# **Nonparametric and Nonlinear Methods**

# Chapter 19

## Density Estimation

### 19.1 Introduction

Sometimes it is useful to estimate a density function of a continuously-distributed variable. As a general rule, density functions can take any shape. This means that density functions are inherently **nonparametric**. The most common method to estimate density functions is with **kernel smoothing** estimators, which are related to the kernel regression estimators explored in Chapter 20.

There are many excellent monographs written on nonparametric density estimation, including Silverman (1986) and Scott (1992). The methods are also covered in detail in Pagan and Ullah (1999) and Li and Racine (2007).

In this chapter we focus on univariate density estimation. The setting is a real-valued random variable  $x_i$  for which we have  $n$  observations. The maintained assumption is that  $x_i$  has a continuous density  $f(x)$ . The goal is to estimate  $f(x)$  either at a single point  $x$  or at a set of points in the interior of the support of  $x_i$ . For purposes of presentation we focus on estimation at a single point  $x$ .

For most of the theoretical treatment we assume that the observations are i.i.d. The methods can also be applied to time series and clustered observations, but the theoretical treatment is more advanced. The case of clustered observations is discussed in Section 19.7.

### 19.2 Histogram Density Estimation

To make things concrete, recall the `cpsmar09` dataset and let's focus on the sample of Asian women, which has  $n = 1149$  observations. Our goal is to estimate the density  $f(x)$  of hourly wages for this group.

A simple and familiar density estimator is a histogram. We divide the range of  $f(x)$  into  $B$  bins of width  $w$  and then count the number of observations  $n_j$  in each bin. The histogram estimator of  $f(x)$  for  $x$  in the  $j^{th}$  bin is

$$\hat{f}(x) = \frac{n_j}{nw}. \quad (19.1)$$

The histogram is the plot of these heights, displayed as rectangles. The scaling is set so that the sum of the area of the rectangles is  $\sum_{j=1}^B w n_j / nw = 1$ , and therefore the histogram estimator is a valid density.

To illustrate, Figure 19.1(a) displays the histogram of the sample described above, using bins of width \$10. For example, the first bar shows that 189 of the 1149 individuals had wages in the range [0, 10] so the height of the histogram is  $189/(1149 * 10) = 0.016$ .

The histogram in Figure 19.1(a) is a rather crude estimate. For example, it is uninformative whether \$11 or \$19 wages are more prevalent. To address this we display a histogram using bins with the smaller width \$1 in Figure 19.1(b). In contrast with part (a) this appears quite noisy, and we might guess that the peaks and valleys are likely just sampling noise. We would like an estimator which avoids the two extremes shown in Figures 19.1. For this we consider smoother estimators in the following section.

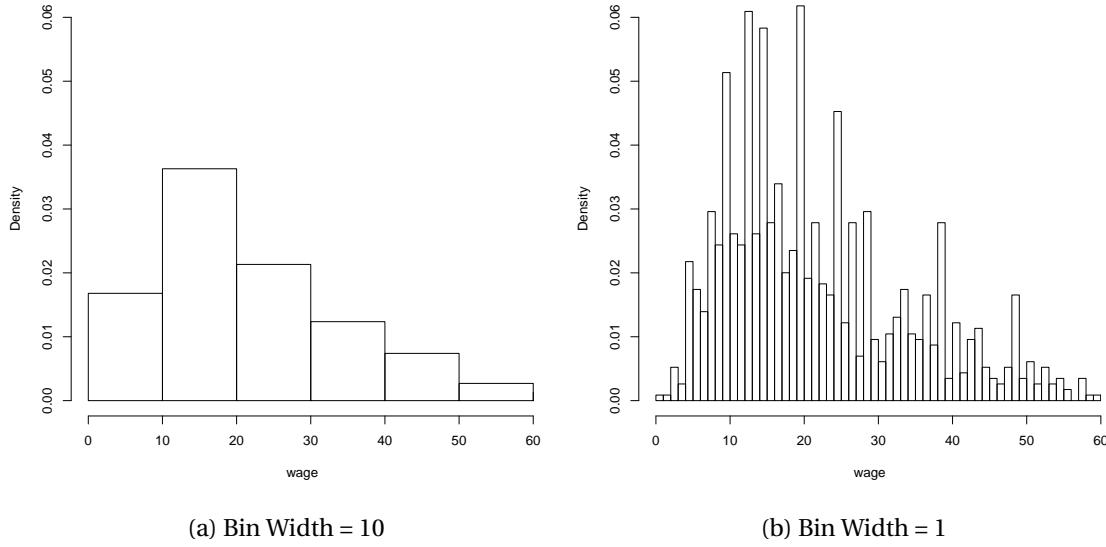


Figure 19.1: Histogram Estimate of Wage Density for Asian Women

### 19.3 Kernel Density Estimator

Continuing the wage density example from the previous section, suppose we want to estimate the density at  $x = \$13$ . Consider the histogram density estimate in Figure 19.1(a). It is based on the frequency of observations in the interval  $[10, 20]$  which is a skewed window about  $x = 13$ . It seems more sensible to center the window at  $x = 13$ , for example to use  $[8, 18]$  instead of  $[10, 20]$ . It also seems sensible to give more weight to observations close to  $x = 13$  and less to those at the edge of the window.

These considerations give rise to what is called the **kernel density estimator** of  $f(x)$ :

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n k\left(\frac{x_i - x}{h}\right). \quad (19.2)$$

where  $k(u)$  is a weighting function known as a **kernel function** and  $h > 0$  is a scalar known as a **bandwidth**. The estimator (19.2) is the sample average of the “kernel smooths”  $h^{-1} k\left(\frac{x_i - x}{h}\right)$ . The latter will be seen throughout this chapter and the next as they are used in all kernel smoothing estimators. The estimator (19.2) was first proposed by Rosenblatt (1956) and Parzen (1962), and is often called the Rosenblatt or Rosenblatt-Parzen kernel density estimator.

Kernel density estimators (19.2) can be constructed with any kernel satisfying the following definition.

**Definition 19.1** A (second-order) **kernel function**  $k(u)$  satisfies

1.  $0 \leq k(u) \leq \bar{k} < \infty$ ,
2.  $k(u) = k(-u)$ ,
3.  $\int_{-\infty}^{\infty} k(u) du = 1$ ,
4.  $\int_{-\infty}^{\infty} u^2 k(u) du = 1$ ,
5.  $\int_{-\infty}^{\infty} |u|^r k(u) du < \infty$  for all positive integers  $r$ .

Essentially, a kernel function is a bounded probability density function which is symmetric about zero and has a unit variance. Since  $k(u)$  is symmetric about zero all odd moments equal zero. Assumption 19.1.4, that the kernel has unit variance, is not essential, but a convenient normalization. Assumption 19.1.5 is also not essential for most results, but again is a convenient simplification, and does not exclude any kernel functions used in standard empirical practice.

The estimator (19.2) critically depends on the bandwidth  $h$ .

**Definition 19.2** A **bandwidth** or **tuning parameter**  $h > 0$  is a real number used to control the degree of smoothing of a nonparametric estimator.

Typically, larger values of a bandwidth  $h$  result in smoother estimators and smaller values of  $h$  result in less smooth estimators.

The histogram density estimator (19.1) equals the kernel density estimator (19.2) at the bin midpoints (e.g.  $x = 5$  or  $x = 15$  in Figure 19.1(a)) when  $k(u)$  is a uniform density function. This is known as the **rectangular kernel**.

The kernel density estimator generalizes the histogram estimator in two important ways. First, the window is centered at the point  $x$  rather than by bins, and second, the observations are weighted by the kernel function. Thus, the estimator (19.2) can be viewed as a smoothed histogram.  $\hat{f}(x)$  counts the frequency that observations  $x_i$  are close to  $x$ . The bandwidth  $h$  determines what is meant by “close”, and the kernel  $k(u)$  applies weights based on the distance between  $x_i$  and  $x$ .

There are a large number of functions which satisfy Definition 19.1, and many are programmed as options in statistical packages. We list the three most important in Table 19.1 below: the **rectangular**, **Gaussian**, and **Epanechnikov** kernels. These three kernel functions are displayed in Figure 19.2. In practice it is unnecessary to consider kernels beyond these three.

Table 19.1: Common Second-Order Kernels

Kernel	Formula	$R_k$	$C_k$
Rectangular	$k(u) = \begin{cases} \frac{1}{2\sqrt{3}} & \text{if }  u  < \sqrt{3} \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{2\sqrt{3}}$	1.064
Gaussian	$k(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$	$\frac{1}{2\sqrt{\pi}}$	1.059
Epanechnikov	$k(u) = \begin{cases} \frac{3}{4\sqrt{5}} \left(1 - \frac{u^2}{5}\right) & \text{if }  u  < \sqrt{5} \\ 0 & \text{otherwise} \end{cases}$	$\frac{3\sqrt{5}}{25}$	1.049

In practice we advise against the rectangular kernel as it produces discontinuous density estimates. Better choices are the Epanechnikov and Gaussian kernels which give more weight to observations  $x_i$  near the point of evaluation  $x$ . In most practical applications these two kernels will provide very similar density estimates, with the Gaussian somewhat smoother. In practice the Gaussian kernel is a convenient choice it produces a density estimator which possesses derivatives of all orders.

Kernel estimators are invariant to rescaling the kernel function and bandwidth. That is, the estimator (19.2) using a kernel  $k(u)$  and bandwidth  $h$  is equal for any  $b > 0$  to a kernel density estimator using the kernel  $k_b(u) = k(u/b)/b$  with bandwidth  $h/b$ .

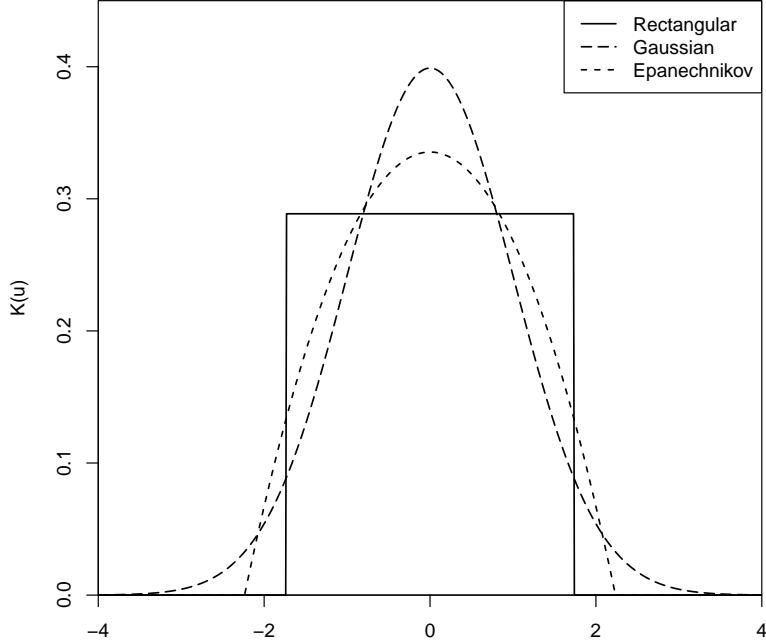


Figure 19.2: Kernel Functions

Kernel density estimators are also invariant to data rescaling. That is, let  $y_i = cx_i$  for some  $c > 0$ . Then the density of  $y_i$  is  $f_y(y) = f_x(y/c)/c$ . If  $\hat{f}_x(x)$  is the estimator (19.2) using the observations  $x_i$  and bandwidth  $h$ , and  $\hat{f}_y(y)$  is the estimator using the scaled observations  $y_i$  with bandwidth  $ch$ , then  $\hat{f}_y(y) = \hat{f}_x(y/c)/c$ , appropriately.

The kernel density estimator (19.2) is a valid density function. Specifically, it is non-negative and integrates to one. To see the latter point,

$$\int_{-\infty}^{\infty} \hat{f}(x) dx = \frac{1}{nh} \sum_{i=1}^n \int_{-\infty}^{\infty} k\left(\frac{x_i - x}{h}\right) dx = \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{\infty} k(u) du = 1$$

where the second equality makes the change-of-variables  $u = (x_i - x)/h$  and the final uses Definition 19.1.3.

To illustrate, Figure 19.3 displays the histogram estimator along with the kernel density estimator using the Gaussian kernel with the bandwidth  $h = 2.14$ . (Bandwidth selection will be discussed in Section 19.10.) You can see that the density estimator is a smoothed version of the histogram, and is single-peaked with a maximum about  $x = \$13$ .

## 19.4 Bias of Density Estimator

In this section we show how to approximate the bias of the density estimator.

Since the kernel density estimator (19.2) is an average of i.i.d. observations its expectation is

$$\mathbb{E}(\hat{f}(x)) = \mathbb{E}\left(\frac{1}{nh} \sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)\right) = \mathbb{E}\left(\frac{1}{h} k\left(\frac{x_i - x}{h}\right)\right).$$

At this point we may feel unsure if we can proceed further, as  $k((x_i - x)/h)$  is a nonlinear function of the

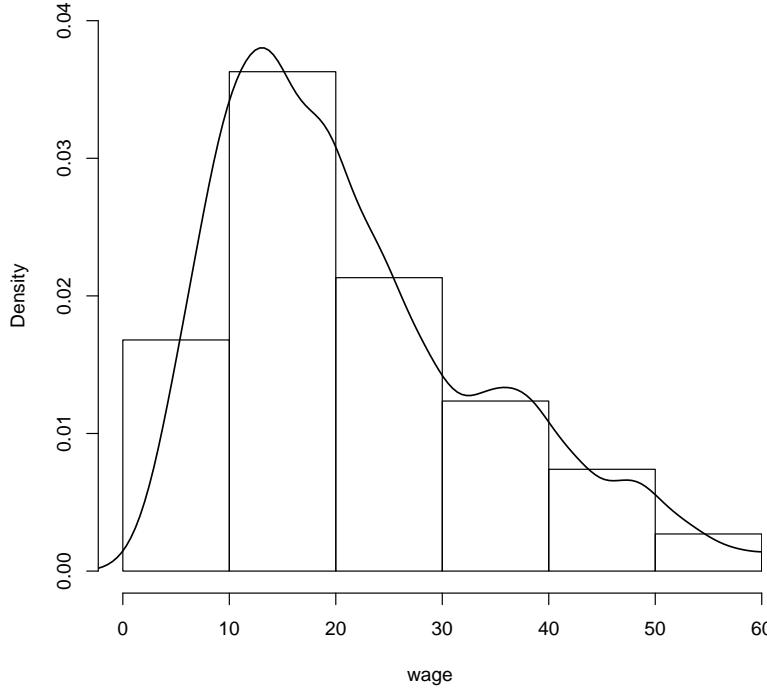


Figure 19.3: Kernel Density Estimator of Wage Density for Asian Women

random variable  $x_i$ . To make progress, we write the expectation as an explicit integral

$$= \int_{-\infty}^{\infty} \frac{1}{h} k\left(\frac{v-x}{h}\right) f(v) dv.$$

The next step is a trick. Make the change-of-variables  $u = (v - x)/h$  so that the expression equals

$$\begin{aligned} &= \int_{-\infty}^{\infty} k(u) f(x + hu) du \\ &= f(x) + \int_{-\infty}^{\infty} k(u) (f(x + hu) - f(x)) du \end{aligned} \tag{19.3}$$

where the final equality uses Definition 19.1.3.

Expression (19.3) shows that the expected value of  $\hat{f}(x)$  is a weighted average of the function  $f(u)$  about the point  $u = x$ . When  $f(x)$  is linear then  $\hat{f}(x)$  will be unbiased for  $f(x)$ . In general, however,  $\hat{f}(x)$  is a biased estimator.

As  $h$  decreases to zero, the bias term in (19.3) tends to zero:

$$\mathbb{E}(\hat{f}(x)) = f(x) + o(1).$$

Intuitively, (19.3) is an average of  $f(u)$  in a local window about  $x$ . If the window is sufficiently small then this average should be close to  $f(x)$ .

Under a stronger smoothness condition we can provide an improved characterization of the bias. Make a second-order Taylor series expansion of  $f(x + hu)$  so that

$$f(x + hu) = f(x) + f'(x)hu + \frac{1}{2}f''(x)h^2u^2 + o(h^2).$$

Substituting, we find that (19.3) equals

$$\begin{aligned}
 &= f(x) + \int_{-\infty}^{\infty} k(u) \left( f'(x) h u + \frac{1}{2} f''(x) h^2 u^2 \right) du + o(h^2) \\
 &= f(x) + f'(x) h \int_{-\infty}^{\infty} u k(u) du + \frac{1}{2} f''(x) h^2 \int_{-\infty}^{\infty} u^2 k(u) du + o(h^2) \\
 &= f(x) + \frac{1}{2} f''(x) h^2 + o(h^2).
 \end{aligned}$$

The final equality uses the facts  $\int_{-\infty}^{\infty} u k(u) du = 0$  and  $\int_{-\infty}^{\infty} u^2 k(u) du = 1$ . We have shown that (19.3) simplifies to

$$\mathbb{E}(\hat{f}(x)) = f(x) + \frac{1}{2} f''(x) h^2 + o(h^2).$$

This is revealing. It shows that the approximate bias of  $\hat{f}(x)$  for  $f(x)$  is  $\frac{1}{2} f''(x) h^2$ . This is consistent with our earlier finding that the bias decreases as  $h$  tends to zero, but is a more constructive characterization. We see that the bias depends on the underlying curvature of  $f(x)$  through its second derivative. If  $f''(x) < 0$  (as it typical at the mode) then the bias is negative, meaning that  $\hat{f}(x)$  is typically less than the true  $f(x)$ . If  $f''(x) > 0$  (as may occur in the tails) then the bias is positive, meaning that  $\hat{f}(x)$  is typically higher than the true  $f(x)$ . This is smoothing bias.

We summarize our findings. Let  $\mathcal{N}$  be a neighborhood of  $x$ .

**Theorem 19.1** If  $f(x)$  is continuous in  $\mathcal{N}$ , then as  $h \rightarrow 0$

$$\mathbb{E}(\hat{f}(x)) = f(x) + o(1). \quad (19.4)$$

If  $f''(x)$  is continuous in  $\mathcal{N}$ , then as  $h \rightarrow 0$

$$\mathbb{E}(\hat{f}(x)) = f(x) + \frac{1}{2} f''(x) h^2 + o(h^2). \quad (19.5)$$

A formal proof is presented in Section 19.18.

The asymptotic unbiasedness result (19.4) holds under the minimal assumption that  $f(x)$  is continuous. The asymptotic expansion (19.5) holds under the stronger assumption that the second derivative is continuous. These are examples of what are often called **smoothness** assumptions. They are interpreted as meaning that the density is not too variable. It is a common feature of nonparametric theory to use smoothness assumptions to obtain asymptotic approximations.

To illustrate the bias of the kernel density estimator, Figure 19.4 displays the density

$$f(x) = \frac{3}{4} \phi(x-4) + \frac{1}{3} \phi\left(\frac{x-7}{3/4}\right)$$

with the solid line. You can see that the density is bimodal, with local peaks at 4 and 7. Now imagine estimating this density using a Gaussian kernel and a bandwidth of  $h = 0.5$  (which turns out to be the reference rule (see Section 19.10) for a sample size  $n = 200$ ). The mean  $\mathbb{E}(\hat{f}(x))$  of this estimator is plotted using the long dashes. You can see that it has the same general shape with  $f(x)$ , with the same local peaks, but the peak and valley are attenuated. The mean is a smoothed version of the actual density  $f(x)$ . The asymptotic approximation  $f(x) + f''(x)h^2/2$  is displayed using the short dashes. You can see that it is similar to the mean  $\mathbb{E}(\hat{f}(x))$ , but is not identical. The difference between  $f(x)$  and  $\mathbb{E}(\hat{f}(x))$  is the bias of the estimator.

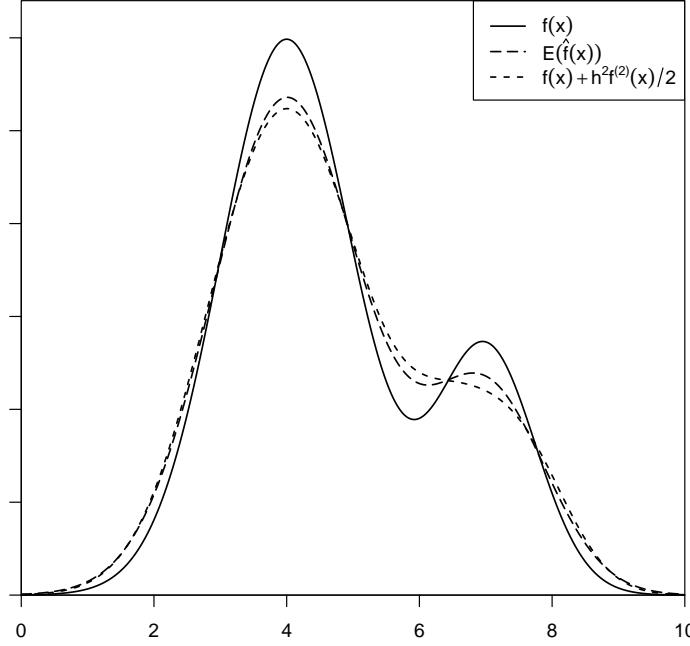


Figure 19.4: Smoothing Bias

## 19.5 Variance of Density Estimator

Since  $\hat{f}(x)$  is a sample average of the kernel smooths and the latter are i.i.d., the exact variance of  $\hat{f}(x)$  is

$$\text{var}(\hat{f}(x)) = \frac{1}{n^2 h^2} \text{var}\left(\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)\right) = \frac{1}{nh^2} \text{var}\left(k\left(\frac{x_i - x}{h}\right)\right).$$

This can be approximated by calculations similar to those used for the bias.

**Theorem 19.2** The exact variance of  $\hat{f}(x)$  is

$$V_{\hat{f}} = \text{var}(\hat{f}(x)) = \frac{1}{nh^2} \text{var}\left(k\left(\frac{x_i - x}{h}\right)\right). \quad (19.6)$$

If  $f(x)$  is continuous in  $\mathcal{N}$ , then as  $h \rightarrow 0$  and  $nh \rightarrow \infty$

$$nh V_{\hat{f}} = \frac{f(x) R_k}{nh} + o\left(\frac{1}{nh}\right) \quad (19.7)$$

where

$$R_k = \int_{-\infty}^{\infty} k(u)^2 du \quad (19.8)$$

is known as the **roughness** of the kernel  $k(u)$ .

The proof is presented in Section 19.18

Equation (19.7) shows that the asymptotic variance of  $\hat{f}(x)$  is inversely proportional to  $nh$ , which can be viewed as the effective sample size. The variance is proportional to the height of the density  $f(x)$  and the kernel roughness  $R_k$ . The values of  $R_k$  for the three kernel functions are displayed in Table 19.1.

## 19.6 Variance Estimation and Standard Errors

The expressions (19.6) and (19.7) can be used to motivate estimators of the variance  $V_{\hat{f}}$ . An estimator based on the finite sample formula (19.6) is the scaled sample variance of the kernel smooths  $h^{-1}k(\frac{x_i-x}{h})$

$$\hat{V}_{\hat{f}}(x) = \frac{1}{n-1} \left( \frac{1}{nh^2} \sum_{i=1}^n k\left(\frac{x_i-x}{h}\right)^2 - \hat{f}(x)^2 \right).$$

An estimator based on the asymptotic formula (19.7) is

$$\hat{V}_{\hat{f}}(x) = \frac{\hat{f}(x)R_k}{nh}. \quad (19.9)$$

Using either estimator, a standard error for  $\hat{f}(x)$  is  $\hat{V}_{\hat{f}}(x)^{1/2}$ .

## 19.7 Clustered Observations

When the observations are clustered we can write the density estimator (19.2) using the notation

$$\hat{f}(x) = \frac{1}{n} \sum_{g=1}^G \sum_{i=1}^{n_g} \frac{1}{h} k\left(\frac{x_{ig}-x}{h}\right).$$

When the clusters are mutually independent this has exact variance

$$\begin{aligned} V_{\hat{f}}(x) &= \frac{1}{n^2} \sum_{g=1}^G \text{var}\left(\sum_{i=1}^{n_g} \frac{1}{h} k\left(\frac{x_{ig}-x}{h}\right)\right) \\ &= \frac{1}{n^2} \sum_{g=1}^G \mathbb{E}\left(\sum_{i=1}^{n_g} \frac{1}{h} k\left(\frac{x_{ig}-x}{h}\right) - n_g f(x)\right)^2. \end{aligned}$$

This can be estimated by

$$\begin{aligned} \hat{V}_{\hat{f}}(x) &= \frac{1}{n^2} \sum_{g=1}^G \left( \sum_{i=1}^{n_g} \frac{1}{h} k\left(\frac{x_{ig}-x}{h}\right) - n_g \hat{f}(x) \right)^2 \\ &= \frac{1}{n^2 h^2} \sum_{g=1}^G \left( \sum_{i=1}^{n_g} k\left(\frac{x_{ig}-x}{h}\right) \right)^2 - \frac{1}{n^2} \sum_{g=1}^G n_g^2 \hat{f}(x)^2. \end{aligned}$$

A clustered standard error for  $\hat{f}(x)$  is  $\hat{V}_{\hat{f}}(x)^{1/2}$ .

## 19.8 IMSE of Density Estimator

A useful measure of precision of a density estimator is its **integrated mean squared error** (IMSE).

$$\text{IMSE} = \int_{-\infty}^{\infty} \mathbb{E}(\hat{f}(x) - f(x))^2 dx.$$

It is the average precision of  $\hat{f}(x)$  over all values of  $x$ . Using Theorems 19.1 and 19.2 we can calculate that it equals

$$\text{IMSE} = \frac{1}{4} R(f'') h^4 + \frac{R_k}{nh} + o(h^4) + o((nh)^{-1})$$

where

$$R(f'') = \int_{-\infty}^{\infty} (f''(x))^2 dx$$

is called the **roughness** of the second derivative  $f''(x)$ . The leading term

$$\text{AIMSE} = \frac{1}{4} R(f'') h^4 + \frac{R_k}{nh} \quad (19.10)$$

is called the **asymptotic integrated mean squared error**. The AIMSE is an asymptotic approximation to the IMSE. In nonparametric theory it is common to use AIMSE to assess precision.

The AIMSE (19.10) shows that  $\hat{f}(x)$  is less accurate when  $R(f'')$  is large, meaning that accuracy deteriorates with increased curvature in  $f(x)$ . The expression also shows that the first term (the squared bias) of the AIMSE is increasing in  $h$ , but the second term (the variance) is decreasing in  $h$ . Thus the choice of  $h$  affects (19.10) with a trade-off between bias and variance.

We can calculate the bandwidth  $h$  which minimizes the AIMSE by solving the first-order condition. (See Exercise 19.2.) The solution is

$$h_0 = \left( \frac{R_k}{R(f'')} \right)^{1/5} n^{-1/5}. \quad (19.11)$$

This bandwidth takes the form  $h_0 = cn^{-1/5}$  so satisfies the intriguing rate  $h_0 \sim n^{-1/5}$ .

A common error is to interpret  $h_0 \sim n^{-1/5}$  as meaning that a user can set  $h = n^{-1/5}$ . This is incorrect and can be a huge mistake in an application. The constant  $c$  is critically important as well.

When  $h \sim n^{-1/5}$  then  $\text{AIMSE} \sim n^{-4/5}$  which means that the density estimator converges at the rate  $n^{-2/5}$ . This is slower than the standard  $n^{-1/2}$  parametric rate. This is a common finding in nonparametric analysis. An interpretation is that nonparametric estimation problems are harder than parametric problems, so more observations are required to obtain accurate estimates.

We summarize our findings.

**Theorem 19.3** If  $f''(x)$  is uniformly continuous, then

$$\text{IMSE} = \frac{1}{4} R(f'') h^4 + \frac{R_k}{nh} + o(h^4) + o((nh)^{-1}).$$

The leading terms (the AIMSE) are minimized by the bandwidth

$$h_0 = \left( \frac{R_k}{R(f'')} \right)^{1/5} n^{-1/5}.$$

## 19.9 Optimal Kernel

Expression (19.10) shows that the choice of kernel function affects the AIMSE only through  $R_k$ . This means that the kernel with the smallest  $R_k$  will have the smallest AIMSE. As shown by Hodges and Lehmann (1956),  $R_k$  is minimized by the Epanechnikov kernel. This means that density estimation with the Epanechnikov kernel is AIMSE efficient. This observation led Epanechnikov (1969) to recommend this kernel for density estimation.

**Theorem 19.4** AIMSE is minimized by the Epanechnikov kernel.

We prove Theorem 19.4 below.

It is also interesting to calculate the efficiency loss obtained by using a different kernel. Inserting the optimal bandwidth (19.11) into the AIMSE (19.10) and a little algebra we find that for any kernel the optimal AIMSE is

$$\text{AIMSE}_0(k) = \frac{5}{4} R(f'')^{1/5} R_k^{4/5}.$$

The square root of the ratio of the optimal AIMSE of the Gaussian kernel to the Epanechnikov kernel is

$$\left( \frac{\text{AIMSE}_0(\text{Gaussian})}{\text{AIMSE}_0(\text{Epanechnikov})} \right)^{1/2} = \left( \frac{R_k(\text{Gaussian})}{R_k(\text{Epanechnikov})} \right)^{2/5} = \left( \frac{1/2\sqrt{\pi}}{3\sqrt{5}/25} \right)^{2/5} \simeq 1.02.$$

Thus the efficiency loss from using the Gaussian kernel relative to the Epanechnikov is only 2%. This is not particularly large. Therefore from an efficiency viewpoint the Epanechnikov is optimal, and the Gaussian is near-optimal.

The Gaussian kernel has other advantages over the Epanechnikov. The Gaussian kernel possesses derivatives of all orders (is infinitely smooth) so kernel density estimates with the Gaussian kernel will also have derivatives of all orders. This is not the case with the Epanechnikov kernel, as its first derivative is discontinuous at the boundary of its support. Consequently estimates calculated using the Gaussian kernel are smoother and particularly well suited for estimation of density derivatives. Another useful feature is that the density estimator  $\hat{f}(x)$  with the Gaussian kernel is non-zero for all  $x$ , which can be a useful feature if the inverse  $\hat{f}(x)^{-1}$  is desired. These considerations lead to the practical recommendation to use the Gaussian kernel.

We now show Theorem 19.4. To do so we use the calculus of variations. Construct the Lagrangian

$$\mathcal{L}(k, \lambda_1, \lambda_2) = \int_{-\infty}^{\infty} k(u)^2 du - \lambda_1 \left( \int_{-\infty}^{\infty} k(u) du - 1 \right) - \lambda_2 \left( \int_{-\infty}^{\infty} u^2 k(u) du - 1 \right).$$

The first term is  $R_k$ . The constraints are that the kernel integrates to one and the second moment is 1. Taking the derivative with respect to  $k(u)$  and setting to zero we obtain

$$\frac{d}{dk(u)} \mathcal{L}(k, \lambda_1, \lambda_2) = (2k(u) - \lambda_1 - \lambda_2 u^2) \mathbf{1}(k(u) \geq 0) = 0.$$

Solving for  $k(u)$  we find the solution

$$k(u) = \frac{1}{2} (\lambda_1 + \lambda_2 u^2) \mathbf{1}(\lambda_1 + \lambda_2 u^2 \geq 0)$$

which is a truncated quadratic.

The constants  $\lambda_1$  and  $\lambda_2$  may be found by setting  $\int_{-\infty}^{\infty} k(u) du = 1$  and  $\int_{-\infty}^{\infty} u^2 k(u) du = 1$ . After some algebra we find the solution is the Epanechnikov kernel as listed in Table 19.1.

## 19.10 Reference Bandwidth

The density estimator (19.2) depends critically on the bandwidth  $h$ . Without a specific rule to select  $h$  the method is incomplete. Consequently an important component of nonparametric estimation methods are data-dependent bandwidth selection rules.

A simple bandwidth selection rule proposed by Silverman (1986) has come to be known as the **reference bandwidth** or **Silverman's Rule-of-Thumb**. It uses the bandwidth (19.11) which is optimal under the simplifying assumption that the true density  $f(x)$  is normal, with a few variations. The rule produces a reasonable bandwidth for many estimation contexts.

The Silverman rule is

$$h_r = \sigma_x C_k n^{-1/5} \tag{19.12}$$

where  $\sigma_x$  is the standard deviation of the distribution of  $x$  and

$$C_k = \left( \frac{8\sqrt{\pi}R_k}{3} \right)^{1/5}.$$

The constant  $C_k$  is determined by the kernel. Its values are recorded in Table 19.1.

The Silverman rule is simple to derive. Using change-of-variables you can calculate that when  $f(x) = \sigma_x^{-1}\phi(x/\sigma_x)$  then  $R(f'') = \sigma_x^{-5}R(\phi'')$ . A technical calculation (see Theorem 19.5 below) shows that  $R(\phi'') = 3/8\sqrt{\pi}$ . Together we obtain the reference estimate  $R(f'') = \sigma_x^{-5}3/8\sqrt{\pi}$ . Inserted into (19.11) we obtain (19.12).

For the Gaussian kernel  $R_k = 1/2\sqrt{\pi}$  so the constant  $C_k$  is

$$C_k = \left( \frac{8\sqrt{\pi}}{3} \frac{1}{2\sqrt{\pi}} \right)^{1/5} = \left( \frac{4}{3} \right)^{1/5} \approx 1.059. \quad (19.13)$$

Thus the Silverman rule (19.12) is often written as

$$h_r = \sigma_x 1.06 n^{-1/5}. \quad (19.14)$$

It turns out that the constant (19.13) is remarkably robust to the choice of kernel. Notice that  $C_k$  depends on the kernel only through  $R_k$ , which is minimized by the Epanechnikov kernel for which  $C_k \approx 1.05$ , and maximized (among single-peaked kernels) by the rectangular kernel for which  $C_k \approx 1.06$ . Thus the constant  $C_k$  is essentially invariant to the specific kernel. Consequently the Silverman rule (19.14) can be used by any kernel with unit variance.

The unknown standard deviation  $\sigma_x$  needs to be replaced with a sample estimator. Using the sample standard deviation  $\hat{\sigma}_x$  we obtain a classical reference rule for the Gaussian kernel, sometimes referred to as the optimal bandwidth under the assumption of normality:

$$h_r = \hat{\sigma}_x 1.06 n^{-1/5}. \quad (19.15)$$

Silverman (1986, Section 3.4.2) recommended a robust estimator for  $\sigma_x$  based on the interquartile range  $\hat{R}$  (the difference between the 0.75 and 0.25 quantiles) divided by 1.34. Silverman suggested the smaller of this and the sample standard deviation

$$\tilde{\sigma}_x = \min [\hat{\sigma}_x, \hat{R}/1.34].$$

This gives rise to a second form of the reference rule

$$h_r = \tilde{\sigma}_x 1.06 n^{-1/5}. \quad (19.16)$$

Silverman (1986) observed that the constant  $C_k = 1.06$  produces a bandwidth which is a bit too large when the density  $f(x)$  is thick-tailed or bimodal. He therefore recommended using a slightly smaller bandwidth in practice, and based on simulation evidence specifically recommended  $C_k = 0.9$ . This leads to a third form of the reference rule

$$h_r = 0.9 \tilde{\sigma}_x n^{-1/5}. \quad (19.17)$$

This rule (19.17) is popular in package implementations and is commonly known as Silverman's Rule of Thumb.

The kernel density estimator implemented with any of the above reference bandwidths is fully data-dependent and thus a valid estimator. (That is, it does not depend on user-selected tuning parameters.) This is a good property.

We close this section by justifying the claim  $R(\phi'') = 3/8\sqrt{\pi}$ . We provide a more general calculation, and present the proof in Section 19.18.

**Theorem 19.5** For any integer  $m \geq 0$ ,

$$R(\phi^{(m)}) = \frac{\mu_{2m}}{2^{m+1}\sqrt{\pi}} \quad (19.18)$$

where  $\mu_{2m} = (2m-1)!! = \mathbb{E}(Z^{2m})$  is the  $2m^{\text{th}}$  moment of the standard normal density.

## 19.11 Sheather-Jones Bandwidth\*

In this section we present a bandwidth selection rule derived by Sheather and Jones (1991) which has much improved performance over the reference rule.

The AIMSE-optimal bandwidth (19.11) depends on the unknown roughness  $R(f'')$ . An improvement on the reference rule may be obtained by replacing  $R(f'')$  with a nonparametric estimator.

Consider the general problem of estimation of  $S_m = \int_{-\infty}^{\infty} (f^{(m)}(x))^2 dx$  for some integer  $m \geq 0$ . By  $m$  applications of integration-by-parts we can calculate that

$$S_m = (-1)^m \int_{-\infty}^{\infty} f^{(2m)}(x) f(x) dx = (-1)^m \mathbb{E}(f^{(2m)}(x_i))$$

where the second equality uses the fact that  $f(x)$  is the density of  $x_i$ . Let  $\hat{f}(x) = (nb_m)^{-1} \sum_{i=1}^n \phi((x_i - x)/b_m)$  be a kernel density estimator using the Gaussian kernel and bandwidth  $b_m$ . An estimator of  $f^{(2m)}(x)$  is

$$\hat{f}^{(2m)}(x) = \frac{1}{nb_m^{2m+1}} \sum_{i=1}^n \phi^{(2m)}\left(\frac{x_i - x}{b_m}\right).$$

A non-parametric estimator of  $S_m$  is

$$\hat{S}_m(b_m) = \frac{(-1)^m}{n} \sum_{i=1}^n \hat{f}^{(2m)}(x_i) = \frac{(-1)^m}{n^2 b_m^{2m+1}} \sum_{i=1}^n \sum_{j=1}^n \phi^{(2m)}\left(\frac{x_i - x_j}{b_m}\right).$$

Jones and Sheather (1991) calculated that the MSE-optimal bandwidth  $b_m$  for the estimator  $\hat{S}_m$  is

$$b_m = \left( \sqrt{\frac{2}{\pi}} \frac{\mu_{2m}}{S_{m+1}} \right)^{1/(3+2m)} n^{-1/(3+2m)} \quad (19.19)$$

where  $\mu_{2m} = (2m-1)!!$  is the  $2m^{\text{th}}$  moment of the normal kernel. The bandwidth (19.19) depends on the unknown  $S_{m+1}$ . One solution is to replace  $S_{m+1}$  with a reference estimate. Given Theorem 19.5 this is  $S_{m+1} = \sigma_x^{-3-2m} \mu_{2m+2} / 2^{m+2} \sqrt{\pi}$ . Substituted into (19.19) and simplifying we obtain the reference bandwidth

$$\tilde{b}_m = \sigma_x \left( \frac{2^{m+5/2}}{2m+1} \right)^{1/(3+2m)} n^{-1/(3+2m)}.$$

Used for estimation of  $S_m$  we obtain the feasible estimator  $\tilde{S}_m = \hat{S}_m(\tilde{b}_m)$ . It turns out that two reference bandwidths of interest are

$$\tilde{b}_2 = 1.24 \sigma_x n^{-1/7}$$

and

$$\tilde{b}_3 = 1.23 \sigma_x n^{-1/9}$$

for  $\tilde{S}_2$  and  $\tilde{S}_3$ .

A **plug-in bandwidth**  $h$  is obtained by replacing the unknown  $S_2 = R(f'')$  in (19.11) with  $\tilde{S}_2$ . Its performance, however, depends critically on the preliminary bandwidth  $b_2$  which depends on the reference rule estimator  $\tilde{S}_3$ .

Sheather and Jones (1991) improved on the plug-in bandwidth with the following algorithm which takes into account the interactions between  $h$  and  $b_2$ . Take the two equations for optimal  $h$  and  $b_2$  with  $S_2$  and  $S_3$  replaced with the reference estimates  $\tilde{S}_2$  and  $\tilde{S}_3$

$$h = \left( \frac{R_k}{\tilde{S}_2} \right)^{1/5} n^{-1/5}$$

$$b_2 = \left( \sqrt{\frac{2}{\pi}} \frac{3}{\tilde{S}_3} \right)^{1/7} n^{-1/7}.$$

Solve the first equation for  $n$  and plug it into the second equation, viewing it as a function of  $h$ . We obtain

$$b_2(h) = \left( \sqrt{\frac{2}{\pi}} \frac{3}{R_k} \frac{\tilde{S}_2}{\tilde{S}_3} \right)^{1/7} h^{5/7}.$$

Now use  $\tilde{b}_2(h)$  to make the estimator  $\hat{S}_2(\tilde{b}_2(h))$  a function of  $h$ . Find the  $h$  which is the solution to the equation

$$h = \left( \frac{R_k}{\hat{S}_2(\tilde{b}_2(h))} \right)^{1/5} n^{-1/5}. \quad (19.20)$$

The solution for  $h$  must be found numerically but it is fast to solve by the Newton-Raphson method. Theoretical and simulation analysis have shown that the resulting bandwidth  $h$  and density estimator  $\hat{f}(x)$  perform quite well in a range of contexts.

When the kernel  $k(u)$  is Gaussian the relevant formulae are

$$b_2(h) = 1.357 \left( \frac{\tilde{S}_2}{\tilde{S}_3} \right)^{1/7} h^{5/7}$$

and

$$h = \frac{0.776}{\hat{S}_2(\tilde{b}_2(h))^{1/5}} n^{-1/5}.$$

## 19.12 Recommendations for Bandwidth Selection

In general it is advisable to try several bandwidths and use judgment. Estimate the density function using each bandwidth. Plot the results and compare. Select your density estimator based on the evidence, your purpose for estimation, and your judgment.

For example, take the empirical example presented at the beginning of this chapter, which are wages for the sub-sample of Asian women. There are  $n = 1149$  observations. Thus  $n^{-1/5} = 0.24$ . The sample standard deviation is  $\hat{\sigma}_x = 20.6$ . This means that the Gaussian optimal rule (19.15) is

$$h = \hat{\sigma}_x 1.06 n^{-1/5} = 5.34.$$

The interquartile range is  $\hat{R} = 18.8$ . The robust estimate of standard deviation is  $\tilde{\sigma}_x = 14.0$ . The rule-of-thumb (19.17) is

$$h = 0.9 \tilde{\sigma}_x n^{-1/5} = 3.08.$$

This is smaller than the Gaussian optimal bandwidth mostly because the robust standard deviation is much smaller than the sample standard deviation.

The Sheather-Jones bandwidth which solves (19.20) is

$$h = 2.14.$$

This is significantly smaller than the other two bandwidths. This is because the empirical roughness estimate  $\hat{S}_2$  is much larger than the normal reference value.

We estimate the density using these three bandwidths and the Gaussian kernel, and display the estimates in Figure 19.5. What we can see is that the estimate using the largest bandwidth (the Gaussian optimal) is the smoothest, and the estimate using the smallest bandwidth (Sheather-Jones) is the least smooth. The Gaussian optimal estimate understates the primary density mode, and overstates the left tail, relative to the other two. The Gaussian optimal estimate seems over-smoothed. The estimates using the rule-of-thumb and the Sheather-Jones bandwidth are reasonably similar, and the choice between the two may be made partly on aesthetics. The rule-of-thumb estimate produces a smoother estimate which may be more appealing to the eye, while the Sheather-Jones estimate produces more detail. My preference leans towards detail, and hence the Sheather-Jones bandwidth. This is the justification for the choice  $h = 2.14$  used for the density estimate which was displayed in Figure 19.3.

If you are working in a package which only produces one bandwidth rule (such as Stata) then it is advisable to experiment by trying alternative bandwidths obtained by adding and subtracting modest deviations (e.g. 20-30%) and then assess the density plots obtained.

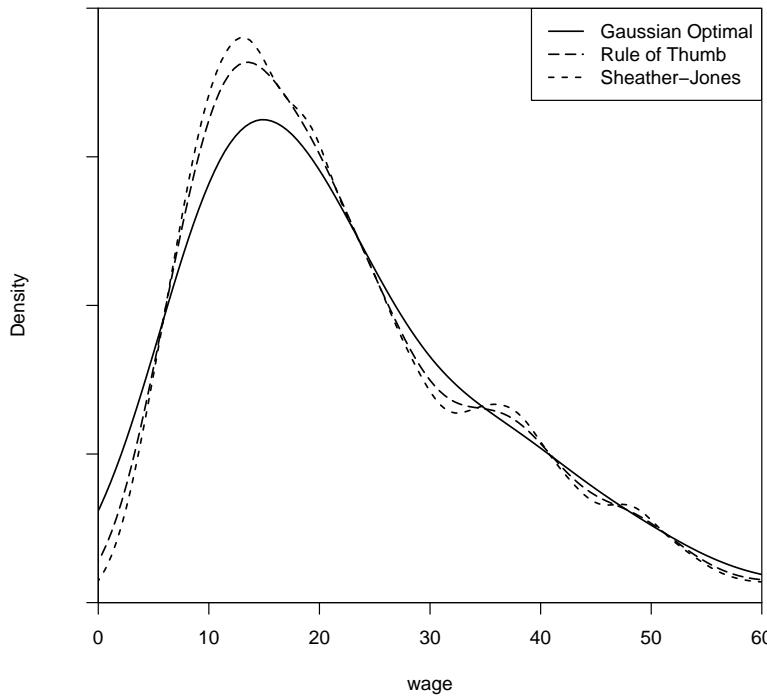


Figure 19.5: Choice of Bandwidth

We can also assess the impact of the choice of kernel function. In Figure 19.6 we display the density estimates calculated using the rectangular, Gaussian, and Epanechnikov kernel functions, and the Sheather-Jones bandwidth (optimized for the Gaussian kernel). The shapes of the three density estimates are very similar, and the Gaussian and Epanechnikov estimates are nearly indistinguishable, with the Gaussian slightly smoother. The estimate using the rectangular kernel, however, is noticeably different. It is erratic and non-smooth. This illustrates how the rectangular kernel is a poor choice for density estimation, and the differences between the Gaussian and Epanechnikov kernels are typically minor.

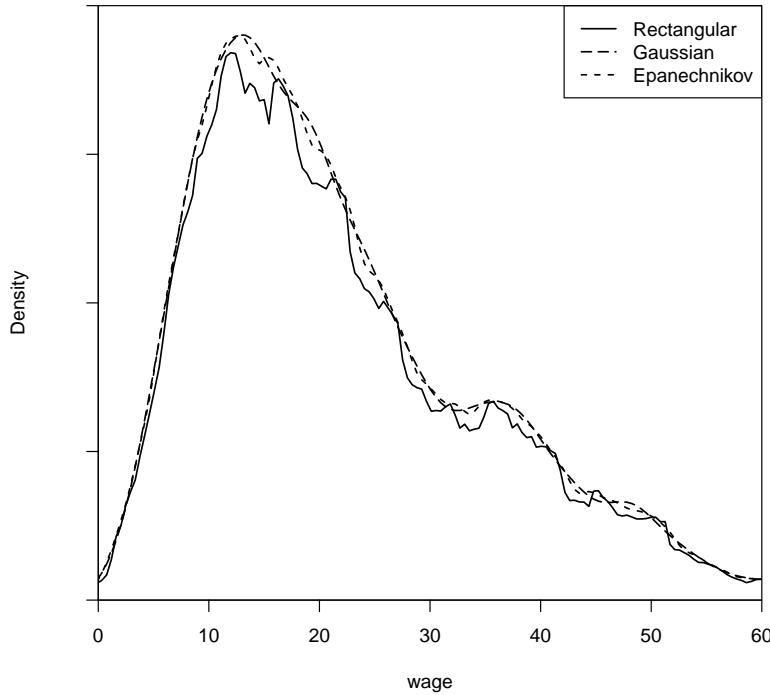


Figure 19.6: Choice of Kernel

### 19.13 Practical Issues in Density Estimation

The most common purpose for a density estimator  $\hat{f}(x)$  is to produce a display such as Figure 19.3. In this case the estimator  $\hat{f}(x)$  is calculated on a grid of values of  $x$  and then plotted. Typically 100 gridpoints is sufficient for a reasonable density plot. However if the density estimate has a section with a steep slope it may be poorly displayed unless more gridpoints are used.

Sometimes it is questionable whether or not a density estimator can be used when the observations are somewhat in between continuous and discrete. For example, many variables are recorded as integers even though the underlying model treats them as continuous. A practical suggestion is to refrain from applying a density estimator unless there are at least 50 distinct values in the dataset.

There is also a practical question about sample size. How large should the sample be to apply a kernel density estimator? The convergence rate is slow, so we should expect to require a larger number of observations than for parametric estimators. I suggest a minimal sample size of  $n = 100$ , and even then estimation precision may be poor.

### 19.14 Computation

In Stata, the kernel density estimator (19.2) can be computed and displayed using the `kdensity` command. By default it uses the Epanechnikov kernel and selects the bandwidth using the reference rule (19.17). One deficiency of the Stata `kdensity` command is that it incorrectly implements the reference rule for kernels with non-unit variances. (This includes all kernel options in Stata other than the Epanechnikov and Gaussian). Consequently the `kdensity` command should only be used with either the Epanechnikov or Gaussian kernel.

R has several commands for density estimation, including the built-in command `density`. By default the latter uses the Gaussian kernel and the reference rule (19.17). The latter can be explicitly specified

using the option `nrd0`. Other kernels and bandwidth selection methods are available, including (19.16) as the option `nrd` and the Sheather-Jones method as the option `SJ`.

Matlab has the built-in function `kdensity`. By default it uses the Gaussian kernel and the reference rule (19.16).

## 19.15 Asymptotic Distribution

In this section we provide asymptotic limit theory for the kernel density estimator (19.2). We first state a consistency result.

**Theorem 19.6** If  $f(x)$  is continuous in  $\mathcal{N}$ , then as  $h \rightarrow 0$  and  $nh \rightarrow \infty$ ,  $\hat{f}(x) \xrightarrow{p} f(x)$ .

This shows that the nonparametric estimator  $\hat{f}(x)$  is consistent for  $f(x)$  under quite minimal assumptions. Theorem 19.6 follows from (19.4) and (19.7).

We now provide an asymptotic distribution theory.

**Theorem 19.7** If  $f''(x)$  is continuous in  $\mathcal{N}$ , then as  $nh \rightarrow \infty$  such that  $h = O(n^{-1/5})$

$$\sqrt{nh} \left( \hat{f}(x) - f(x) - \frac{1}{2} f''(x) h^2 \right) \xrightarrow{d} N\left(0, f(x) R_k\right).$$

The proof is given in Section 19.18.

Theorems 19.1 and 19.2 characterized the asymptotic bias and variance. Theorem 19.7 extends this by applying the Lindeberg central limit theorem to show that the asymptotic distribution is normal.

The convergence result in Theorem 19.7 is different from typical parametric results in two aspects. The first is that the convergence rate is  $\sqrt{nh}$  rather than  $\sqrt{n}$ . This is because the estimator is based on local smoothing, and the effective number of observations for local estimation is  $nh$  rather than the full sample  $n$ . The second notable aspect of the theorem is that the statement includes an explicit adjustment for bias, as the estimator  $\hat{f}(x)$  is centered at  $f(x) + \frac{1}{2} f''(x) h^2$ . This is because the bias is not asymptotically negligible and needs to be acknowledged. The presence of a bias adjustment is typical in the asymptotic theory for kernel estimators.

Theorem 19.7 adds the extra technical condition that  $h = O(n^{-1/5})$ . This strengthens the assumption  $h \rightarrow 0$  by saying it must decline at least at the rate  $n^{-1/5}$ . This condition ensures that the remainder from the bias approximation is asymptotically negligible. It can be weakened somewhat if the smoothness assumptions on  $f(x)$  are strengthened.

## 19.16 Undersmoothing

A technical way to eliminate the bias term in Theorem 19.7 is by using an **undersmoothing** bandwidth. This is a bandwidth  $h$  which converges to zero faster than the optimal rate  $n^{-1/5}$ , thus  $nh^5 = o(1)$ . In practice this means that  $h$  is smaller than the optimal bandwidth so the estimator  $\hat{f}(x)$  is AIMSE inefficient. An undersmoothing bandwidth can be obtained by setting  $h = n^{-\alpha} h_r$  where  $h_r$  is a reference or plug-in bandwidth and  $\alpha > 0$ .

With a smaller bandwidth the estimator has reduced bias and increased variance. Consequently the bias is asymptotically negligible.

**Theorem 19.8** If  $f''(x)$  is continuous in  $\mathcal{N}$ , then as  $nh \rightarrow \infty$  such that  $nh^5 = o(1)$

$$\sqrt{nh}(\hat{f}(x) - f(x)) \xrightarrow{d} N(0, f(x)R_k).$$

This theorem looks identical to Theorem 19.7 with the notable difference that the bias term is omitted. At first, this appears to be a “better” distribution result, as it is certainly preferred to have (asymptotically) unbiased estimators. However this is an incomplete understanding. Theorem 19.7 (with the bias term included) is a better distribution result precisely because it captures the asymptotic bias. Theorem 19.8 is inferior precisely because it avoids characterizing the bias. Another way of thinking about it is that Theorem 19.7 is a more honest characterization of the distribution than Theorem 19.8.

It is worth noting that the assumption  $nh^5 = o(1)$  is the same as  $h = o(n^{-1/5})$ . Some authors will state it one way, and some the other. The assumption means that the estimator  $\hat{f}(x)$  is converging at a slower rate than optimal, and is thus AIMSE inefficient.

While the undersmoothing assumption  $nh^5 = o(1)$  technically eliminates the bias from the asymptotic distribution, it does not actually eliminate the finite sample bias. Thus it is better in practice to view an undersmoothing bandwidth as producing an estimator with “low bias” rather than “zero bias”.

## 19.17 Application

We close the chapter with an empirical illustration. We consider the Duflo, Dupas and Kremer (2011) investigation of the effect of student tracking on testscores. Recall that the core model was a least-squares regression of a standardized version of the variable *testscore* on the dummy variable *tracking*. We can examine the impact on the entire distribution of non-standardized testscores by comparing the estimated densities of testscores for the subsamples with and without tracking. In this application we focus on the sub-sample of girls. In Exercise 19.8 we repeat the application for the sub-sample of boys.

The sub-samples of tracked and non-tracked girls are similar (each has approximately 1400 observations with *testscore* standard deviation of about 9.3). To compare the densities it therefore makes sense to use the same bandwidth for each sample. We first computed the rule-of-thumb bandwidth for each sub-sample, obtaining 1.94 and 1.99 respectively. We then computed the Sheather-Jones bandwidth for each sub-sample, obtaining 1.57 and 1.29 respectively. We estimated the two sub-sample densities using the Gaussian kernel and the three bandwidths 1.96, 1.57, and 1.29. The estimates with the largest bandwidth appear over-smoothed and those with the smallest bandwidth appear under-smoothed, leading us to select the estimates with the middle bandwidth 1.57.

We display the density estimates using  $h = 1.57$  in Figure 19.7. You can see that the *testscore* distribution is highly skewed with a thick right tail. You can also see that the effect of tracking on testscores is more than a simple location shift. In particular, the density of untracked testscores has a significant mass of students with very low testscores. This hump shifts meaningfully to the right for tracked students. At the upper end of the distribution the difference between the densities seems smaller. This means that tracking appears to have a particular effect of improving scores for students with the lowest initial performance.

This illustrates how examination of density estimates can augment regression analysis.

## 19.18 Technical Proofs\*

For simplicity all formal results assume that the kernel  $k(u)$  has bounded support, that is, for some  $a < \infty$ ,  $k(u) = 0$  for  $|u| > a$ . This includes most kernels used in applications with the exception of the Gaussian kernel. The results apply as well to the Gaussian kernel but with a more detailed argument.

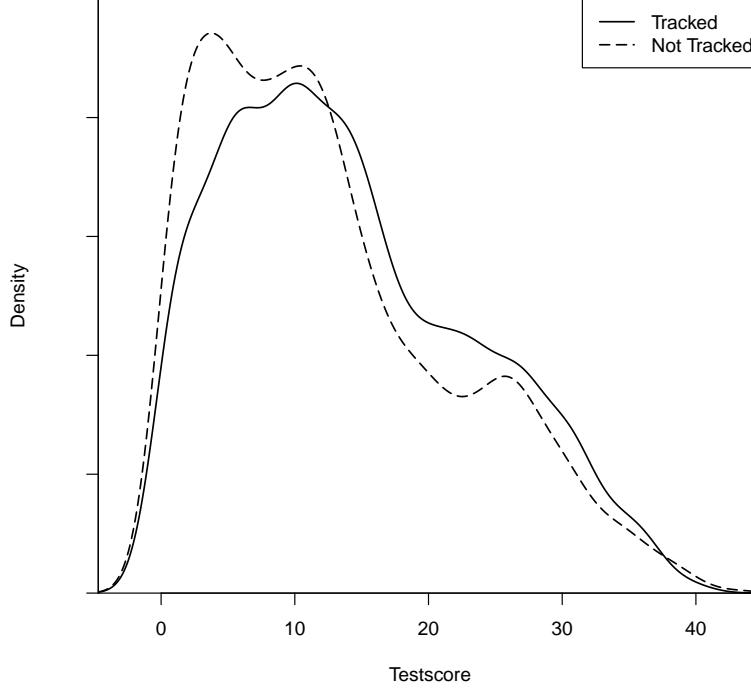


Figure 19.7: Density Estimates of Test Scores for Girls

**Proof of Theorem 19.1.** We first show (19.4). Fix  $\varepsilon > 0$ . Since  $f(x)$  is continuous in some neighborhood  $\mathcal{N}$  there exists a  $\delta > 0$  such that  $|v| \leq \delta$  implies  $|f(x + v) - f(x)| \leq \varepsilon$ . Set  $h \leq \delta/a$ . Then  $|u| \leq a$  implies  $|hu| \leq \delta$  and  $|f(x + hu) - f(x)| \leq \varepsilon$ . Then using (19.3)

$$\begin{aligned} |\mathbb{E}(\hat{f}(x) - f(x))| &= \left| \int_{-a}^a k(u)(f(x + hu) - f(x)) du \right| \\ &\leq \int_{-a}^a k(u) |f(x + hu) - f(x)| du \\ &\leq \varepsilon \int_{-a}^a k(u) du \\ &= \varepsilon. \end{aligned}$$

Since  $\varepsilon$  is arbitrary this shows that  $|\mathbb{E}(\hat{f}(x) - f(x))| = o(1)$  as  $h \rightarrow 0$ , as claimed.

We next show (19.5). By the mean-value theorem

$$\begin{aligned} f(x + hu) &= f(x) + f'(x)hu + \frac{1}{2}f''(x + hu^*)h^2u^2 \\ &= f(x) + f'(x)hu + \frac{1}{2}f''(x)h^2u^2 + \frac{1}{2}(f''(x + hu^*) - f''(x))h^2u^2 \end{aligned}$$

where  $u^*$  lies between 0 and  $u$ . Substituting into (19.3) and using  $\int_{-\infty}^{\infty} k(u) u du = 0$  and  $\int_{-\infty}^{\infty} k(u) u^2 du = 1$  we find

$$\mathbb{E}(\hat{f}(x)) = f(x) + \frac{1}{2}f''(x)h^2 + h^2R(h)$$

where

$$R(h) = \frac{1}{2} \int_{-\infty}^{\infty} (f''(x + hu^*) - f''(x)) u^2 k(u) du.$$

It remains to show that  $R(h) = o(1)$  as  $h \rightarrow 0$ . Fix  $\varepsilon > 0$ . Since  $f''(x)$  is continuous in some neighborhood  $\mathcal{N}$  there exists a  $\delta > 0$  such that  $|v| \leq \delta$  implies  $|f''(x+v) - f''(x)| \leq \varepsilon$ . Set  $h \leq \delta/a$ . Then  $|u| \leq a$  implies  $|hu^*| \leq |hu| \leq \delta$  and  $|f''(x+hu^*) - f''(x)| \leq \varepsilon$ . Then

$$|R(h)| \leq \frac{1}{2} \int_{-\infty}^{\infty} |f''(x+hu^*) - f''(x)| u^2 k(u) du \leq \frac{\varepsilon}{2}.$$

Since  $\varepsilon$  is arbitrary this shows that  $R(h) = o(1)$ . This completes the proof.  $\blacksquare$

**Proof of Theorem 19.2.** As mentioned at the beginning of the section, for simplicity assume  $k(u) = 0$  for  $|u| > a$ .

Equation (19.6) was shown in the text. We now show (19.7). By a derivation similar to that for Theorem 19.1, since  $f(x)$  is continuous in  $\mathcal{N}$

$$\begin{aligned} \frac{1}{h} \mathbb{E} \left( k \left( \frac{x_i - x}{h} \right)^2 \right) &= \int_{-\infty}^{\infty} \frac{1}{h} k \left( \frac{v - x}{h} \right)^2 f(v) dv \\ &= \int_{-\infty}^{\infty} k(u)^2 f(x + hu) du \\ &= \int_{-\infty}^{\infty} k(u)^2 f(x) du + o(1) \\ &= f(x) R_k + o(1). \end{aligned}$$

Then since the observations are i.i.d. and using (19.5)

$$\begin{aligned} nh \text{var}(\hat{f}(x)) &= \frac{1}{h} \text{var} \left( k \left( \frac{x_i - x}{h} \right) \right) \\ &= \frac{1}{h} \mathbb{E} \left( k \left( \frac{x_i - x}{h} \right)^2 \right) - h \left( \mathbb{E} \left( \frac{1}{h} k \left( \frac{x_i - x}{h} \right) \right) \right)^2 \\ &= f(x) R_k + o(1) \end{aligned}$$

as stated.  $\blacksquare$

**Proof of Theorem 19.5 .** By  $m$  applications of integration-by-parts, the fact  $\phi^{(2m)}(x) = He_{2m}(x)\phi(x)$  where  $He_{2m}(x)$  is the  $2m^{\text{th}}$  Hermite polynomial, the fact  $\phi(x)^2 = \phi(\sqrt{2}x)/\sqrt{2\pi}$ , the change-of-variables  $u = x/\sqrt{2}$ , an explicit expression for the Hermite polynomial, the normal moment  $\int_{-\infty}^{\infty} u^{2mj} \phi(u) du = (2m-1)!! = (2m)!/(2^m m!)$ , the Binomial Theorem, and finally  $(2m)!/(2^m m!) = \mu_{2m}$ , we find

$$\begin{aligned} R(\phi^{(m)}) &= \int_{-\infty}^{\infty} \phi^{(m)}(x) \phi^{(m)}(x) dx \\ &= (-1)^m \int_{-\infty}^{\infty} He_{2m}(x) \phi(x)^2 dx \\ &= \frac{(-1)^m}{\sqrt{2\pi}} \int_{-\infty}^{\infty} He_{2m}(x) \phi(\sqrt{2}x) dx \\ &= \frac{(-1)^m}{2\sqrt{\pi}} \int_{-\infty}^{\infty} He_{2m}(u/\sqrt{2}) \phi(u) du \\ &= \frac{(-1)^m}{2\sqrt{\pi}} \int_{-\infty}^{\infty} \sum_{j=0}^m \frac{(2m)!}{j!(2m-2j)!2^m} (-1)^j u^{2m-2j} \phi(u) du \\ &= \frac{(-1)^m (2m)!}{2^{2m+1} m! \sqrt{\pi}} \sum_{j=0}^m \frac{m!}{j!(m-j)!} (-1)^j \\ &= \frac{(2m)!}{2^{2m+1} m! \sqrt{\pi}} \\ &= \frac{\mu_{2m}}{2^{m+1} \sqrt{\pi}} \end{aligned}$$

as claimed. ■

**Proof of Theorem 19.7.** Define

$$y_{ni} = h^{-1/2} \left( k\left(\frac{x_i - x}{h}\right) - \mathbb{E}\left(k\left(\frac{x_i - x}{h}\right)\right) \right)$$

so that

$$\sqrt{nh} (\hat{f}(x) - \mathbb{E}(\hat{f}(x))) = \sqrt{n}\bar{y}.$$

We verify the conditions for the Lindeberg CLT (Theorem 6.12). It is necessary to verify the Lindeberg condition as Lyapunov's condition fails.

In the notation of Theorem 6.12,  $\bar{\sigma}_n^2 = \text{var}(\sqrt{n}\bar{y}) \rightarrow R_k f(x)$  as  $h \rightarrow 0$ . Notice that since the kernel function is positive and finite,  $0 \leq k(u) \leq \bar{k}$ , say, then  $y_{ni}^2 \leq h^{-1}\bar{k}^2$ . Fix  $\varepsilon > 0$ . Then

$$\lim_{n \rightarrow \infty} \mathbb{E}(y_{ni}^2 \mathbf{1}(y_{ni}^2 > \varepsilon n)) \leq \lim_{n \rightarrow \infty} \mathbb{E}\left(y_{ni}^2 \mathbf{1}\left(\bar{k}^2/\varepsilon > nh\right)\right) = 0$$

the final equality since  $nh > \bar{k}^2/\varepsilon$  for sufficiently large  $n$ . This establishes the Lindeberg condition (6.5). The Lindeberg CLT (Theorem 6.12) shows that

$$\sqrt{nh} (\hat{f}(x) - \mathbb{E}(\hat{f}(x))) = \sqrt{n}\bar{y} \xrightarrow{d} N(0, f(x)R_k).$$

Equation (19.5) established

$$\mathbb{E}(\hat{f}(x)) = f(x) + \frac{1}{2} f''(x)h^2 + o(h^2).$$

Since  $h = O(h^{-1/5})$

$$\begin{aligned} \sqrt{nh} \left( \hat{f}(x) - f(x) - \frac{1}{2} f''(x) \kappa_k^2 h^2 \right) &= \sqrt{nh} (\hat{f}(x) - \mathbb{E}(\hat{f}(x))) + o(1) \\ &\xrightarrow{d} N(0, f(x)R_k). \end{aligned}$$

This completes the proof. ■

## Exercises

**Exercise 19.1** If  $x_i^*$  is a random variable with density  $\hat{f}(x)$  from (19.2), show that

- (a)  $\mathbb{E}(x_i^*) = \bar{x}_n$ .
- (b)  $\text{var}(x_i^*) = \hat{\sigma}_x^2 + h^2$ .

**Exercise 19.2** Show that (19.11) minimizes (19.10).

Hint: Differentiate (19.10) with respect to  $h$  and set to 0. This is the first-order condition for optimization. Solve for  $h$ . Check the second-order condition to verify that this is a minimum.

**Exercise 19.3** Suppose  $f(x)$  is the uniform density on  $[0, 1]$ . What does (19.11) suggest should be the optimal bandwidth  $h$ ? How do you interpret this?

**Exercise 19.4** You estimate a density for expenditures measured in dollars, and then re-estimate measuring in millions of dollars, but use the same bandwidth  $h$ . How do you expect the density plot to change? What bandwidth should use so that the density plots have the same shape?

**Exercise 19.5** You have a sample of wages for 1000 men and 1000 women. You estimate the density functions  $\hat{f}_m(x)$  and  $\hat{f}_w(x)$  for the two groups using the same bandwidth  $h$ . You then take the average  $\hat{f}(x) = (\hat{f}_m(x) + \hat{f}_w(x))/2$ . How does this compare to applying the density estimator to the combined sample?

**Exercise 19.6** You increase your sample from  $n = 1000$  to  $n = 2000$ . For univariate density estimation, how does the AIMSE-optimal bandwidth change? If the sample increases from  $n = 1000$  to  $n = 10,000$ ?

**Exercise 19.7** Using the asymptotic formula (19.9) to calculate standard errors  $s(x)$  for  $\hat{f}(x)$ , find an expression which indicates when  $\hat{f}(x) - 2s(x) < 0$ , which means that the asymptotic 95% confidence interval contains negative values. For what values of  $x$  is this likely (that is, around the mode or towards the tails)? If you generate a plot of  $\hat{f}(x)$  with confidence bands, and the latter include negative values, how should you interpret this?

**Exercise 19.8** Take the DDK2011 dataset and the subsample of boys. Estimate the density of *testscores* separately by tracked and not tracked. Are the graphs similar to those for girls in Figure 19.7?

**Exercise 19.9** Take the cps09mar dataset and the subsample of individuals with *education*=20 (professional degree or doctorate), with *experience* between 0 and 40 years.

- (a) Estimate the density of *wages* separately for men and women. Plot on the same graph to compare. Comment.
- (b) Estimate the density of *experience* separately for men and women. Plot on the same graph to compare. Comment on the difference between the density of *wages* and *experience*.

**Exercise 19.10** Take the Invest1993 dataset and the subsample of observations with  $Q \leq 5$ . Estimate the densities of the variables *I* and *Q*.

# Chapter 20

## Nonparametric Regression

### 20.1 Introduction

We now turn to nonparametric estimation of the conditional expectation function

$$\mathbb{E}(y_i | \mathbf{x}_i = \mathbf{x}) = m(\mathbf{x}).$$

Unless an economic model restricts the form of  $m(\mathbf{x})$  to a parametric function,  $m(\mathbf{x})$  can take any nonlinear shape and is therefore **nonparametric**. In this chapter we discuss nonparametric kernel smoothing estimators of  $m(\mathbf{x})$ . These are related to the density estimators explored in the previous chapter. In Chapter 21 we explore estimation by series and sieve methods.

There are many excellent monographs written on nonparametric regression estimation, including Härdle (1990), Fan and Gijbels (1996), Pagan and Ullah (1999), and Li and Racine (2007).

To get started, suppose that there is a single real-valued regressor  $x_i$ . We consider the case of vector-valued regressors later. The nonparametric regression model with a real-valued regressor is

$$\begin{aligned} y_i &= m(x_i) + e_i \\ \mathbb{E}(e_i | x_i) &= 0 \\ \mathbb{E}(e_i^2 | x_i) &= \sigma^2(x_i). \end{aligned}$$

We assume that we have  $n$  observations for the pair  $(y_i, x_i)$ . The goal is to estimate  $m(x)$  either at a single point  $x$  or at a set of points. For most of our theory we focus on estimation at a single point  $x$  which is in the interior of the support of  $x_i$ .

In addition to the conventional regression assumptions, we assume that both  $m(x)$  and  $f(x)$  (the marginal density of  $x_i$ ) are continuous in  $x$ . For our theoretical treatment we assume that the observations are i.i.d. The methods extend to the case of time series but the theory is more advanced. An excellent treatment for the case of dependent data is Fan and Yao (2003). We discuss clustered observations in Section 20.20.

### 20.2 Binned Means Estimator

For clarity, fix the point  $x$  and consider estimation of  $m(x)$ . This is the mean of  $y_i$  for random pairs  $(y_i, x_i)$  such that  $x_i = x$ . If the distribution of  $x_i$  were discrete then we could estimate  $m(x)$  by taking the average of the sub-sample of observations  $y_i$  for which  $x_i = x$ . But when  $x_i$  is continuous then the probability is zero that  $x_i$  exactly equals  $x$ . So there is no sub-sample of observations with  $x_i = x$  and this estimation idea is infeasible. However, if  $m(x)$  is continuous then it should be possible to get a good approximation by taking the average of the observations for which  $x_i$  is close to  $x$ , perhaps for the observations for which  $|x_i - x| \leq h$  for some small  $h > 0$ . As for the case of density estimation we call  $h$  a

**bandwidth.** This **binned means estimator** can be written as

$$\hat{m}(x) = \frac{\sum_{i=1}^n \mathbf{1}(|x_i - x| \leq h) y_i}{\sum_{i=1}^n \mathbf{1}(|x_i - x| \leq h)}. \quad (20.1)$$

This is a step function estimator of the regression function  $m(x)$ .

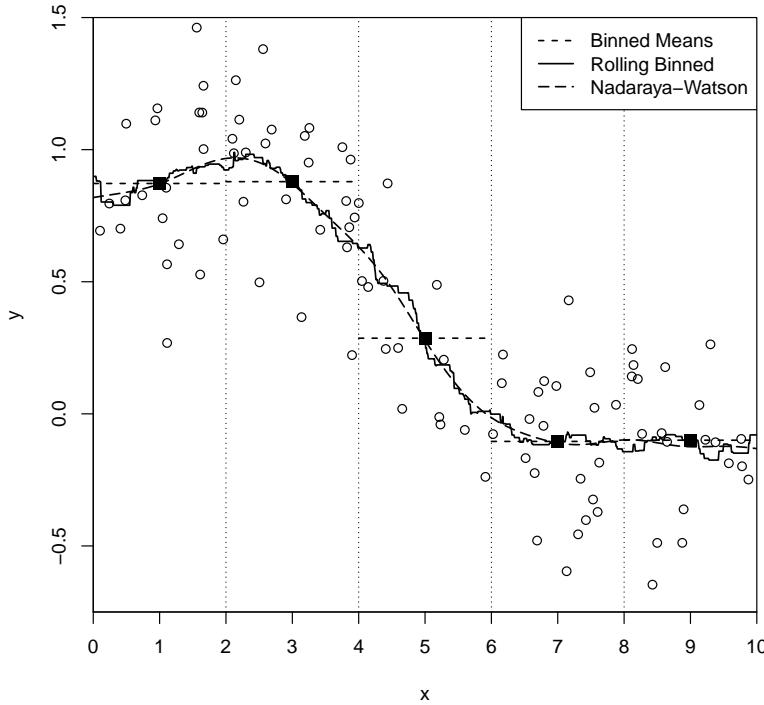


Figure 20.1: Scatter of  $(y_i, x_i)$  and Nadaraya-Watson Regression

To visualize, Figure 20.1 displays a scatter plot of 100 random pairs  $(y_i, x_i)$  generated by simulation. The observations are displayed as the open circles. The estimator (20.1) of  $m(x)$  at  $x = 1$  with  $h = 1$  is the average of the  $y_i$  for the observations such that  $x_i$  falls in the interval  $[0 \leq x_i \leq 2]$ . This estimator is  $\hat{m}(1)$  and is shown on Figure 20.1 by the first solid square. We repeat the calculation (20.1) for  $x = 3, 5, 7$ , and 9, which is equivalent to partitioning the support of  $x_i$  into the bins  $[0, 2], [2, 4], [4, 6], [6, 8]$ , and  $[8, 10]$ . These bins are shown in Figure 20.1 by the vertical dotted lines, and the estimates (20.1) by the solid squares.

The binned estimator  $\hat{m}(x)$  is the step function which is constant within each bin and equals the binned mean. In Figure 20.1 it is displayed by the horizontal dashed lines which pass through the solid squares. This estimate roughly tracks the central tendency of the scatter of the observations  $(y_i, x_i)$ . However, the huge jumps at the edges of the partitions are disconcerting, counter-intuitive, and clearly an artifact of the discrete binning.

If we take another look at the estimation formula (20.1) there is no reason why we need to evaluate (20.1) only on a coarse grid. We can evaluate  $\hat{m}(x)$  for any set of values of  $x$ . In particular, we can evaluate (20.1) on a fine grid of values of  $x$  and thereby obtain a smoother estimate of the CEF. This estimator is displayed in Figure 20.1 with the solid line. We call this estimator “Rolling Binned Means”. This is a generalization of the binned estimator and by construction passes through the solid squares. It turns out that this is a special case of the Nadaraya-Watson estimator considered in the next section. This estimator, while less abrupt than the Binned Means estimator, is still quite jagged.

## 20.3 Kernel Regression

One deficiency with the estimator (20.1) is that it is a step function in  $x$ , even when evaluated on a fine grid. That is why its plot in Figure 20.1 is jagged. The source of the discontinuity is that the weights  $w_i(x)$  are constructed from indicator functions which are themselves discontinuous. If instead the weights are constructed from a continuous kernel function (see Definition 19.1) then  $\hat{m}(x)$  will also be continuous in  $x$ .

A generalization of (20.1) is obtained by replacing the indicator function with a kernel function from Definition 19.1:

$$\hat{m}_{\text{nw}}(x) = \frac{\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right) y_i}{\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)}. \quad (20.2)$$

The estimator (20.2) is known as the **Nadaraya-Watson** estimator, the **kernel regression** estimator, or the **local constant** estimator, and was introduced independently by Nadaraya (1964) and Watson (1964).

The rolling binned means estimator (20.1) is the Nadarya-Watson estimator with the rectangular kernel. The Nadaraya-Watson estimator (20.2) can be constructed with any standard kernel, and is typically estimated using the Gaussian or Epanechnikov kernel. In general we recommend the Gaussian kernel since it produces an estimator  $\hat{m}_{\text{nw}}(x)$  which possesses derivatives of all orders.

The bandwidth  $h$  plays a similar role in kernel regression as in kernel density estimation. Namely, larger values of  $h$  will result in estimates  $\hat{m}_{\text{nw}}(x)$  which are smoother in  $x$ , and smaller values of  $h$  will result in estimates which are more erratic. It might be helpful to consider the two extreme cases  $h \rightarrow 0$  and  $h \rightarrow \infty$ . As  $h \rightarrow 0$  we can see that  $\hat{m}_{\text{nw}}(x_i) \rightarrow y_i$  (if the values of  $x_i$  are unique), so that  $\hat{m}_{\text{nw}}(x)$  is simply the scatter of  $y_i$  on  $x_i$ . In contrast, as  $h \rightarrow \infty$  then  $\hat{m}_{\text{nw}}(x) \rightarrow \bar{y}$ , the sample mean. For intermediate values of  $h$ ,  $\hat{m}_{\text{nw}}(x)$  will smooth between these two extreme cases.

The estimator (20.2) using the Gaussian kernel and  $h = 1/\sqrt{3}$  is also displayed in Figure 20.1 with the long dashes. As you can see, this estimator appears to be much smoother than that using the binned estimator, but tracks exactly the same path. The bandwidth  $h = 1/\sqrt{3}$  for the Gaussian kernel is equivalent to the bandwidth  $h = 1$  for the binned estimator because the latter is a kernel estimator using the rectangular kernel scaled to have a standard deviation of 1/3.

## 20.4 Local Linear Estimator

The Nadaraya-Watson (NW) estimator is often called a **local constant** estimator as it locally (about  $x$ ) approximates  $m(x)$  as a constant function. One way to see this is to observe that  $\hat{m}(x)$  solves the minimization problem

$$\hat{m}_{\text{nw}}(x) = \underset{m}{\operatorname{argmin}} \sum_{i=1}^n k\left(\frac{x_i - x}{h}\right) (y_i - m)^2.$$

This is a weighted regression of  $y_i$  on an intercept only.

This means that the NW estimator is making the local approximation  $m(x_i) \approx m(x)$  for  $x_i \approx x$ , which means it is making the approximation

$$y_i = m(x_i) + e_i \approx m(x) + e_i.$$

The NW estimator is a local estimator of this approximate model using weighted least squares.

This interpretation suggests that we can construct alternative nonparametric estimators of  $m(x)$  by alternative local approximations. Many such local approximations are possible. A popular choice is the **Local Linear** (LL) approximation. Instead of the approximation  $m(x_i) \approx m(x)$ , LL uses the linear approximation  $m(x_i) \approx m(x) + m'(x)(x_i - x)$ . Thus

$$\begin{aligned} y_i &= m(x_i) + e_i \\ &\approx m(x) + m'(x)(x_i - x) + e_i. \end{aligned}$$

The LL estimator then applies weighted least squares similarly to the NW estimator.

One way to represent the LL estimator is as the solution to the minimization problem

$$\{\hat{m}_{\text{LL}}(x), \hat{m}'_{\text{LL}}(x)\} = \underset{\alpha, \beta}{\operatorname{argmin}} \sum_{i=1}^n k\left(\frac{x_i - x}{h}\right) (y_i - \alpha - \beta(x_i - x))^2.$$

Another is to write the approximating model as

$$y_i \approx \mathbf{z}_i(x)' \boldsymbol{\beta}(x) + e_i$$

where  $\boldsymbol{\beta}(x) = (m(x), m'(x))'$  and

$$\mathbf{z}_i(x) = \begin{pmatrix} 1 \\ x_i - x \end{pmatrix}.$$

This is a linear regression with regressor vector  $\mathbf{z}_i(x)$  and coefficient vector  $\boldsymbol{\beta}(x)$ . Applying weighted least squares with the kernel weights we obtain the LL estimator

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{LL}}(x) &= \left( \sum_{i=1}^n k\left(\frac{x_i - x}{h}\right) \mathbf{z}_i(x) \mathbf{z}_i(x)' \right)^{-1} \sum_{i=1}^n k\left(\frac{x_i - x}{h}\right) \mathbf{z}_i(x) y_i \\ &= (\mathbf{Z}' \mathbf{K} \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{K} \mathbf{y} \end{aligned}$$

where  $\mathbf{K} = \text{diag}\{k((x_1 - x)/h), \dots, k((x_n - x)/h)\}$ ,  $\mathbf{Z}$  is the stacked  $\mathbf{z}_i(x)'$ , and  $\mathbf{y}$  is the stacked  $y_i$ . This expression generalizes the Nadaraya-Watson estimator as the latter is obtained by setting  $\mathbf{z}_i(x) = 1$ . Notice that the matrices  $\mathbf{Z}$  and  $\mathbf{K}$  depend on  $x$  and  $h$ .

The local linear estimator was first suggested by Stone (1977) and came into prominence through the work of Fan (1992, 1993).

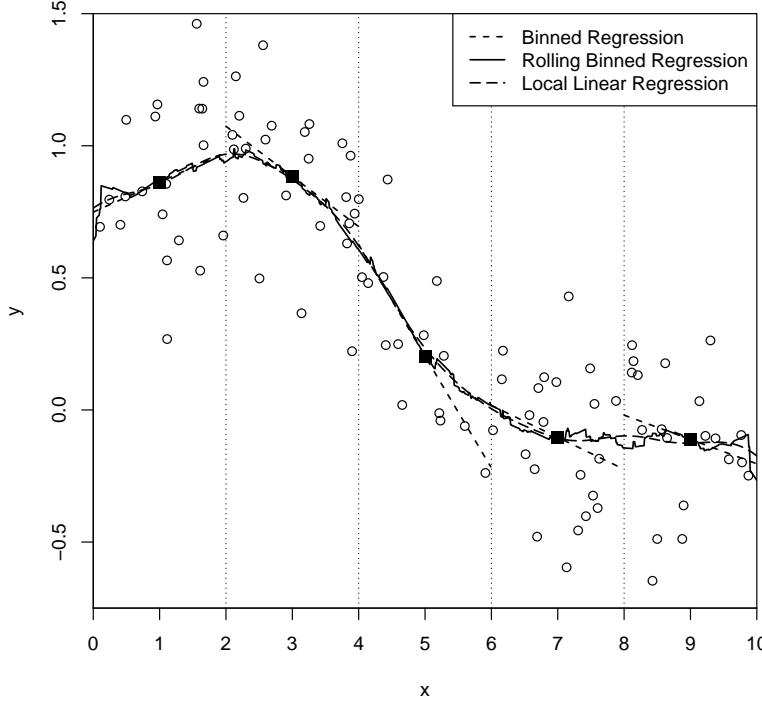
To visualize, Figure 20.2 displays the scatter plot of the same 100 observations from Figure 20.1 divided into the same five bins. A linear regression is fit to the observations in each bin. These five fitted regression lines are displayed by the short dashed lines. This “binned regression estimator” produces a flexible approximation for the mean function, but has large jumps at the edges of the partitions. The midpoints of each of these five regression lines are displayed by the solid squares, and could be viewed as the target estimate for the binned regression estimator. A rolling version of the binned regression estimator moves these estimation windows continuously across the support of  $x$ , and is displayed by the solid line. This corresponds to the local linear estimator with a rectangular kernel and a bandwidth of  $h = 1/\sqrt{3}$ . By construction, this line passes through the solid squares. To obtain a smoother estimator, we replace the rectangular with the Gaussian kernel (using the same bandwidth  $h = 1/\sqrt{3}$ ). We display these estimates with the long dashes. This has the same shape as the rectangular kernel estimate (rolling binned regression) but is visually much smoother. We label this the “Local Linear” estimator since it is the standard implementation.

One interesting feature is that as  $h \rightarrow \infty$ , the LL estimator approaches the full-sample linear least-squares estimator  $\hat{m}_{\text{LL}}(x) \rightarrow \hat{\alpha} + \hat{\beta}x$ . That is because as  $h \rightarrow \infty$  all observations receive equal weight regardless of  $x$ . In this sense we can see that the LL estimator is a flexible generalization of the linear OLS estimator.

Another useful property of the LL estimator is that it simultaneously provides estimates of the regression function  $m(x)$  and its slope  $m'(x)$  at  $x$ .

## 20.5 Local Polynomial Estimator

The NW and LL estimators are both special cases of the **local polynomial estimator**. The idea is to approximate the regression function  $m(x)$  by a polynomial of fixed degree  $p$ , and then estimate locally using the kernel weights.

Figure 20.2: Scatter of  $(y_i, x_i)$  and Local Linear Regression

The approximating model is a  $p^{th}$  order Taylor series approximation

$$\begin{aligned} y_i &= m(x_i) + e_i \\ &\simeq m(x) + m'(x)(x_i - x) + \cdots + m^{(p)}(x) \frac{(x_i - x)^p}{p!} + e_i \\ &= \mathbf{z}_i(x)' \boldsymbol{\beta}(x) + e_i \end{aligned}$$

where

$$\mathbf{z}_i(x) = \begin{pmatrix} 1 \\ x_i - x \\ \vdots \\ \frac{(x_i - x)^p}{p!} \end{pmatrix} \quad \boldsymbol{\beta}(x) = \begin{pmatrix} m(x) \\ m'(x) \\ \vdots \\ m^{(p)}(x) \end{pmatrix}.$$

The estimator is

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{LP}(x) &= \left( \sum_{i=1}^n k \left( \frac{x_i - x}{h} \right) \mathbf{z}_i(x) \mathbf{z}_i(x)' \right)^{-1} \left( \sum_{i=1}^n k \left( \frac{x_i - x}{h} \right) \mathbf{z}_i(x) y_i \right) \\ &= (\mathbf{Z}' \mathbf{K} \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{K} \mathbf{y}. \end{aligned}$$

Notice that this expression includes the Nadaraya-Watson and local linear estimators as special cases with  $p = 0$  and  $p = 1$ , respectively.

There is a trade-off between the polynomial order  $p$  and the local smoothing bandwidth  $h$ . By increasing  $p$  we improve the model approximation and thereby can use a larger bandwidth  $h$ . On the other hand, increasing  $p$  increases estimation variance.

## 20.6 Asymptotic Bias

Since  $\mathbb{E}(y_i | x_i) = m(x_i)$ , the conditional mean of the Nadaraya-Watson estimator is

$$\begin{aligned}\mathbb{E}(\hat{m}_{\text{nw}}(x) | \mathbf{X}) &= \frac{\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)\mathbb{E}(y_i | x_i)}{\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)} \\ &= \frac{\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)m(x_i)}{\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)}.\end{aligned}\tag{20.3}$$

We can simplify this expression as  $n \rightarrow \infty$ .

The following regularity conditions will be maintained through the chapter. Let  $f(x)$  denote the marginal density of  $x_i$  and let  $\sigma^2(x) = \mathbb{E}(e_i^2 | x_i = x)$  denote the conditional variance of  $e_i = y_i - m(x_i)$ .

### Assumption 20.1

1.  $h \rightarrow 0$ .
2.  $nh \rightarrow \infty$ .
3.  $m(x)$ ,  $f(x)$  and  $\sigma^2(x)$  are continuous in some neighborhood  $\mathcal{N}$  of  $x$ .
4.  $f(x) > 0$ .

These conditions are similar to those used for the asymptotic theory for kernel density estimation. The assumptions that  $h \rightarrow 0$  and  $nh \rightarrow \infty$  means that the bandwidth gets small yet the number of observations in the estimation window about  $x$  diverges to infinity. Assumption 20.1.3 are minimal smoothness conditions on the conditional mean  $m(x)$ , marginal density  $f(x)$  and conditional variance  $\sigma^2(x)$ . Assumption 20.1.4 specifies that the marginal density is non-zero. This is required since we are estimating the conditional mean at  $x$ , so there needs to be a non-trivial number of observations for  $x_i$  near  $x$ .

**Theorem 20.1** Suppose Assumption 20.1 holds and  $m''(x)$  and  $f'(x)$  are continuous in  $\mathcal{N}$ . Then

$$1. \mathbb{E}(\hat{m}_{\text{nw}}(x) | \mathbf{X}) = m(x) + h^2 B_{\text{nw}}(x) + o_p(h^2) + O_p\left(\sqrt{\frac{h}{n}}\right)$$

where

$$B_{\text{nw}}(x) = \frac{1}{2}m''(x) + f(x)^{-1}f'(x)m'(x).$$

$$2. \mathbb{E}(\hat{m}_{\text{LL}}(x) | \mathbf{X}) = m(x) + h^2 B_{\text{LL}}(x) + o_p(h^2) + O_p\left(\sqrt{\frac{h}{n}}\right)$$

where

$$B_{\text{LL}}(x) = \frac{1}{2}m''(x).$$

The proof for the Nadaraya-Watson estimator is presented in Section 20.25. For a proof for the local linear estimator see Fan and Gijbels (1996).

In addition to Assumption 20.1, Theorem 20.1 adds additional smoothness conditions on  $m(x)$  and  $f(x)$ .

We call the terms  $h^2 B_{\text{nw}}(x)$  and  $h^2 B_{\text{LL}}(x)$  the **asymptotic bias** of the estimators.

Theorem 20.1 shows that the asymptotic bias of the Nadaraya-Watson and local linear estimators is proportional to the squared bandwidth  $h^2$  (the degree of smoothing) and to the functions  $B_{\text{nw}}(x)$  and  $B_{\text{LL}}(x)$ . The asymptotic bias of the local linear estimator depends on the curvature (second derivative) of the CEF function  $m(x)$  similarly to the asymptotic bias of the kernel density estimator in Theorem 19.1. When  $m''(x) < 0$  then  $\hat{m}_{\text{LL}}(x)$  is downwards biased. When  $m''(x) > 0$  then  $\hat{m}_{\text{LL}}(x)$  is upwards biased. Local averaging smooths  $m(x)$ , inducing bias, and this bias is increasing in the level of curvature of  $m(x)$ . This is called **smoothing bias**.

The asymptotic bias of the Nadaraya-Watson estimator adds a second term which depends on the first derivatives of  $m(x)$  and  $f(x)$ . This is because the Nadaraya-Watson estimator is a local average. If the density is upward sloped at  $x$  (if  $f'(x) > 0$ ) then there are (on average) more observations to the right of  $x$  than to the left, so a local average will be biased if  $m(x)$  has a non-zero slope. In contrast the bias of the local linear estimator does not depend on the local slope  $m'(x)$  since it locally fits a linear regression. The fact that the bias of the local linear estimator has fewer terms than the bias of the Nadaraya-Watson estimator (and is invariant to the slope  $m'(x)$ ) justifies the claim that the local linear estimator has generically reduced bias relative to Nadaraya-Watson.

We illustrate asymptotic smoothing bias in Figure 20.3. The solid line is the true conditional mean for the data displayed in Figures 20.1 and 20.2. The dashed lines are the asymptotic approximations to the expectation  $m(x) + h^2 B(x)$  for bandwidths  $h = 1/2$ ,  $h = 1$ , and  $h = 3/2$ . (The asymptotic biases of the NW and LL estimators are the same since  $x_i$  has a uniform distribution.) You can see that there is minimal bias for the smallest bandwidth, but considerable bias for the largest. The dashed lines are smoothed versions of the conditional mean, attenuating the peaks and valleys.

Smoothing bias is a natural by-product of non-parametric estimation of non-linear functions. It can only be reduced by using a small bandwidth. As we see in the following section this will result in high estimation variance.

## 20.7 Asymptotic Variance

From (20.3) we deduce that

$$\hat{m}_{\text{nw}}(x) - \mathbb{E}(\hat{m}_{\text{nw}}(x) | \mathbf{X}) = \frac{\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right) e_i}{\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)}.$$

Since the denominator is a function only of  $x_i$ , and the numerator is linear in  $e_i$ , we can calculate that the finite sample variance of  $\hat{m}_{\text{nw}}(x)$  is

$$\text{var}(\hat{m}_{\text{nw}}(x) | \mathbf{X}) = \frac{\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)^2 \sigma^2(x_i)}{\left(\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)\right)^2}. \quad (20.4)$$

We can simplify this expression as  $n \rightarrow \infty$ . Let  $\sigma^2(x) = \mathbb{E}(e_i^2 | x_i = x)$  denote the conditional variance of  $e_i = y_i - m(x_i)$ .

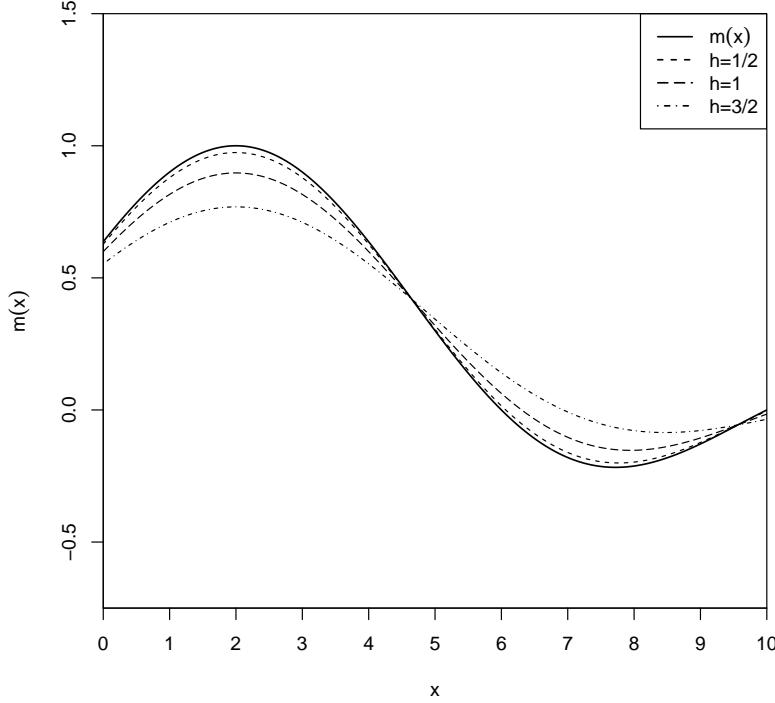


Figure 20.3: Asymptotic Smoothing Bias

**Theorem 20.2** Under Assumption 20.1,

1.  $\text{var}(\hat{m}_{\text{nw}}(x) | \mathbf{X}) = \frac{R_k \sigma^2(x)}{f(x) nh} + o\left(\frac{1}{nh}\right).$
2.  $\text{var}(\hat{m}_{\text{LL}}(x) | \mathbf{X}) = \frac{R_k \sigma^2(x)}{f(x) nh} + o\left(\frac{1}{nh}\right).$

The proof for the Nadaraya-Watson estimator is presented in Section 20.25. For the local linear estimator see Fan and Gijbels (1996).

We call the leading terms in Theorem 20.2 the **asymptotic variance** of the estimators. Theorem 20.2 shows that the asymptotic variance of the two estimators are identical. The asymptotic variance is proportional to the roughness  $R_k$  of the kernel  $k(u)$  and to the conditional variance  $\sigma^2(x)$  of the regression error. It is inversely proportional to the effective number of observations  $nh$  and to the marginal density  $f(x)$ . This expression reflects the fact that the estimators are local estimators. The precision of  $\hat{m}(x)$  is low for regions where  $e_i$  has a large conditional variance and/or  $x_i$  has a low density (where there are relatively few observations).

## 20.8 AIMSE

One implication of Theorem 20.8 is that we can define the asymptotic MSE of  $\hat{m}(x)$  as the sum of the squared asymptotic bias and asymptotic variance:

$$\text{AMSE}(x) \stackrel{\text{def}}{=} h^4 B(x)^2 + \frac{R_k \sigma^2(x)}{nh f(x)}$$

where  $B(x) = B_{\text{nw}}(x)$  for the Nadaraya-Watson estimator and  $B(x) = B_{\text{LL}}(x)$  for the local linear estimator. This is the asymptotic MSE for  $\hat{m}(x)$  for a single point  $x$ .

A global measure of fit can be obtained by integrating AMSE( $x$ ). It is standard to weight the AMSE by  $f(x)w(x)$  for some integrable weight function  $w(x)$ . This is called the asymptotic integrated MSE (AIMSE). Let  $S$  be the support of  $x_i$  (the region where  $f(x) > 0$ ).

$$\begin{aligned}\text{AIMSE} &= \int_S \left( h^4 B(x)^2 + \frac{R_k \sigma^2(x)}{nh f(x)} \right) f(x) w(x) dx \\ &= h^4 \bar{B} + \frac{R_k}{nh} \bar{\sigma}^2\end{aligned}\quad (20.5)$$

where

$$\begin{aligned}\bar{B} &= \int_S B(x)^2 f(x) w(x) dx \\ \bar{\sigma}^2 &= \int_S \sigma^2(x) w(x) dx.\end{aligned}$$

The weight function  $w(x)$  can be omitted if  $S$  is bounded. Otherwise, a common choice is  $w(x) = \mathbf{1}_{(\xi_1 \leq x \leq \xi_2)}$ . An integrable weight function is needed when  $x_i$  has unbounded support to ensure that  $\bar{\sigma}^2 < \infty$ .

The form of the AIMSE is similar to that for kernel density estimation. It has two terms (squared bias and variance). The first is increasing in the bandwidth  $h$  and the second is decreasing in  $h$ . Thus the choice of  $h$  affects AIMSE with a trade-off between these two components. Similarly to density estimation, we can calculate the bandwidth which minimizes the AIMSE. (See Exercise 20.2.) The solution is given in the following theorem.

**Theorem 20.3** The bandwidth which minimizes the AIMSE (20.5) is

$$h_0 = \left( \frac{R_k \bar{\sigma}^2}{4\bar{B}} \right)^{1/5} n^{-1/5}. \quad (20.6)$$

With  $h \sim n^{-1/5}$  then  $\text{AIMSE}(\hat{m}(x)) = O(n^{-4/5})$ .

This result characterizes the AIMSE-optimal bandwidth. This bandwidth satisfies the rate  $h = cn^{-1/5}$  which is the same rate as for kernel density estimation. The optimal constant  $c$  depends on the kernel  $k$ , the weighted average squared bias  $\bar{B}$ , and the weighted average variance  $\bar{\sigma}^2$ . The constant  $c$  is different from that for density estimation. A common mis-interpretation is to set  $h = n^{-1/5}$ , which is equivalent to setting  $c = 1$  and is completely arbitrary. Instead, an empirical bandwidth selection rule should be used in practice. Another common error is to use the Silverman Rule-of-Thumb  $h = 0.9\hat{\sigma}_x n^{-1/5}$ . This is appropriate for estimation of the density of  $x_i$ , but is irrelevant for estimation of the conditional mean of  $y_i$  given  $x_i$ .

The AIMSE (20.5) depends on the kernel  $k(u)$  only through the constant  $R_k$ . Since the Epanechnikov kernel has the smallest value of  $R_k$ , it is also the kernel which produces the smallest AIMSE for the NW and LL estimators.

**Theorem 20.4** The AIMSE (20.5) is minimized by the Epanechnikov kernel for the Nadaraya-Watson and Local Linear regression estimators.

Despite this result, we recommend the Gaussian kernel for regression estimation for the same reasons as for density estimation. The Gaussian kernel is nearly as efficient as the Epanechnikov and produces smoother estimates. The latter is especially important as we are often interested in marginal effects.

## 20.9 Boundary Bias

One strong advantage of the local linear over the Nadaraya-Watson estimator is that the LL has better performance at the boundary of the support of  $x_i$ . The NW estimator has excessive smoothing bias near the boundaries. In many contexts in econometrics the boundaries are of great interest. In these contexts it is strongly recommended to use the local linear estimator (or any local polynomial estimator with  $p \geq 1$ ).

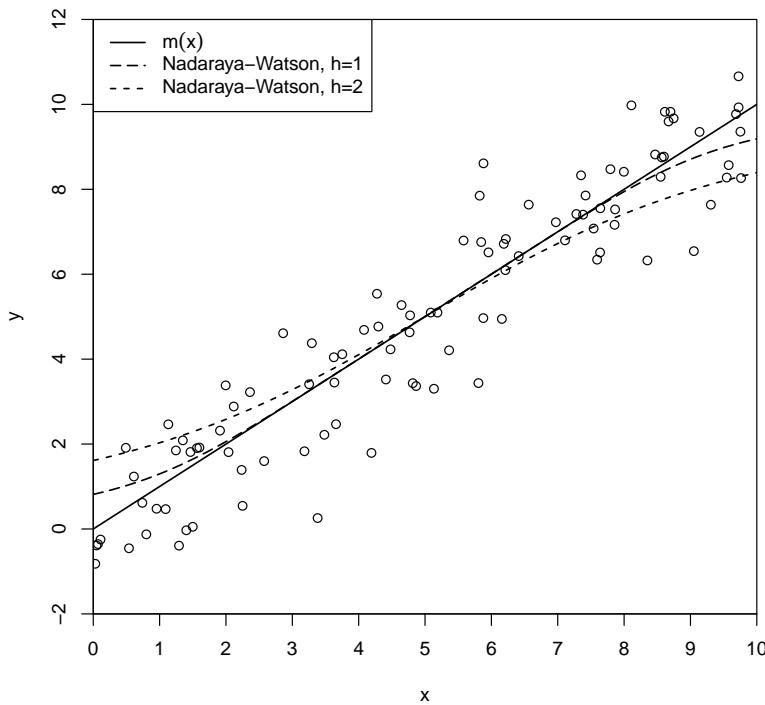


Figure 20.4: Boundary Bias

To understand the problem it may be helpful to example Figure 20.4. This shows a scatter plot of 100 observations generated as  $x_i \sim U[0, 10]$  and  $y_i \sim N(x_i, 1)$  so that  $m(x) = x$ . Suppose we are interested the conditional mean  $m(0)$  at the lower boundary  $x = 0$ . If we use a Nadaraya-Watson estimator it equals a weighted average of the  $y_i$  observations for small values of  $|x_i|$ . Since  $x_i \geq 0$ , these are all observations for which  $m(x_i) \geq m(0)$ , and therefore  $\hat{m}_{\text{nw}}(0)$  is biased upwards. Symmetrically, the Nadaraya-Watson estimator at the upper boundary  $x = 10$  is a weighted average of observations for which  $m(x_i) \leq m(10)$  and therefore  $\hat{m}_{\text{nw}}(10)$  is biased downwards.

In contrast, the local linear estimators  $\hat{m}_{\text{LL}}(0)$  and  $\hat{m}_{\text{LL}}(10)$  are unbiased in this example since  $m(x)$  is linear in  $x$ . The local linear estimator fits a linear regression line. Since the mean is correctly specified there is no estimation bias.

The exact bias<sup>1</sup> of the NW estimator is shown in Figure 20.4 by the dashed lines. The long dashes is the mean  $\mathbb{E}(\hat{m}_{\text{nw}}(x))$  for  $h = 1$  and the short dashes is the mean  $\mathbb{E}(\hat{m}_{\text{nw}}(x))$  for  $h = 2$ . We can see that the

<sup>1</sup>Calculated by simulation from 10,000 simulation replications.

bias is substantial. For  $h = 2$  the bias is visible for all values of  $x$ . For the smaller bandwidth  $h = 1$  the bias is minimal for  $x$  in the central range of the support, but is still quite substantial for  $x$  near the boundaries.

To calculate the asymptotic smoothing bias at the boundary we can revisit the proof of Theorem 20.1.1 which calculated the asymptotic bias at interior points. Equation (20.28) calculates the bias of the numerator of the estimator, expressed as an integral over the marginal density. Evaluated at a lower boundary point this density is only positive for  $u \geq 0$ , so the integral is over the positive region  $[0, \infty)$ . This applies as well to equation (20.30) and the equations which follow. In this case the leading term of this expansion is the first term (20.31) which is proportional to  $h$  rather than  $h^2$ . Completing the calculations we find the following.

**Theorem 20.5** Let the support of  $x_i$  be  $S = [\underline{x}, \bar{x}]$ . Suppose Assumption 20.1 holds and  $m''(x)$ ,  $\sigma^2(x)$  and  $f'(x)$  are right continuous at  $\underline{x}$ , left continuous at  $\bar{x}$ , and  $f(\underline{x}+) > 0$  and  $f(\bar{x}-) > 0$ . Then

1.  $\mathbb{E}(\hat{m}_{\text{nw}}(\underline{x}) | \mathbf{X}) = m(\underline{x}) + hm'(\underline{x})\mu_k + o_p(h) + O_p\left(\sqrt{\frac{h}{n}}\right)$   
 $\mathbb{E}(\hat{m}_{\text{nw}}(\bar{x}) | \mathbf{X}) = m(\bar{x}) - hm'(\bar{x})\mu_k + o_p(h) + O_p\left(\sqrt{\frac{h}{n}}\right)$   
where  $\mu_k = 2 \int_0^\infty uk(u)du$ .
2.  $\mathbb{E}(\hat{m}_{\text{LL}}(\underline{x}) | \mathbf{X}) = m(\underline{x}) + h^2 m''(\underline{x})/2 + o_p(h^2) + O_p\left(\sqrt{\frac{h}{n}}\right)$   
 $\mathbb{E}(\hat{m}_{\text{LL}}(\bar{x}) | \mathbf{X}) = m(\bar{x}) + h^2 m''(\bar{x})/2 + o_p(h^2) + O_p\left(\sqrt{\frac{h}{n}}\right)$

Theorem 20.5 shows that the asymptotic bias of the NW estimator at the boundary is  $O(h)$  and depends on the slope of  $m(x)$  at the boundary. This means that when the slope is positive the NW estimator is upward biased at the lower boundary and downward biased at the upper boundary. In contrast, the asymptotic bias of the LL estimator at the boundary is the same as at interior points, is  $O(h^2)$  and is invariant to the slope of  $m(x)$ . Our interpretation of Theorem 20.5 is that the NW estimator will tend to have much higher bias near boundary points.

Taking Theorems 20.1, 20.2 and 20.5 together, the local linear estimator has superior asymptotic properties relative to the NW estimator. For a given bandwidth  $h$  the two estimators have the same asymptotic variance, but have different bias properties. At interior points both estimators have asymptotic biases of order  $O(h^2)$  but at boundary points the asymptotic bias of the NW estimator is  $O(h)$ , which is of higher order. Furthermore, at interior points the bias of the LL estimator is invariant to the slope of  $m(x)$  and its asymptotic bias only depends on the second derivative, while the bias of the NW estimator depends on both the first and second derivatives. For these reasons, it is generally recommended to use the local linear estimator rather than the Nadaraya-Watson estimator.

## 20.10 Reference Bandwidth

The NW, LL and LP estimators depend on a bandwidth, and without an empirical rule for selection of  $h$  the methods are incomplete. It is useful to have a reference bandwidth which mimics the optimal bandwidth in a simplified setting and provides a baseline for further investigations.

Theorem 20.3 and a little re-writing reveals that the optimal bandwidth takes the form

$$h_0 = \left(\frac{R_k}{4}\right)^{1/5} \left(\frac{\bar{\sigma}^2}{nB}\right)^{1/5} \simeq 0.58 \left(\frac{\bar{\sigma}^2}{nB}\right)^{1/5} \quad (20.7)$$

where the approximation holds for all single-peaked kernels by similar calculations<sup>2</sup> as in Section 19.10.

As we discussed in Section 19.10, Silverman developed a reference bandwidth  $h = 0.9\hat{\sigma}_x n^{-1/5}$  for density estimation. A common error is to use this rule for regression estimation. This is a mistake as the two smoothing problems are quite different, and there is no reason to expect a bandwidth appropriate for density estimation will be a good bandwidth for regression estimation.

However, a reference approach can be used to develop a rule-of-thumb for regression estimation. In particular, Fan and Gijbels (1996, Section 4.2) develop what they call the ROT (rule of thumb) bandwidth for the local linear estimator. We now describe their derivation.

First, set  $w(x) = \mathbf{1}(\xi_1 \leq x \leq \xi_2)$ . Second, form a preliminary estimator of the regression function  $m(x)$  using a  $q^{th}$ -order polynomial regression

$$m(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_q x^q.$$

(In particular they suggest  $q = 4$  but this is not essential to their recommendation.). By least-squares we obtain the coefficient estimates  $\hat{\beta}_0, \dots, \hat{\beta}_q$  and implied second derivative  $\hat{m}''(x) = 2\hat{\beta}_2 + 6\hat{\beta}_3 x + 12\hat{\beta}_4 x^2$  (for the case  $q = 4$ ). Third, notice that  $\bar{B}$  can be written as an expectation

$$\bar{B} = \mathbb{E}(B(x_i)^2 w(x_i)) = \mathbb{E}\left(\left(\frac{1}{2}m''(x_i)\right)^2 \mathbf{1}(\xi_1 \leq x_i \leq \xi_2)\right).$$

A moment estimator is

$$\hat{B} = \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{2}\hat{m}''(x_i)\right)^2 \mathbf{1}(\xi_1 \leq x_i \leq \xi_2). \quad (20.8)$$

Third, assume that the regression error is homoskedastic  $\mathbb{E}(e_i^2 | x_i) = \sigma^2$  so that  $\bar{\sigma}^2 = \sigma^2(\xi_2 - \xi_1)$ . Estimate  $\sigma^2$  by the error variance estimate  $\hat{\sigma}^2$  from the preliminary regression. Plugging these into (20.7) we obtain the reference bandwidth

$$h_{\text{rot}} = 0.58 \left( \frac{\hat{\sigma}^2(\xi_2 - \xi_1)}{n\hat{B}} \right)^{1/5}. \quad (20.9)$$

Fan and Gijbels (1996) call this the rule-of-thumb (ROT) bandwidth.

Fan and Gijbels developed similar rules for higher-order odd local polynomial estimators, but not for the local constant (Nadaraya-Watson) estimator. However, we can derive a ROT for the NW as well by using a reference model for the marginal density  $f(x)$ . A particularly convenient choice is the uniform density, under which  $f'(x) = 0$  and the optimal bandwidths for NW and LL coincide. This motivates using (20.9) as a ROT bandwidth for both the LL and NW estimators.

We now comment on the choice of the weight region  $[\xi_1, \xi_2]$ . When  $x_i$  has bounded support then  $[\xi_1, \xi_2]$  can be set equal to this support. Otherwise,  $[\xi_1, \xi_2]$  can be set equal to the region of interest for  $\hat{m}(x)$ , or the endpoints can be set to equal fixed quantiles (e.g. 0.05 and 0.95) of the distribution of  $x_i$ .

To illustrate, take the data shown in Figures 20.1 and 20.2. If we fit 4<sup>th</sup> order polynomial we find  $\hat{m}(x) = .49 + .70x - .28x^2 - .033x^3 - .0012x^4$  which implies  $\hat{m}''(x) = -.56 - .20x - .014x^2$ . Setting  $[\xi_1, \xi_2] = [0, 10]$  to equal to the support of  $x_i$ , we find  $\hat{B} = 0.00889$ . The residuals from this polynomial regression have variance  $\hat{\sigma}^2 = 0.0687$ . Plugging these into (20.9) we find  $h_{\text{rot}} = 0.551$ , which is similar to the one used in Figures 20.1 and 20.2.

## 20.11 Nonparametric Residuals and Prediction Errors

Given any nonparametric regression estimator  $\hat{m}(x)$  the fitted regression at  $x = x_i$  is  $\hat{m}(x_i)$  and the fitted residual is

$$\hat{e}_i = y_i - \hat{m}(x_i).$$

As a general rule, but especially when the bandwidth  $h$  is small, it is hard to view  $\hat{e}_i$  as a good measure of the fit of the regression. For the NW and LL estimators, as  $h \rightarrow 0$  then  $\hat{m}(x_i) \rightarrow y_i$  and therefore  $\hat{e}_i \rightarrow 0$ .

---

<sup>2</sup>The constant  $(R_k/4)^{1/5}$  is bounded between 0.58 and 0.59.

This is clear overfitting as the true error  $e_i$  is not zero. In general, since  $\hat{m}(x_i)$  is a local average which includes  $y_i$ , the fitted value will be necessarily close to  $y_i$  and the residual  $\hat{e}_i$  small, and the degree of this overfitting increases as  $h$  decreases.

A standard solution is to measure the fit of the regression at  $x = x_i$  by re-estimating the model excluding the  $i^{th}$  observation. Let  $\tilde{m}_{-i}(x)$  be the leave-one-out nonparametric estimator computed without observation  $i$ . For example, for Nadaraya-Watson regression, this is

$$\tilde{y}_i = \tilde{m}_{-i}(x) = \frac{\sum_{j \neq i} k\left(\frac{x_j - x}{h}\right) y_j}{\sum_{j \neq i} k\left(\frac{x_j - x}{h}\right)}.$$

Notationally, the “ $-i$ ” subscript is used to indicate that the  $i^{th}$  observation is omitted.

The leave-one-out predicted value for  $y_i$  at  $x = x_i$  is

$$\tilde{y}_i = \tilde{m}_{-i}(x_i)$$

and the leave-one-out prediction error is

$$\tilde{e}_i = y_i - \tilde{y}_i. \quad (20.10)$$

Since  $\tilde{y}_i$  is not a function of  $y_i$ , there is no tendency for  $\tilde{y}_i$  to overfit for small  $h$ . Consequently,  $\tilde{e}_i$  is a good measure of the fit of the estimated nonparametric regression.

When possible the leave-one-out prediction errors should be used instead of the residuals  $\hat{e}_i$ .

## 20.12 Cross-Validation Bandwidth Selection

The most popular method in applied statistics to select bandwidths is cross-validation. The general idea is to estimate the model fit based on leave-one-out estimation. Here we describe the method as typically applied for regression estimation. The method applies to NW, LL and LP estimation, as well as other nonparametric estimators.

To be explicit about the dependence of the estimator on the bandwidth, let us write an estimator of  $m(x)$  with a given bandwidth  $h$  as  $\hat{m}(x, h)$ .

Ideally, we would like to select  $h$  to minimize the integrated mean-squared error (IMSE) of  $\hat{m}(x, h)$  as a estimate of  $m(x)$ :

$$\text{IMSE}_n(h) = \int_S \mathbb{E}((\hat{m}(x, h) - m(x))^2) f(x) w(x) dx$$

where  $f(x)$  is the marginal density of  $x_i$  and  $w(x)$  is an integrable weight function. The weight  $w(x)$  is the same as used in (20.5) and can be omitted when  $x_i$  has bounded support.

The difference  $\hat{m}(x, h) - m(x)$  at  $x = x_i$  can be estimated by the leave-one-out prediction errors (20.10)

$$\tilde{e}_i(h) = y_i - \tilde{m}_{-i}(x_i, h)$$

where we are being explicit about the dependence on the bandwidth  $h$ . A reasonable estimator of  $\text{IMSE}_n(h)$  is the weighted average mean squared prediction errors

$$\text{CV}(h) = \frac{1}{n} \sum_{i=1}^n \tilde{e}_i(h)^2 w(x_i). \quad (20.11)$$

This function of  $h$  is known as the **cross-validation criterion**. Once again, if  $x_i$  has bounded support then the weights  $w(x_i)$  can be omitted and this is typically done in practice.

It turns out that the cross-validation criterion is an unbiased estimator of the IMSE plus a constant for a sample with  $n - 1$  observations.

**Theorem 20.6**

$$\mathbb{E}(\text{CV}(h)) = \bar{\sigma}^2 + \text{IMSE}_{n-1}(h) \quad (20.12)$$

where  $\bar{\sigma}^2 = \mathbb{E}(e_i^2 w(x_i))$ .

The proof of Theorem 20.6 is presented in Section 19.18.

Since  $\bar{\sigma}^2$  is a constant independent of the bandwidth  $h$ ,  $\mathbb{E}(\text{CV}(h))$  is simply a shifted version of  $\text{IMSE}_{n-1}(h)$ . In particular, the  $h$  which minimizes  $\mathbb{E}(\text{CV}(h))$  and  $\text{IMSE}_{n-1}(h)$  are identical. When  $h$  is large the bandwidth which minimizes  $\text{IMSE}_{n-1}(h)$  and  $\text{IMSE}_n(h)$  are nearly identical, so  $\text{CV}(h)$  is essentially unbiased as an estimator of  $\text{IMSE}_n(h) + \bar{\sigma}^2$ . This considerations lead to the recommendation to select  $h$  as the value which minimizes  $\text{CV}(h)$ .

The cross-validation bandwidth  $\hat{h}$  is the value which minimizes  $\text{CV}(h)$

$$h_{\text{cv}} = \underset{h \geq h_\ell}{\operatorname{argmin}} \text{CV}(h) \quad (20.13)$$

for some  $h_\ell > 0$ . The restriction  $h \geq h_\ell$  can be imposed so that  $\text{CV}(h)$  is not evaluated over unreasonably small bandwidths.

There is not an explicit solution to the minimization problem (20.13), so it must be solved numerically. One method is grid search. Create a grid of values for  $h$ , e.g.  $[h_1, h_2, \dots, h_J]$ , evaluate  $\text{CV}(h_j)$  for  $j = 1, \dots, J$ , and set

$$h_{\text{cv}} = \underset{h \in [h_1, h_2, \dots, h_J]}{\operatorname{argmin}} \text{CV}(h).$$

Evaluation using a coarse grid is typically sufficient for practical application. Plots of  $\text{CV}(h)$  against  $h$  are a useful diagnostic tool to verify that the minimum of  $\text{CV}(h)$  has been obtained. Another method for obtaining the solution (20.13) is numerical optimization.

It is possible for the solution (20.13) to be unbounded, that is,  $\text{CV}(h)$  is decreasing for large  $h$  so that  $h_{\text{cv}} = \infty$ . This is okay. It simply means that the regression estimator simplifies to its full-sample version. For Nadaraya-Watson estimator this is  $\hat{m}_{\text{nw}}(x) = \bar{y}$ . For the local linear estimator this is  $\hat{m}_{\text{LL}}(x) = \hat{\alpha} + \hat{\beta}x$ .

For NW and LL estimation, the criterion (20.11) requires leave-one-out estimation of the conditional mean at each observation  $x_i$ . This is different from calculation of the estimator  $\hat{m}(x)$  as the latter is typically done at a set of fixed values of  $x$  for purposes of display.

To illustrate, Figure 20.5 displays the cross-validation criteria  $\text{CV}(h)$  for the Nadaraya-Watson and Local Linear estimators using the data from Figure 20.1, both using the Gaussian kernel. The CV functions are computed on a grid on  $[h_{\text{rot}}/3, 3h_{\text{rot}}]$  with 200 gridpoints. The CV-minimizing bandwidths are  $h_{\text{nw}} = 0.830$  for the Nadaraya-Watson estimator and  $h_{\text{LL}} = 0.764$  for the local linear estimator. These are slightly higher than the rule of thumb  $h_{\text{rot}} = 0.551$  value calculated earlier. Figure 20.5 shows the minimizing bandwidths by the arrows.

The CV criterion can also be used to select between different nonparametric estimators. The CV-selected estimator is the one with the lowest minimized CV criterion. For example, in Figure 20.5, you can see that the LL estimator has a minimized CV criterion of 0.0699 which is lower than the minimum 0.0703 obtained by the NW estimator. Since the LL estimator achieves a lower value of the CV criterion, LL is the CV-selected estimator. The difference, however, is small, indicating that the two estimators achieve similar IMSE.

Figure 20.6 displays the local linear estimates  $\hat{m}(x)$  using the ROT and CV bandwidths along with the true conditional mean  $m(x)$ . The estimators track the true function quite well, and the difference between the bandwidths is relatively minor in this application.

## 20.13 Asymptotic Distribution

We first provide a consistency result for the Nadaraya-Watson estimator.

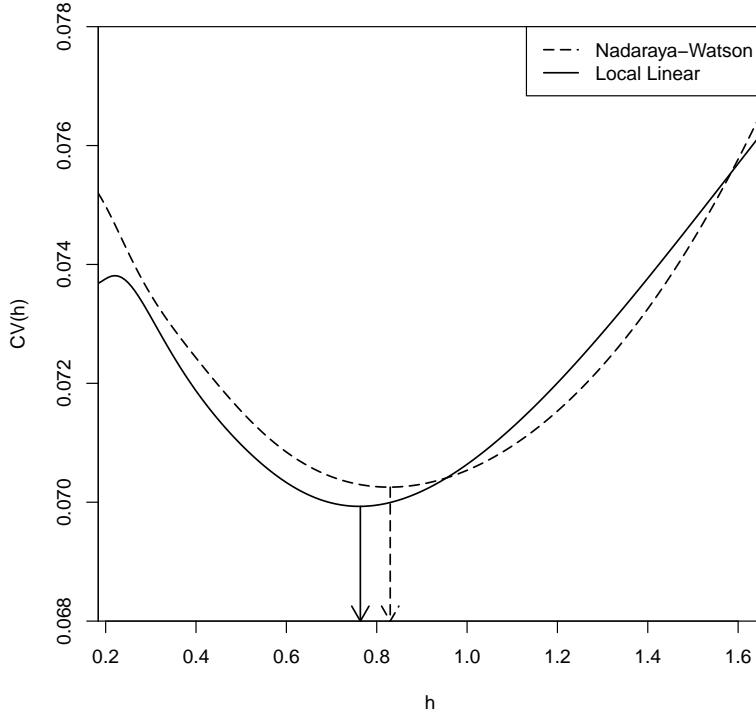


Figure 20.5: Cross-Validation Criteria, Nadaraya-Watson Regression and Local Linear Regression

**Theorem 20.7** Under Assumption 20.1, then  $\hat{m}_{\text{nw}}(x) \xrightarrow{p} m(x)$  and  $\hat{m}_{\text{LL}}(x) \xrightarrow{p} m(x)$ .

A proof for the Nadaraya-Watson estimator is presented in Section 20.25. For the local linear estimator see Fan and Gijbels (1996).

Theorem 20.7 shows that the estimators are consistent for  $m(x)$  under very mild continuity assumptions. In particular, no smoothness conditions on  $m(x)$  are required beyond continuity.

We next present an asymptotic distribution result. The following shows that the kernel regression estimators are asymptotically normal with a non-parametric rate of convergence, a non-trivial asymptotic bias, and a non-degenerate asymptotic variance.

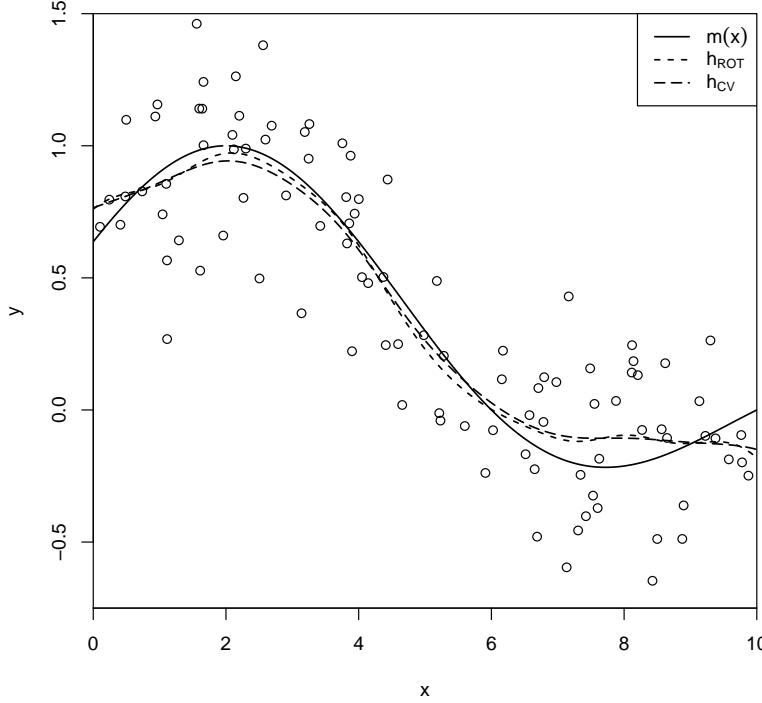


Figure 20.6: Nonparametric Estimates using data-dependent (CV) bandwidths

**Theorem 20.8** Suppose Assumption 20.1 holds. Assume in addition that  $m''(x)$  and  $f'(x)$  are continuous in  $\mathcal{N}$ , that for some  $r > 2$  and  $x \in \mathcal{N}$ ,

$$\mathbb{E}(|e_i|^r | x_i = x) \leq \bar{\sigma} < \infty, \quad (20.14)$$

and

$$nh^5 = O(1). \quad (20.15)$$

Then

$$\sqrt{nh}(\hat{m}_{\text{nw}}(x) - m(x) - h^2 B_{\text{nw}}(x)) \xrightarrow{d} N\left(0, \frac{R_k \sigma^2(x)}{f(x)}\right). \quad (20.16)$$

Similarly,

$$\sqrt{nh}(\hat{m}_{\text{LL}}(x) - m(x) - h^2 B_{\text{LL}}(x)) \xrightarrow{d} N\left(0, \frac{R_k \sigma^2(x)}{f(x)}\right).$$

A proof for the Nadaraya-Watson estimator appears in Section 19.18. For the local linear estimator, see Fan and Gijbels (1996).

Relative to Theorem 20.7, Theorem 20.8 requires stronger smoothness conditions on the conditional mean and marginal density. There are also two technical regularity conditions. The first is a conditional moment bound (20.14) (which is used to verify the Lindeberg condition for the CLT) and the second is the bandwidth bound  $nh^5 = O(1)$ . The latter means that the bandwidth must decline to zero at least at the rate  $n^{-1/5}$ , and is used<sup>3</sup> to ensure that higher-order bias terms do not enter the asymptotic distribution

<sup>3</sup>This could be weakened if stronger smoothness conditions are assumed. For example, if  $m^{(4)}(x)$  and  $f^{(3)}(x)$  are continuous

(20.16).

There are several interesting features about the asymptotic distribution which are noticeably different than for parametric estimators. First, the estimators converge at the rate  $\sqrt{nh}$  not  $\sqrt{n}$ . Since  $h \rightarrow 0$ ,  $\sqrt{nh}$  diverges slower than  $\sqrt{n}$ , thus the nonparametric estimators converge more slowly than a parametric estimator. Second, the asymptotic distribution contains a non-negligible bias term  $h^2 B(x)$ . Third, the distribution (20.16) is identical in form to that for the kernel density estimator (Theorem 19.7).

The fact that the estimators converge at the rate  $\sqrt{nh}$  has led to the interpretation of  $nh$  as the “effective sample size”. This is because the number of observations being used to construct  $\hat{m}(x)$  is proportional to  $nh$ , not  $n$  as for a parametric estimator.

It is helpful to understand that the nonparametric estimator has a reduced convergence rate relative to parametric asymptotic theory because the object being estimated –  $m(x)$  – is nonparametric. This is harder than estimating a finite dimensional parameter, and thus comes at a cost.

Unlike parametric estimation, the asymptotic distribution of the nonparametric estimator includes a term representing the bias of the estimator. The asymptotic distribution (20.16) shows the form of this bias. It is proportional to the squared bandwidth  $h^2$  (the degree of smoothing) and to the function  $B_{\text{nw}}(x)$  or  $B_{\text{LL}}(x)$  which depends on the slope and curvature of the CEF  $m(x)$ . Interestingly, when  $m(x)$  is constant then  $B_{\text{nw}}(x) = B_{\text{LL}}(x) = 0$  and the kernel estimator has no asymptotic bias. The bias is essentially increasing in the curvature of the CEF function  $m(x)$ . This is because the local averaging smooths  $m(x)$ , and the smoothing induces more bias when  $m(x)$  is curved.

The asymptotic variance of  $\hat{m}(x)$  is inversely proportional to the marginal density  $f(x)$ . This means that  $\hat{m}(x)$  has relatively low precision for regions where  $x_i$  has a low density. This makes sense since these are regions where there are relatively few observations. An implication is that the nonparametric estimator  $\hat{m}(x)$  will be relatively inaccurate in the tails of the distribution of  $x_i$ .

## 20.14 Undersmoothing

In Section 19.16 we showed that the bias term in the asymptotic distribution of the kernel density estimator can be eliminated if the bandwidth is selected to converge to zero faster than the optimal rate  $n^{-1/5}$ , thus  $h = o(n^{-1/5})$ . The same holds for kernel regression. This is called an **undersmoothing** bandwidth.

Similarly, with a smaller bandwidth the regression estimator has reduced bias and increased variance. Consequently the bias is asymptotically negligible if the bandwidth converges to zero faster than  $n^{-1/5}$ .

**Theorem 20.9** Under the conditions of Theorem 20.8, and in addition  $nh^5 = o(1)$ ,

$$\begin{aligned}\sqrt{nh}(\hat{m}_{\text{nw}}(x) - m(x)) &\xrightarrow{d} N\left(0, \frac{R_k \sigma^2(x)}{f(x)}\right) \\ \sqrt{nh}(\hat{m}_{\text{LL}}(x) - m(x)) &\xrightarrow{d} N\left(0, \frac{R_k \sigma^2(x)}{f(x)}\right).\end{aligned}$$

This result has the same advantages and disadvantages as discussed in Section 19.16. In particular, undersmoothing results in a less efficient estimator. While a smaller bandwidth reduces bias it does not actually eliminate the bias in any actual application. In this sense the undersmoothing distribution theory in Theorem 20.9 is misleading.

---

then (20.15) can be weakened to  $nh^9 = O(1)$ , which means that the bandwidth must decline to zero at least at the rate  $n^{-1/9}$ .

## 20.15 Conditional Variance Estimation

The conditional variance is

$$\sigma^2(x) = \text{var}(y_i | x_i = x) = \mathbb{E}(e_i^2 | x_i = x).$$

There are a number of contexts where it is desirable to estimate  $\sigma^2(x)$  including prediction intervals and confidence intervals for the estimated mean function. In general the conditional variance function is nonparametric as economic models rarely specify the form of  $\sigma^2(x)$ . Thus estimation of  $\sigma^2(x)$  is typically done nonparametrically.

Since  $\sigma^2(x)$  is the CEF of  $e_i^2$  given  $x_i$ , it can be estimated by a nonparametric regression of  $e_i^2$  on  $x_i$ . For example, the ideal NW estimator (if  $e_i$  were observed) is

$$\bar{\sigma}^2(x) = \frac{\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right) e_i^2}{\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)}.$$

Since the errors  $e_i$  are not observed, we need to replace them with an estimator. A simple choice are the residuals  $\hat{e}_i = y_i - \hat{m}(x_i)$ . A better choice are the leave-one-out prediction errors  $\tilde{e}_i = y_i - \hat{m}_{-i}(x_i)$ . The latter are recommended for variance estimation as they are not subject to overfitting. With this substitution the NW estimator of the conditional variance is

$$\hat{\sigma}^2(x) = \frac{\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right) \tilde{e}_i^2}{\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)}. \quad (20.17)$$

This estimator depends on a bandwidth  $h$ , but there is no reason for this bandwidth to be the same as that used to estimate the conditional mean. The ROT or cross-validation using  $\tilde{e}_i^2$  as the dependent variable can be used to select the bandwidth for estimation of  $\hat{\sigma}^2(x)$  separately from cross-validation for estimation of  $\hat{m}(x)$ .

There is one subtle difference between CEF and conditional variance estimation. The conditional variance is inherently non-negative  $\sigma^2(x) \geq 0$  and it is desirable for the estimator to satisfy this property. Interestingly, the NW estimator (20.17) is necessarily non-negative, since it is a smoothed average of the non-negative squared residuals, but the LL estimator is not guaranteed to be non-negative for all  $x$ . Furthermore, the NW estimator has as a special case the homoskedastic estimator  $\hat{\sigma}^2(x) = \hat{\sigma}^2$  (full sample variance) which may be a relevant selection. For these reasons, the NW estimator may be preferred for conditional variance estimation.

Fan and Yao (1998) derive the asymptotic distribution of the estimator (20.17). They obtain the surprising result that the asymptotic distribution of the two-step estimator  $\hat{\sigma}^2(x)$  is identical to that of the one-step idealized estimator  $\bar{\sigma}^2(x)$ .

## 20.16 Variance Estimation and Standard Errors

It is relatively straightforward to calculate the exact conditional variance of the Nadaraya-Watson, local linear, or local polynomial estimator. They can be written as

$$\begin{aligned} \hat{\beta}(x) &= (\mathbf{Z}' \mathbf{K} \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{K} \mathbf{y}) \\ &= (\mathbf{Z}' \mathbf{K} \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{K} \mathbf{m}) + (\mathbf{Z}' \mathbf{K} \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{K} \mathbf{e}) \end{aligned}$$

where  $\mathbf{m}$  is the  $n \times 1$  vector of means  $m(x_i)$ . The first component is a function only of the regressors and the second is linear in the error  $\mathbf{e}$ . Thus conditionally on the regressors  $\mathbf{X}$ ,

$$V_{\hat{\beta}}(x) = \text{var}(\hat{\beta} | \mathbf{X}) = (\mathbf{Z}' \mathbf{K} \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{K} \mathbf{D} \mathbf{K} \mathbf{Z}) (\mathbf{Z}' \mathbf{K} \mathbf{Z})^{-1}$$

where  $\mathbf{D} = \text{diag}(\sigma^2(x_1), \dots, \sigma^2(x_n))$ .

A White-type estimator can be formed by replacing  $\sigma^2(x_i)$  with the squared residuals  $\tilde{e}_i^2$  or prediction errors  $\hat{e}_i^2$

$$\hat{V}_{\hat{\beta}}(x) = (\mathbf{Z}' \mathbf{K} \mathbf{Z})^{-1} \left( \sum_{i=1}^n k \left( \frac{x_i - x}{h} \right)^2 \mathbf{z}_i(x) \mathbf{z}_i(x)' \tilde{e}_i^2 \right) (\mathbf{Z}' \mathbf{K} \mathbf{Z})^{-1}.$$

A second estimator is obtained by replacing  $\sigma^2(x_i)$  with an estimator such as (20.17)

$$\hat{V}_{\hat{\beta}}(x) = (\mathbf{Z}' \mathbf{K} \mathbf{Z})^{-1} \left( \sum_{i=1}^n k \left( \frac{x_i - x}{h} \right)^2 \mathbf{z}_i(x) \mathbf{z}_i(x)' \hat{\sigma}^2(x_i) \right) (\mathbf{Z}' \mathbf{K} \mathbf{Z})^{-1}.$$

A third replaces  $\hat{\sigma}^2(x_i)$  with the estimator  $\hat{\sigma}^2(x)$

$$\begin{aligned} \hat{V}_{\hat{\beta}}(x) &= (\mathbf{Z}' \mathbf{K} \mathbf{Z})^{-1} \left( \sum_{i=1}^n k \left( \frac{x_i - x}{h} \right)^2 \mathbf{z}_i(x) \mathbf{z}_i(x)' \right) (\mathbf{Z}' \mathbf{K} \mathbf{Z})^{-1} \hat{\sigma}^2(x) \\ &= (\mathbf{Z}' \mathbf{K} \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{K}^2 \mathbf{Z}) (\mathbf{Z}' \mathbf{K} \mathbf{Z})^{-1} \hat{\sigma}^2(x). \end{aligned} \quad (20.18)$$

A fourth uses the asymptotic formula

$$\hat{V}_{\hat{m}(x)} = \frac{R_k \hat{\sigma}^2(x)}{nh \hat{f}(x)}$$

with  $\hat{\sigma}^2(x)$  from (20.17) and  $\hat{f}(x)$  from (19.2).

For local linear and local polynomial estimators the estimator  $\hat{V}_{\hat{m}(x)}$  is the first diagonal element of the matrix  $\hat{V}_{\hat{\beta}}(x)$ . For any of the variance estimators a standard error for  $\hat{m}(x)$  is the square root of  $\hat{V}_{\hat{m}(x)}$ .

## 20.17 Confidence Bands

For either density or conditional mean estimation we can construct asymptotic confidence intervals. For the density function  $f(x)$  an asymptotic 95% confidence interval is

$$\hat{f}(x) \pm 1.96 \sqrt{\hat{V}_{\hat{f}}(x)}. \quad (20.19)$$

For the mean function  $m(x)$  an asymptotic 95% confidence interval is

$$\hat{m}(x) \pm 1.96 \sqrt{\hat{V}_{\hat{m}(x)}}. \quad (20.20)$$

These confidence intervals can be plotted along with  $\hat{f}(x)$  or  $\hat{m}(x)$  to assess precision.

It should be noted, however, that these confidence intervals have two unusual properties. First, they are pointwise in  $x$ , meaning that they are designed to have coverage probability at each  $x$ , not uniformly across  $x$ . Thus they are typically called **pointwise confidence intervals**.

Second, because they do not account for the bias, they are not asymptotically valid confidence intervals for  $f(x)$  or  $m(x)$ . Rather, they are asymptotically valid confidence intervals for the pseudo-true (smoothed) value, e.g.  $f(x) + \frac{1}{2} f''(x)h^2$ . One way of thinking about this is that the confidence intervals account for the variance of the estimator but not its bias. A technical trick which solves this problem is to assume an undersmoothing bandwidth. In this case the above confidence intervals are technically asymptotically valid. This is only a technical trick as it does not really eliminate the bias only assumes it away. The plain fact is that once we honestly acknowledge that the true CEF is nonparametric, it then follows that any finite sample estimate will have finite sample bias, and this bias will be inherently unknown and thus difficult to incorporate into confidence intervals.

Despite these unusual properties we can still use the intervals (20.19) and (20.20) to display uncertainty and as a check on the precision of the estimates.

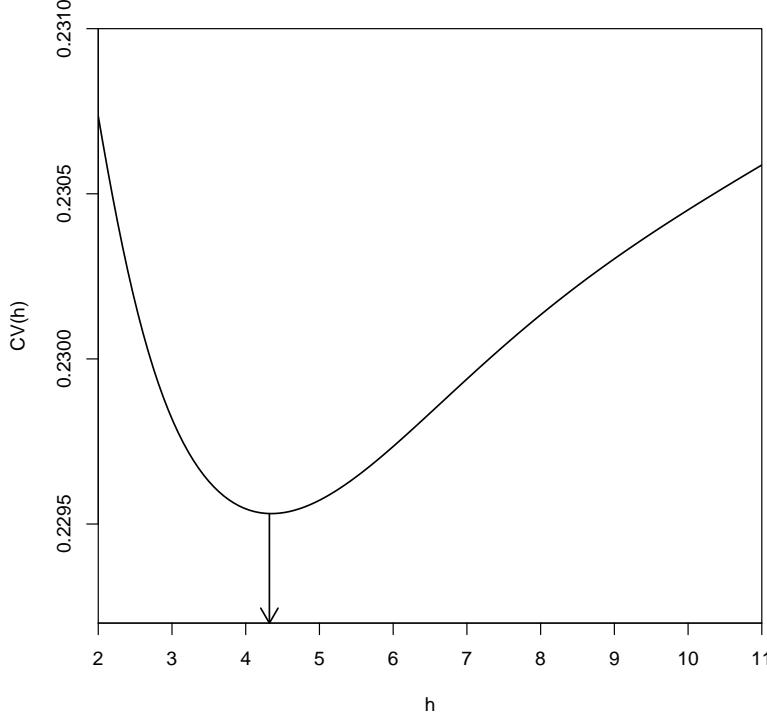


Figure 20.7: Cross-Validation Criteria for Wage Regression

## 20.18 The Local Nature of Kernel Regression

The kernel regression estimators (Nadaraya-Watson, Local Linear, and Local Polynomial) are all essentially local estimators in that given  $h$  the estimator  $\hat{m}(x)$  is a function only of the sub-sample for which  $x_i$  is close to  $x$ . The other observations do not directly affect the estimator. This is reflected in the distribution theory as well. Theorem 20.7 shows that  $\hat{m}(x)$  is consistent for  $m(x)$  if the latter is continuous at  $x$ . Theorem 20.8 shows that the asymptotic distribution of  $\hat{m}(x)$  depends only on the functions  $m(x)$ ,  $f(x)$  and  $\sigma^2(x)$  at the point  $x$ . The distribution does not depend on the global behavior of  $m(x)$ .

Global features do affect the estimator  $\hat{m}(x)$ , however, through the bandwidth  $h$ . The bandwidth selection methods described here are global in nature as they attempt to minimize AIMSE. Local bandwidths (designed to minimize the AMSE at a single point  $x$ ) can alternatively be employed, but these are less commonly used, in part because such bandwidth estimators have high imprecision. Picking local bandwidths adds extra noise.

Furthermore, selected bandwidths may be meaningfully large, so that the estimation window may be a large portion of the sample. In this case estimation is neither local nor fully global.

## 20.19 Application to Wage Regression

We illustrate the methods with an application to the CPS data set. We are interested in the nonparametric regression of  $\log(wage)$  on  $experience$ . To illustrate we take the subsample of black men with 12 years of education (high school graduates). This sample has 762 observations.

We first need to decide on the region of interest (range of experience) for which we will calculate the regression estimator. We select the range  $[0, 40]$  since most observations (90%) have experience levels below 40 years.

To avoid boundary bias, we use the local linear estimator.

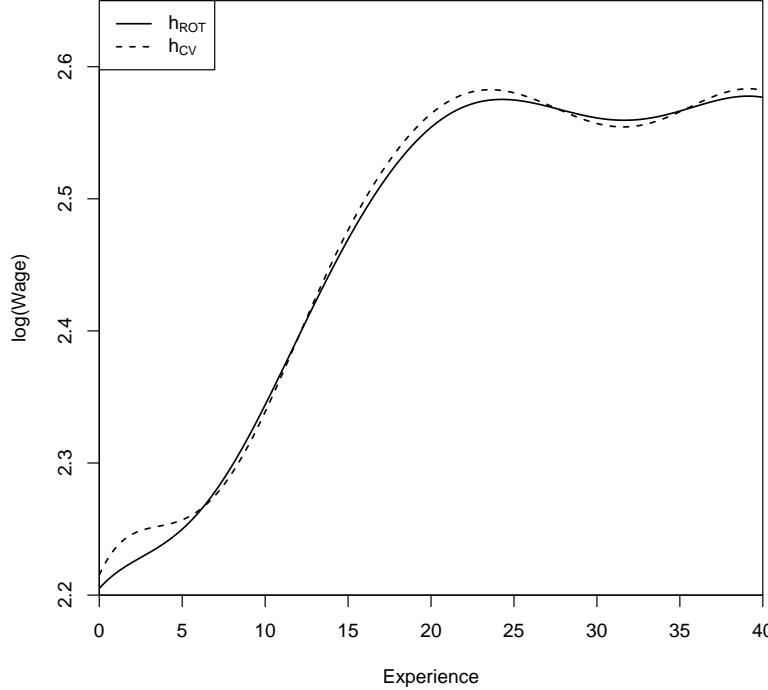


Figure 20.8: Local Linear Regressions of  $\log(\text{wage})$  on experience

We next calculate the Fan-Ghybels rule-of-thumb bandwidth (20.9) and find  $h_{\text{rot}} = 5.14$ . We then calculate the cross-validation criterion, using the rule-of-thumb as a baseline. The CV criterion is displayed in Figure 20.7. The minimizer is  $h_{\text{cv}} = 4.32$  which is somewhat smaller than the ROT bandwidth.

We calculate the local linear estimator using both bandwidths and display the estimates in Figure 20.8. The regression functions are increasing for experience levels up to 20 years, and then become flat. While the functions are roughly concave, they are noticeably different than a traditional quadratic specification. Comparing the estimates, the smaller CV-selected bandwidth produces a regression estimate which is a bit too wavy, while the ROT bandwidth produces a regression estimate which is much smoother, yet captures the same essential features. Based on this inspection we select the estimate based on the ROT bandwidth (the solid line in Figure 20.8).

Our next step is to calculate the conditional variance function. We calculate the ROT bandwidth for a regression using the squared leave-one-out residuals (prediction errors), and find  $h_{\text{rot}} = 6.77$  which is larger than the bandwidth used for conditional mean estimation. We next calculate the cross-validation functions for conditional variance estimation (regression of squared prediction errors on *experience*) using both NW and LL regression. The CV functions are displayed in Figure 20.9. The CV plots are quite interesting. For the LL estimator the CV function has a local minimum around  $h = 5$  but the global minimizer is unbounded. The CV function for the NW estimator is globally decreasing with an unbounded minimizer. The NW also achieves a considerably lower CV value than the LL estimator. This means that the CV-selected variance estimator is the NW estimator with  $h = \infty$ , which is the simple full-sample estimator  $\hat{\sigma}^2$  calculated with the prediction errors.

We next compute standard errors for the regression function estimates, using formula (20.18) with the estimator  $\hat{\sigma}^2$  just described. In Figure 20.10 we display the estimated regression (the same as Figure 20.8 using the ROT bandwidth), along with 95% asymptotic confidence bands computed as in (20.20). By displaying the confidence bands we can see that there is considerable imprecision in the estimator for low experience levels. We can still see that the estimates and confidence bands show that the experience profile is increasing up to about 20 years of experience, and then flattens above 20 years. The estimates

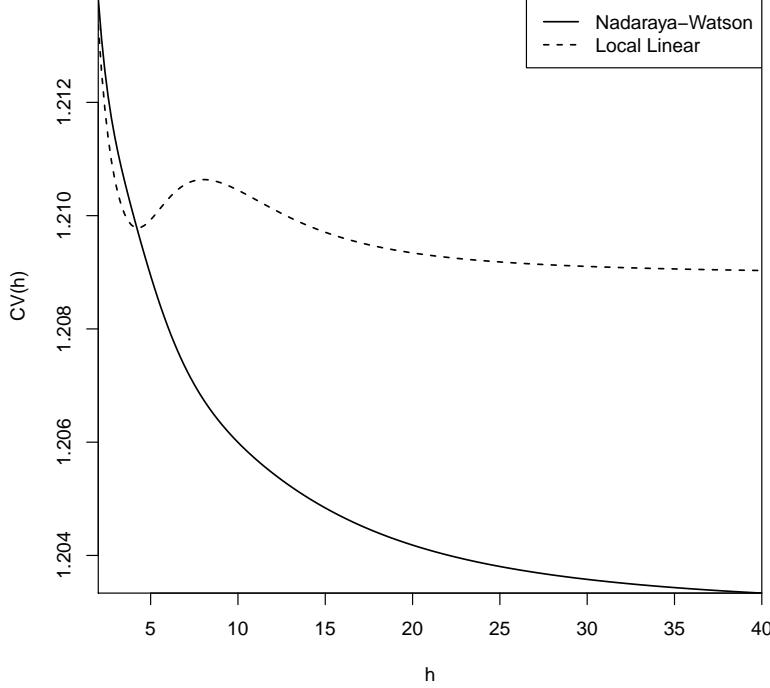


Figure 20.9: Cross-Validation Functions for Conditional Variance Estimators

imply that for this population (black men who are high school graduates) the average wage rises for the first 20 years of work experience (from 18 to 38 years of age) and then flattens, with no further increases in average wages for the next 20 years of work experience (from 38 to 58 years of age).

## 20.20 Clustered Observations

Clustered observations take the form  $(y_{ig}, x_{ig})$  for individuals  $i = 1, \dots, n_g$  in cluster  $g = 1, \dots, G$ . The model is

$$y_{ig} = m(x_{ig}) + e_{ig}$$

$$\mathbb{E}(e_{ig} | \mathbf{X}_g) = 0.$$

where  $\mathbf{X}_g$  is the stacked  $x_{ig}$ . The assumption is that the clusters are mutually independent. Dependence within each cluster is unstructured.

Write

$$\mathbf{z}_{ig}(x) = \begin{pmatrix} 1 \\ x_{ig} - x \end{pmatrix}.$$

Stack  $y_{ig}$ ,  $e_{ig}$  and  $\mathbf{z}_{ig}(x)$  into cluster-level variables  $\mathbf{y}_g$ ,  $\mathbf{e}_g$  and  $\mathbf{Z}_g(x)$ . Let  $\mathbf{K}_g(x) = \text{diag}\left\{k\left(\frac{x_{ig} - x}{h}\right)\right\}$ . The local linear estimator can be written as

$$\begin{aligned} \hat{\boldsymbol{\beta}}(x) &= \left( \sum_{g=1}^G \sum_{i=1}^{n_g} k\left(\frac{x_{ig} - x}{h}\right) \mathbf{z}_{ig}(x) \mathbf{z}_{ig}(x)' \right)^{-1} \left( \sum_{g=1}^G \sum_{i=1}^{n_g} k\left(\frac{x_{ig} - x}{h}\right) \mathbf{z}_{ig}(x) y_{ig} \right) \\ &= \left( \sum_{g=1}^G \mathbf{Z}_g(x)' \mathbf{K}_g(x) \mathbf{Z}_g(x) \right)^{-1} \left( \sum_{g=1}^G \mathbf{Z}_g(x)' \mathbf{K}_g(x) \mathbf{y}_g \right). \end{aligned} \quad (20.21)$$

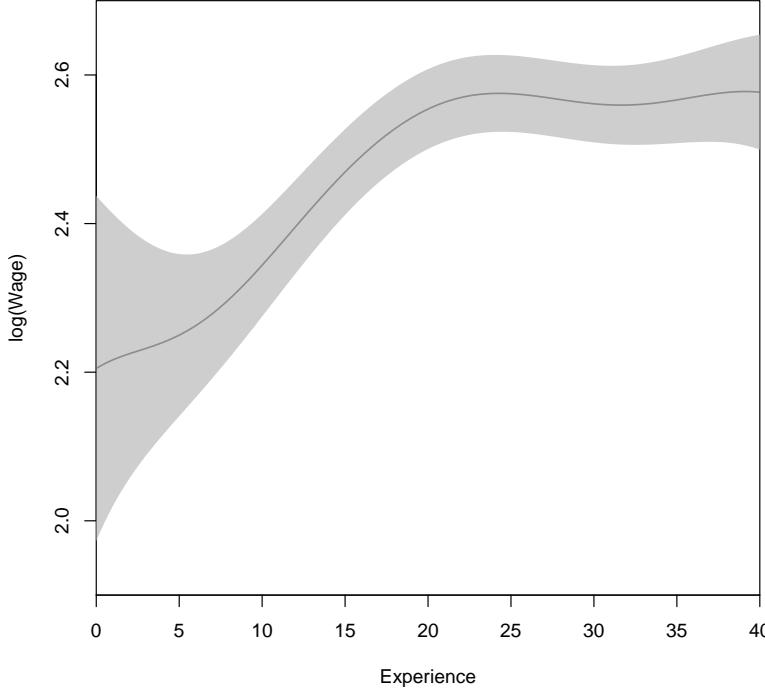


Figure 20.10: Regression  $\log(\text{wage})$  on experience, with 95% Pointwise Confidence Bands

The local linear estimator  $\hat{m}(x) = \hat{\beta}_1(x)$  is the intercept in (20.21).

The natural method to obtain prediction errors is by delete-cluster regression. The delete-cluster estimator of  $\beta$  is

$$\tilde{\beta}_{(-g)}(x) = \left( \sum_{j \neq g} \mathbf{Z}_j(x)' \mathbf{K}_j(x) \mathbf{Z}_j(x) \right)^{-1} \left( \sum_{j \neq g} \mathbf{Z}_j(x)' \mathbf{K}_j(x) \mathbf{y}_j \right). \quad (20.22)$$

The delete-cluster estimator of  $m(x)$  is the intercept  $\tilde{m}_1(x) = \tilde{\beta}_{1(-g)}(x)$  from (20.22). The delete-cluster prediction error for observation  $ig$  is

$$\tilde{e}_{ig} = y_{ig} - \tilde{\beta}_{1(-g)}(x_{ig}). \quad (20.23)$$

Let  $\tilde{\mathbf{e}}_g$  be the stacked  $\tilde{e}_{ig}$  for cluster  $g$ .

The variance of (20.21), conditional on the regressors  $\mathbf{X}$ , is

$$\mathbf{V}_{\hat{\beta}}(x) = \left( \sum_{g=1}^G \mathbf{Z}_g(x)' \mathbf{K}_g(x) \mathbf{Z}_g(x) \right)^{-1} \left( \sum_{g=1}^G \mathbf{Z}_g(x)' \mathbf{K}_g(x) \mathbf{S}_g(x) \mathbf{K}_g(x) \mathbf{Z}_g(x) \right) \left( \sum_{g=1}^G \mathbf{Z}_g(x)' \mathbf{K}_g(x) \mathbf{Z}_g(x) \right)^{-1} \quad (20.24)$$

where

$$\mathbf{S}_g = \mathbb{E}(\mathbf{e}_g \mathbf{e}_g' | \mathbf{X}_g).$$

The covariance matrix (20.24) can be estimated by replacing  $\mathbf{S}_g$  with an estimator of  $\mathbf{e}_g \mathbf{e}_g'$ . Based on analogy with regression estimation we suggest the delete-cluster prediction errors  $\tilde{\mathbf{e}}_g$  as they are not subject to over-fitting. This covariance matrix estimator using this choice is

$$\widehat{\mathbf{V}}_{\hat{\beta}}(x) = \left( \sum_{g=1}^G \mathbf{Z}_g(x)' \mathbf{K}_g(x) \mathbf{Z}_g(x) \right)^{-1} \left( \sum_{g=1}^G \mathbf{Z}_g(x) \mathbf{K}_g(x) \tilde{\mathbf{e}}_g \tilde{\mathbf{e}}_g' \mathbf{K}_g(x) \mathbf{Z}_g(x) \right) \left( \sum_{g=1}^G \mathbf{Z}_g(x)' \mathbf{K}_g(x) \mathbf{Z}_g(x) \right)^{-1}. \quad (20.25)$$

The standard error for  $\hat{m}(x)$  is the square root of the first diagonal element of  $\widehat{V}_{\hat{\beta}}(x)$ .

There is no current theory on how to select the bandwidth  $h$  for nonparametric regression using clustered observations. The Fan-Ghybels ROT bandwidth  $h_{\text{rot}}$  is designed for independent observations, so is likely to be a crude choice in the case of clustered observations. Standard cross-validation has similar limitations. A practical alternative is to select the bandwidth  $h$  to minimize a delete-cluster cross-validation criterion. While there is no formal theory to justify this choice, it seems like a reasonable option. The delete-cluster CV criterion is

$$\text{CV}(h) = \frac{1}{n} \sum_{g=1}^G \sum_{i=1}^{n_g} \tilde{e}_{ig}^2$$

where  $\tilde{e}_{ig}$  are the delete-cluster prediction errors (20.23). The delete-cluster CV bandwidth is the value which minimizes this function:

$$h_{\text{cv}} = \underset{h \geq h_\ell}{\operatorname{argmin}} \text{CV}(h).$$

As for the case of conventional cross-validation, it may be valuable to plot  $\text{CV}(h)$  against  $h$  to verify that the minimum has been obtained and to assess sensitivity.

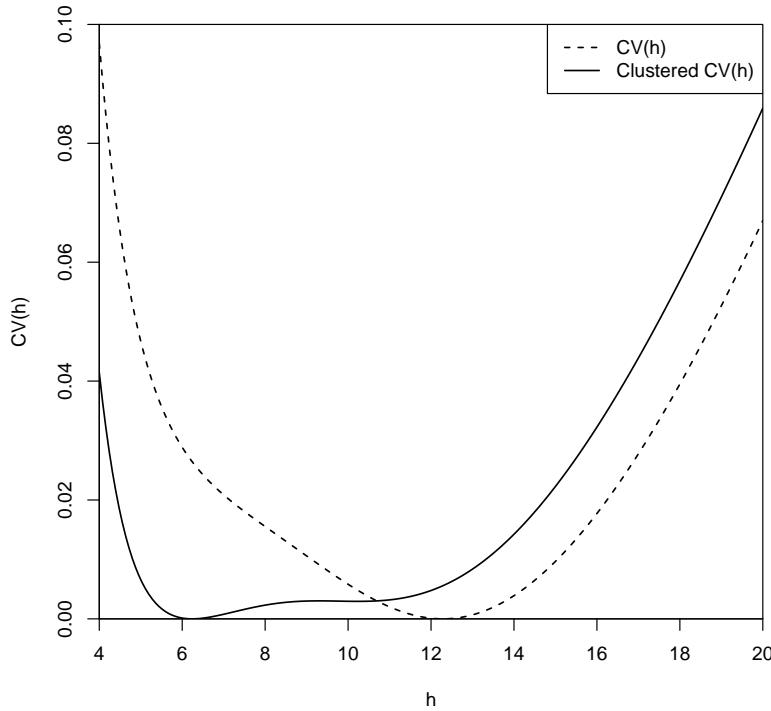


Figure 20.11: Cross-Validation Functions

## 20.21 Application to Testscores

We illustrate kernel regression with clustered observations by using the Duflo, Dupas and Kremer (2011) investigation of the effect of student tracking on testscores. Recall that the core question was effect of *testscore* on the dummy variable *tracking*. A set of controls were included, including a continuous variable *percentile* which recorded the student's initial test score (as a percentile) used for classroom assignment. We investigate the authors' specification of this control using local linear regression.

We took the subsample of 1487 girls who experienced tracking, and estimated the regression of *testscores* on *percentile*. For this application we used unstandardized<sup>4</sup> test scores which range from 0 to about 40. We used local linear regression with a Gaussian kernel.

First consider bandwidth selection. The Fan-Ghybs ROT and conventional cross-validation bandwidths are  $h_{\text{rot}} = 6.7$  and  $h_{\text{cv}} = 12.3$ . We then calculated the clustered cross-validation criterion, which has minimizer  $h_{\text{cv}} = 6.2$ . To understand the differences, we plot the standard and clustered cross-validation functions in Figure 20.11. In order to plot on the same graph we normalize each by subtracting their minimized value (so each is minimized at zero). What we can see from Figure 20.11 is that while the conventional CV criterion is sharply minimized at  $h = 12.3$ , the clustered CV criterion is essentially flat between 5 and 11. This means that the clustered CV criterion has difficulty discriminating between these bandwidth choices

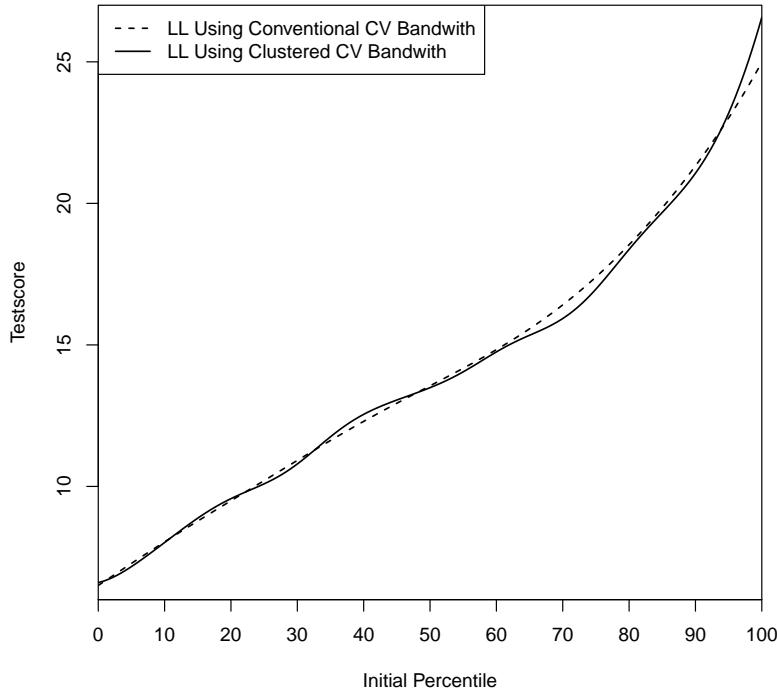


Figure 20.12: TestScore as a Function of Initial Percentile

To compare the estimated regression functions, in Figure 20.12 we plot the estimated regression functions which use the bandwidths selected by conventional and clustered cross-validation. Inspecting the plots, the estimator using the conventional CV bandwidth is smoother than the estimator using the smaller clustered CV bandwidth. The most noticeable difference arises at the right end of the plot, which shows the expected test score for the students who had the very best preliminary test scores. The estimator using the clustered CV bandwidth shows a meaningful upturn for students with initial testscore percentile above 90%. Based on this evidence we select the local linear estimator  $\hat{m}_{\text{LL}}(x)$  using the clustered cross-validation bandwidth  $h_{\text{cv}} = 6.2$ .

Using this bandwidth we estimate the delete-cluster prediction errors  $\tilde{\epsilon}_g$  and use these to calculate the standard errors for the local linear estimator  $\hat{m}_{\text{LL}}(x)$  using formula (20.25). These standard errors are roughly twice as large as those calculated using the non-clustered formula. We use the standard errors to calculate 95% asymptotic pointwise confidence bands as in (20.20).

<sup>4</sup>In Section 4.21, following Duflo, Dupas and Kremer (2011) the dependent variable was standardized testscores (normalized to have mean zero and variance one).

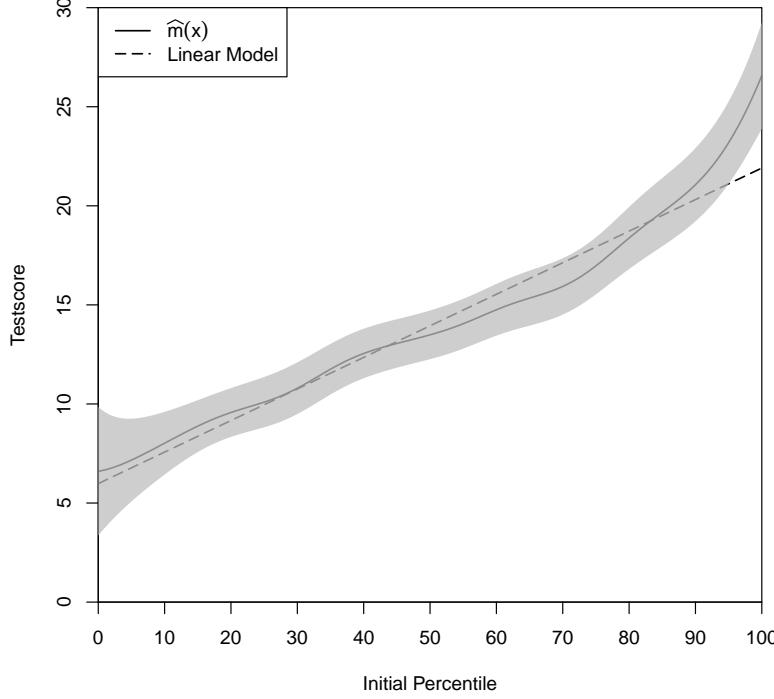


Figure 20.13: TestScore as a Function of Initial Percentile with Confidence Bands

Figure 20.13 shows our estimated regression function and pointwise 95% confidence bands. Also plotted for comparison is an estimated linear regression line. The local linear estimator is very similar to the global linear regression estimator for initial percentiles below 80%. But for initial percentiles above 80% the two lines diverge. The confidence bands suggest that these differences are statistically meaningful. Students with initial testscores at the top of the initial distribution have higher final testscores on average than predicted by a linear specification.

## 20.22 Multiple Regressors

Our analysis has focus on the case of real-valued  $x_i$  for simplicity of exposition, but the methods of kernel regression extend easily to the multiple regressor case, at the cost of a reduced rate of convergence. In this section we consider the case of estimation of the conditional expectation function

$$\mathbb{E}(y_i | \mathbf{x}_i = \mathbf{x}) = m(\mathbf{x})$$

when

$$\mathbf{x}_i = \begin{pmatrix} x_{1i} \\ \vdots \\ x_{di} \end{pmatrix}$$

is a  $d$ -vector.

For any evaluation point  $\mathbf{x}$  and observation  $i$ , define the kernel weights

$$k_i(\mathbf{x}) = k\left(\frac{x_{1i} - x_1}{h_1}\right)k\left(\frac{x_{2i} - x_2}{h_2}\right)\cdots k\left(\frac{x_{di} - x_d}{h_d}\right),$$

a  $d$ -fold product kernel. The kernel weights  $k_i(\mathbf{x})$  assess if the regressor vector  $\mathbf{x}_i$  is close to the evaluation point  $\mathbf{x}$  in the Euclidean space  $\mathbb{R}^d$ .

These weights depend on a set of  $d$  bandwidths,  $h_j$ , one for each regressor. We can group them together into a single vector for notational convenience:

$$\mathbf{h} = \begin{pmatrix} h_1 \\ \vdots \\ h_d \end{pmatrix}.$$

Given these weights, the Nadaraya-Watson estimator takes the form

$$\hat{m}(\mathbf{x}) = \frac{\sum_{i=1}^n k_i(\mathbf{x})y_i}{\sum_{i=1}^n k_i(\mathbf{x})}.$$

For the local-linear estimator, define

$$\mathbf{z}_i(\mathbf{x}) = \begin{pmatrix} 1 \\ \mathbf{x}_i - \mathbf{x} \end{pmatrix}$$

and then the local-linear estimator can be written as  $\hat{m}(\mathbf{x}) = \hat{\alpha}(\mathbf{x})$  where

$$\begin{pmatrix} \hat{\alpha}(\mathbf{x}) \\ \hat{\beta}(\mathbf{x}) \end{pmatrix} = \left( \sum_{i=1}^n k_i(\mathbf{x}) \mathbf{z}_i(\mathbf{x}) \mathbf{z}_i(\mathbf{x})' \right)^{-1} \sum_{i=1}^n k_i(\mathbf{x}) \mathbf{z}_i(\mathbf{x}) y_i \\ = (\mathbf{Z}' \mathbf{K} \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{K} \mathbf{y}$$

where  $\mathbf{K} = \text{diag}\{k_1(x), \dots, k_n(x)\}$ .

In multiple regressor kernel regression, cross-validation remains a recommended method for bandwidth selection. The leave-one-out residuals  $\tilde{e}_i$  and cross-validation criterion  $\text{CV}(\mathbf{h})$  are defined identically as in the single regressor case. The only difference is that now the CV criterion is a function over the  $d$ -dimensional bandwidth  $\mathbf{h}$ . This means that numerical minimization needs to be done more efficiently than by a simple grid search.

The asymptotic distribution of the estimators in the multiple regressor case is an extension of the single regressor case. Let  $f(\mathbf{x})$  denote the marginal density of  $\mathbf{x}_i$ ,  $\sigma^2(\mathbf{x}) = \mathbb{E}(e_i^2 | \mathbf{x}_i = \mathbf{x})$  denote the conditional variance of  $e_i = y_i - m(\mathbf{x}_i)$ , and set  $|\mathbf{h}| = h_1 h_2 \cdots h_d$ .

**Proposition 20.1** Let  $\hat{m}(\mathbf{x})$  denote either the Nadarya-Watson or Local Linear estimator of  $m(\mathbf{x})$ . As  $n \rightarrow \infty$  and  $h_j \rightarrow 0$  such that  $n|\mathbf{h}| \rightarrow \infty$ ,

$$\sqrt{n|\mathbf{h}|} \left( \hat{m}(\mathbf{x}) - m(\mathbf{x}) - \sum_{j=1}^d h_j^2 B_j(\mathbf{x}) \right) \xrightarrow{d} N\left(0, \frac{R_k^d \sigma^2(\mathbf{x})}{f(\mathbf{x})}\right).$$

For the Nadaraya-Watson estimator

$$B_j(\mathbf{x}) = \frac{1}{2} \frac{\partial^2}{\partial x_j^2} m(\mathbf{x}) + f(\mathbf{x})^{-1} \frac{\partial}{\partial x_j} f(\mathbf{x}) \frac{\partial}{\partial x_j} m(\mathbf{x})$$

and for the Local Linear estimator

$$B_j(\mathbf{x}) = \frac{1}{2} \frac{\partial^2}{\partial x_j^2} m(\mathbf{x}).$$

We do not provide regularity condition or a formal proof of the result but instead refer interested readers to Fan and Gijbels (1996).

## 20.23 Curse of Dimensionality

The term “curse of dimensionality” is used to describe the phenomenon that the convergence rate of nonparametric estimators slows as the dimension increases.

For the multiple regressor case we define the AIMSE as the integral of the squared bias plus variance, integrating with respect to  $f(\mathbf{x})w(\mathbf{x})$  where  $w(\mathbf{x})$  is an integrable weight function. For notational simplicity consider the case that there is a single common bandwidth  $h$ . In this case the AIMSE of  $\hat{m}(\mathbf{x})$  equals

$$\text{AIMSE} = h^4 \int_S \left( \sum_{j=1}^d B_j(\mathbf{x}) \right)^2 f(\mathbf{x}) w(\mathbf{x}) d\mathbf{x} + \frac{R_k^d}{nh^d} \int_S \sigma^2(\mathbf{x}) w(\mathbf{x}) d\mathbf{x}.$$

We see that the squared bias is of order  $h^4$ , the same as in the single regressor case. The variance, however, is of larger order  $(nh^d)^{-1}$ .

If pick the bandwith to minimizing the AIMSE, we find that it takes the form  $h = cn^{-1/(4+d)}$  for some constant  $c$ . This generalizes the formula for the one-dimensional case. The rate  $n^{-1/(4+d)}$  is slower than the  $n^{-1/5}$  rate. This effectively means that with multiple regressors a larger bandwidth is required.

When the bandwidth is set as  $h = cn^{-1/(4+d)}$  then the AIMSE is of order  $O(n^{-4/(4+d)})$ . This is a slower rate of convergence than in the one-dimensional case.

**Theorem 20.10** In the multiple regression problem, the bandwidth which minimizes the AIMSE is of order  $h \sim n^{-1/(4+d)}$ . With  $h \sim n^{-1/(4+d)}$  then  $\text{AIMSE} = O(n^{-4/(4+d)})$ .

See Exercise 20.6.

We see that the optimal AIMSE rate  $O(n^{-4/(4+d)})$  depends on the dimension  $d$ . As  $d$  increases this rate slows. Thus the precision of kernel regression estimators worsens with multiple regressors. The reason is the estimator  $\hat{m}(\mathbf{x})$  is a local average of the  $y_i$  for observations such that  $\mathbf{x}_i$  is close to  $\mathbf{x}$ , and when there are multiple regressors the number of such observations is inherently smaller.

This phenomenon – that the rate of convergence of nonparametric estimation decreases as the dimension increases – is called the **curse of dimensionality**. It is common across most nonparametric estimation problems and is not specific to kernel regression.

## 20.24 Computation

Stata has two commands which implement kernel regression: `lpoly` and `npregress`. `npregress` is only available in Stata 15 or higher. `lpoly` implements local polynomial estimation for any  $p$ , including Nadaraya-Watson (the default) and local linear estimation, and selects the bandwidth using the Fan-Gijbels ROT method. It uses the Epanechnikov kernel by default, but the Gaussian can be selected as an option. The `lpoly` command automatically displays the estimated mean function along with 95% confidence bands with standard errors computed using (20.18).

The Stata command `npregress` estimates local linear (the default) regression or Nadaraya-Watson regression. By default it selects the bandwidth by cross-validation. It uses the Epanechnikov kernel by default, but the Gaussian can be selected as an option. Confidence intervals may be calculated using the percentile bootstrap. A display of the estimated mean and 95% confidence bands at specific points (computed using the percentile bootstrap) may be obtained with the postestimation command `margins`.

There are several R packages which implement kernel regression. One flexible choice is `npreg` available in the `np` package. Its default method is Nadaraya-Watson estimation using a Gaussian kernel with bandwidth selected by cross-validation. There are options which allow local linear and local polynomial estimation, alternative kernels, and alternative bandwidth selection methods.

## 20.25 Technical Proofs\*

For all technical proofs we make the simplifying assumption that the kernel function  $k(u)$  has bounded support, thus  $k(u) = 0$  for  $|u| > a$ . The results extend to the Gaussian kernel but with additional technical arguments.

**Proof of Theorem 20.1.1.** Equation (20.3) shows that

$$\mathbb{E}(\widehat{m}_{\text{nw}}(x) | \mathbf{X}) = m(x) + \frac{\widehat{b}(x)}{\widehat{f}(x)} \quad (20.26)$$

where  $\widehat{f}(x)$  is the kernel density estimator (19.2) of  $f(x)$  and

$$\widehat{b}(x) = \frac{1}{nh} \sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)(m(x_i) - m(x)). \quad (20.27)$$

Theorem 19.6 established that  $\widehat{f}(x) \xrightarrow{p} f(x)$ . The proof is completed by showing that  $\widehat{b}(x) = h^2 f(x) B_{\text{nw}}(x) + o_p(h^2 + 1/\sqrt{nh})$ .

Since  $\widehat{b}(x)$  is a sample average it has the expectation

$$\begin{aligned} \mathbb{E}(\widehat{b}(x)) &= \frac{1}{h} \mathbb{E}\left(k\left(\frac{x_i - x}{h}\right)(m(x_i) - m(x))\right) \\ &= \int_{-\infty}^{\infty} \frac{1}{h} k\left(\frac{v - x}{h}\right)(m(v) - m(x)) f(v) dv \\ &= \int_{-\infty}^{\infty} k(u) (m(x + hu) - m(x)) f(x + hu) du. \end{aligned} \quad (20.28)$$

The second equality writes the expectation as an integral with respect to the density of  $x_i$ . The third uses the change-of-variables  $v = x + hu$ . We next use the two Taylor series expansions

$$\begin{aligned} m(x + hu) - m(x) &= m'(x)hu + \frac{1}{2}m''(x)h^2u^2 + o(h^2) \\ f(x + hu) &= f(x) + f'(x)hu + o(h). \end{aligned} \quad (20.29)$$

Inserted into (20.28) we find that (20.28) equals

$$\int_{-\infty}^{\infty} k(u) \left( m'(x)hu + \frac{1}{2}m''(x)h^2u^2 + o(h^2) \right) (f(x) + f'(x)hu + o(h)) du \quad (20.30)$$

$$= h \left( \int_{-\infty}^{\infty} uk(u) du \right) m'(x) (f(x) + o(h)) \quad (20.31)$$

$$+ h^2 \left( \int_{-\infty}^{\infty} u^2 k(u) du \right) \left( \frac{1}{2}m''(x)f(x) + m'(x)f'(x) \right)$$

$$+ h^3 \left( \int_{-\infty}^{\infty} u^3 k(u) du \right) \frac{1}{2}m''(x)f'(x) + o(h^2)$$

$$= h^2 \left( \frac{1}{2}m''(x)f(x) + m'(x)f'(x) \right) + o(h^2)$$

$$= h^2 B_{\text{nw}}(x)f(x) + o(h^2).$$

The second equality uses the fact that the kernel  $k$  integrates to one, its odd moments are zero, and the kernel variance is one. We have shown that  $\mathbb{E}(\widehat{b}(x)) = B_{\text{nw}}(x)f(x)h^2 + o(h^2)$ .

Now consider the variance of  $\hat{b}(x)$ . Since  $\hat{b}(x)$  is a sample average of independent components and the variance is smaller than the second moment

$$\begin{aligned}\text{var}(\hat{b}(x)) &= \frac{1}{nh^2} \text{var}\left(k\left(\frac{x_i - x}{h}\right)(m(x_i) - m(x))\right) \\ &\leq \frac{1}{nh^2} \mathbb{E}\left(k\left(\frac{x_i - x}{h}\right)^2 (m(x_i) - m(x))^2\right) \\ &= \frac{1}{nh} \int_{-\infty}^{\infty} k(u)^2 (m(x+hu) - m(x))^2 f(x+hu) du \\ &= \frac{1}{nh} \int_{-\infty}^{\infty} u^2 k(u)^2 du (m'(x))^2 f(x) (h^2 + o(1)) \\ &\leq \frac{h}{n} \bar{k}(m'(x))^2 f(x) + o\left(\frac{h}{n}\right).\end{aligned}\tag{20.32}$$

The second equality writes the expectation as an integral. The third uses (20.29). The final inequality uses  $k(u) \leq \bar{k}$  from Definition 19.1.1 and the fact that the kernel variance is one. This shows that

$$\text{var}(\hat{b}(x)) \leq O\left(\frac{h}{n}\right).$$

Together we conclude that

$$\hat{b}(x) = h^2 f(x) B_{\text{nw}}(x) + o(h^2) + O_p\left(\sqrt{\frac{h}{n}}\right)$$

and

$$\frac{\hat{b}(x)}{\hat{f}(x)} = h^2 B_{\text{nw}}(x) + o_p(h^2) + O_p\left(\sqrt{\frac{h}{n}}\right)\tag{20.33}$$

Together with (20.26) this implies Theorem 20.1.1. ■

**Proof of Theorem 20.2.1.** Equation (20.4) states that

$$nh \text{var}(\hat{m}_{\text{nw}}(x) | \mathbf{X}) = \frac{\hat{v}(x)}{\hat{f}(x)^2}$$

where

$$\hat{v}(x) = \frac{1}{nh} \sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)^2 \sigma^2(x_i)$$

and  $\hat{f}(x)$  is the kernel density estimator (19.2) of the marginal density  $f(x)$ . Theorem 19.6 established that  $\hat{f}(x) \xrightarrow{p} f(x)$ . The proof is completed by showing that  $\hat{v}(x) \xrightarrow{p} R_k \sigma^2(x) f(x)$ .

First, writing the expectation as an integral with respect to the marginal density of  $x_i$ , making the change-of-variables  $v = x + hu$ , and appealing to the continuity of  $\sigma^2(x)$  and  $f(x)$  at  $x$ ,

$$\begin{aligned}\mathbb{E}(\hat{v}(x)) &= \int_{-\infty}^{\infty} \frac{1}{h} k\left(\frac{v-x}{h}\right)^2 \sigma^2(v) f(v) dv \\ &= \int_{-\infty}^{\infty} k(u)^2 \sigma^2(x+hu) f(x+hu) du \\ &= \int_{-\infty}^{\infty} k(u)^2 \sigma^2(x) f(x) + o(1) \\ &= R_k \sigma^2(x) f(x).\end{aligned}$$

Second, since  $\widehat{v}(x)$  is an average of independent random variables and the variance is smaller than the second moment

$$\begin{aligned} nh \text{var}(\widehat{v}(x)) &= \frac{1}{h} \text{var}\left(k\left(\frac{x_i - x}{h}\right)^2 \sigma^2(x_i)\right) \\ &\leq \frac{1}{h} \int_{-\infty}^{\infty} k\left(\frac{v - x}{h}\right)^4 \sigma^4(v) f(v) dv \\ &= \int_{-\infty}^{\infty} k(u)^4 \sigma^4(x + hu) f(x + hu) du \\ &\leq \bar{k}^2 R_k \sigma^4(x) f(x) + o(1) \end{aligned}$$

so  $\text{var}(\widehat{v}(x)) \rightarrow 0$ .

We deduce from Markov's inequality that  $\widehat{v}(x) \xrightarrow{P} R_k \sigma^2(x) f(x)$ , completing the proof. ■

**Proof of Theorem 20.6.** Observe that  $m(x_i) - \tilde{m}_{-i}(x_i, h)$  is a function only of  $(x_1, \dots, x_n)$  and  $(e_1, \dots, e_n)$  excluding  $e_i$ , and is thus uncorrelated with  $e_i$ . Since  $\tilde{e}_i(h) = m(x_i) - \tilde{m}_{-i}(x_i, h) + e_i$ , then

$$\begin{aligned} \mathbb{E}(\text{CV}(h)) &= \mathbb{E}(\tilde{e}_i(h)^2 w(x_i)) \\ &= \mathbb{E}(e_i^2 w(x_i)) + \mathbb{E}((\tilde{m}_{-i}(x_i, h) - m(x_i))^2 w(x_i)) \\ &\quad + 2\mathbb{E}((\tilde{m}_{-i}(x_i, h) - m(x_i)) w(x_i) e_i) \\ &= \bar{\sigma}^2 + \mathbb{E}((\tilde{m}_{-i}(x_i, h) - m(x_i))^2 w(x_i)). \end{aligned} \quad (20.34)$$

The second term is an expectation over the random variables  $x_i$  and  $\tilde{m}_{-i}(x, h)$ , which are independent as the second is not a function of the  $i^{th}$  observation. Thus taking the conditional expectation given the sample excluding the  $i^{th}$  observation, this is the expectation over  $x_i$  only, which is the integral with respect to its density

$$\mathbb{E}_{-i}((\tilde{m}_{-i}(x_i, h) - m(x_i))^2 w(x_i)) = \int (\tilde{m}_{-i}(x, h) - m(x))^2 f(x) w(x) dx.$$

Taking the unconditional expectation yields

$$\begin{aligned} \mathbb{E}((\tilde{m}_{-i}(x_i, h) - m(x_i))^2 w(x_i)) &= \mathbb{E} \int (\tilde{m}_{-i}(x, h) - m(x))^2 f(x) w(x) dx \\ &= \text{IMSE}_{n-1}(h) \end{aligned}$$

where this is the IMSE of a sample of size  $n - 1$  as the estimator  $\tilde{m}_{-i}$  uses  $n - 1$  observations. Combined with (20.34) we obtain (20.12), as desired. ■

**Proof of Theorem 20.7.** We can write the Nadaraya-Watson estimator as

$$\hat{m}_{\text{nw}}(x) = m(x) + \frac{\hat{b}(x)}{\hat{f}(x)} + \frac{\hat{g}(x)}{\hat{f}(x)} \quad (20.35)$$

where  $\hat{f}(x)$  is the kernel density estimator (19.2),  $\hat{b}(x)$  is defined in (20.27), and

$$\hat{g}(x) = \frac{1}{nh} \sum_{i=1}^n k\left(\frac{x_i - x}{h}\right) e_i. \quad (20.36)$$

Since  $\hat{f}(x) \xrightarrow{P} f(x) > 0$  by Theorem 19.6, the proof is completed by showing  $\hat{b}(x) \xrightarrow{P} 0$  and  $\hat{g}(x) \xrightarrow{P} 0$ .

Take  $\hat{b}(x)$ . From (20.28) and the continuity of  $m(x)$  and  $f(x)$

$$\mathbb{E}(\hat{b}(x)) = \int_{-\infty}^{\infty} k(u) (m(x + hu) - m(x)) f(x + hu) du = o(1)$$

as  $h \rightarrow \infty$ . From (20.32),

$$nh \text{var}(\hat{b}(x)) \leq \int_{-\infty}^{\infty} k(u)^2 (m(x+hu) - m(x))^2 f(x+hu) du = o(1)$$

as  $h \rightarrow \infty$ . Thus  $\text{var}(\hat{b}(x)) \rightarrow 0$ . By Markov's inequality we conclude  $\hat{b}(x) \xrightarrow{p} 0$ .

Take  $\hat{g}(x)$ . Since  $\hat{g}(x)$  is linear in  $e_i$  and  $\mathbb{E}(e_i | x_i) = 0$ , we find  $\mathbb{E}(\hat{g}(x)) = 0$ . Since  $\hat{g}(x)$  is an average of independent random variables, the variance is smaller than the second moment, and the definition  $\sigma^2(x_i) = \mathbb{E}(e_i^2 | x_i)$

$$\begin{aligned} nh \text{var}(\hat{g}(x)) &= \frac{1}{h} \text{var}\left(k\left(\frac{x_i - x}{h}\right) e_i\right) \\ &\leq \frac{1}{h} \mathbb{E}\left(k\left(\frac{x_i - x}{h}\right)^2 e_i^2\right) \\ &= \frac{1}{h} \mathbb{E}\left(k\left(\frac{x_i - x}{h}\right)^2 \sigma^2(x_i)\right) \\ &= \int_{-\infty}^{\infty} k(u)^2 \sigma^2(x+hu) f(x+hu) du \\ &= R_k \sigma^2(x) f(x) + o(1) \end{aligned} \quad (20.37)$$

since  $\sigma^2(x)$  and  $f(x)$  are continuous in  $x$ . Thus  $\text{var}(\hat{g}(x)) \rightarrow 0$ . By Markov's inequality we conclude Thus  $\hat{g}(x) \xrightarrow{p} 0$ , completing the proof. ■

**Proof of Theorem 20.8.** From (20.35), Theorem 19.6, and (20.33) we have

$$\begin{aligned} \sqrt{nh}(\hat{m}_{\text{nw}}(x) - m(x) - h^2 B_{\text{nw}}(x)) &= \sqrt{nh}\left(\frac{\hat{g}(x)}{\hat{f}(x)}\right) + \sqrt{nh}\left(\frac{\hat{b}(x)}{\hat{f}(x)} - h^2 B_{\text{nw}}(x)\right) \\ &= \sqrt{nh}\left(\frac{\hat{g}(x)}{\hat{f}(x)}\right)(1 + o_p(1)) + \sqrt{nh}\left(o_p(h^2) + O_p\left(\sqrt{\frac{h}{n}}\right)\right) \\ &= \sqrt{nh}\left(\frac{\hat{g}(x)}{\hat{f}(x)}\right)(1 + o_p(1)) + \left(o_p(\sqrt{nh^5}) + O_p(h)\right) \\ &= \sqrt{nh}\left(\frac{\hat{g}(x)}{\hat{f}(x)}\right) \end{aligned}$$

where the final equality holds since  $\sqrt{nh}\hat{g}(x) = O_p(1)$  by (20.37) and the assumption  $nh^5 = O(1)$ . The proof is completed by showing  $\sqrt{nh}\hat{g}(x) \xrightarrow{d} N(0, R_k \sigma^2(x) f(x))$ .

Define  $y_{ni} = h^{-1/2} k\left(\frac{x_i - x}{h}\right) e_i$  which is mean zero. Then we can write  $\sqrt{nh}\hat{g}(x) = \sqrt{ny}$ . We verify the conditions for the Lindeberg CLT (Theorem 6.12). The summands  $y_{ni}$  are independent and mean zero. In the notation of Theorem 6.12, set  $\bar{\sigma}_n^2 = \text{var}(\sqrt{ny}) \rightarrow R_k f(x) \sigma^2(x)$  as  $h \rightarrow 0$ . The CLT holds if we can verify the Lindeberg condition.

It turns out that this is a quite advanced calculation and will not interest most readers. It is provided for those interested in a complete derivation.

Fix  $\varepsilon > 0$  and  $\delta > 0$ . Since  $k(u)$  is bounded we can write  $k(u) \leq \bar{k}$ . Let  $nh$  be sufficiently large so that

$$\left(\frac{\varepsilon nh}{\bar{k}}\right)^{r-2} \geq \frac{\bar{\sigma}}{\delta}.$$

The conditional moment bound (20.14) implies that for  $x \in \mathcal{N}$ ,

$$\begin{aligned} \mathbb{E}\left(e_i^2 \mathbf{1}\left(e_i^2 > \varepsilon nh/\bar{k}\right) | x_i = x\right) &= \mathbb{E}\left(\frac{|e_i|^r}{|e_i|^{r-2}} \mathbf{1}\left(e_i^2 > \varepsilon nh/\bar{k}\right) | x_i = x\right) \\ &\leq \mathbb{E}\left(\frac{|e_i|^r}{(\varepsilon nh/\bar{k})^{(r-2)/2}} | x_i = x\right) \\ &\leq \delta. \end{aligned}$$

Since  $y_{ni}^2 \leq h^{-1}\bar{k}e_i^2$  we find

$$\begin{aligned}
\mathbb{E}(y_{ni}^2 \mathbf{1}(y_{ni}^2 > \varepsilon n)) &\leq \frac{1}{h} \mathbb{E}\left(k\left(\frac{x_i - x}{h}\right)^2 e_i^2 \mathbf{1}(e_i^2 > \varepsilon nh/\bar{k})\right) \\
&= \frac{1}{h} \mathbb{E}\left(k\left(\frac{x_i - x}{h}\right)^2 \mathbb{E}\left(e_i^2 \mathbf{1}(e_i^2 > \varepsilon nh/\bar{k}) | x_i\right)\right) \\
&= \int_{-\infty}^{\infty} k(u)^2 \mathbb{E}\left(e_i^2 \mathbf{1}(e_i^2 > \varepsilon nh/\bar{k}) | x_i = x + hu\right) f(x + hu) du \\
&\leq \delta \int_{-\infty}^{\infty} k(u)^2 f(x + hu) du \\
&= \delta R_k f(x) + o(1) \\
&= o(1)
\end{aligned}$$

since  $\delta$  is arbitrary. This is the Lindeberg condition (6.5). The Lindeberg CLT (Theorem 6.12) shows that

$$\sqrt{nh}\hat{g}(x) = \sqrt{n}y \xrightarrow{d} N(0, R_k \sigma^2(x) f(x)).$$

This completes the proof. ■

## Exercises

**Exercise 20.1** For kernel regression, suppose you rescale  $y$ , for example replace  $y_i$  with  $100y_i$ , how should the bandwidth  $h$  change? To answer this, first address how the functions  $m(x)$  and  $\sigma^2(x)$  change under rescaling, and then calculate how  $\bar{B}$  and  $\bar{\sigma}^2$  change. Deduce how the optimal  $h_0$  changes due to rescaling  $y_i$ . Does your answer make intuitive sense?

**Exercise 20.2** Show that (20.6) minimizes the AIMSE (20.5).

**Exercise 20.3** Describe in words how the bias of the local linear estimator changes over regions of convexity and concavity in  $m(x)$ . Does this make intuitive sense?

**Exercise 20.4** Suppose the true regression function is linear  $m(x) = \alpha + \beta x$  and we estimate the function using the Nadaraya-Watson estimator. Calculate the bias function  $B(x)$ . Suppose  $\beta > 0$ . For which regions is  $B(x) > 0$  and for which regions is  $B(x) < 0$ ? Now suppose that  $\beta < 0$  and re-answer the question. Can you intuitively explain why the NW estimator is positively and negatively biased for these regions?

**Exercise 20.5** Suppose  $m(x) = \alpha$  is a constant function. Find the AIMSE-optimal bandwith (20.6) for NW estimation? Explain.

**Exercise 20.6** Prove Theorem 20.10: Show that when  $d \geq 1$  the AIMSE optimal bandwidth takes the form  $h_0 = cn^{-1/(4+d)}$  and AIMSE is  $O(n^{-4/(4+d)})$ .

**Exercise 20.7** Take the DDK2011 dataset and the subsample of boys who experienced tracking. As in Section 20.21, use the Local Linear estimator to estimate the regression of *testscores* on *percentile*, but now with the subsample of boys. Plot with 95% confidence intervals. Comment on the similarities and differences with the estimate for the subsample of girls.

**Exercise 20.8** Take the cps09mar dataset and the subsample of individuals with *education*=20 (professional degree or doctorate), with *experience* between 0 and 40 years.

- (a) Use Nadaraya-Watson to estimate the regression of  $\log(wage)$  on *experience*, separately for men and women. Plot with 95% confidence intervals. Comment on how the estimated wage profiles vary with experience. In particular, do you think the evidence suggests that expected wages fall for experience levels above 20 for this education group?
- (b) Repeat using the Local Linear estimator. How do the estimates and confidence intervals change?

**Exercise 20.9** Take the Invest1993 dataset and the subsample of observations with  $Q \leq 5$ .

- (a) Use Nadaraya-Watson to estimate the regression of *I* on *Q*. Plot with 95% confidence intervals.
- (b) Repeat using the Local Linear estimator.
- (c) Is there evidence to suggest that the regression function is non-linear?

**Exercise 20.10** The RR2010 dataset is from Reinhart and Rogoff (2010). It contains observations on annual U.S. GDP growth rates, inflation rates, and the debt/gdp ratio for the long time span 1791-2009. The paper made the strong claim that gdp growth slows as debt/gdp increases, and in particular that this relationship is nonlinear with debt negatively affecting growth for debt ratios exceeding 90%. Their full dataset includes 44 countries, our extract only includes the United States.

- (a) Use Nadaraya-Watson to estimate the regression of gdp growth on the debt ratio. Plot with 95% confidence intervals.

- (b) Repeat using the Local Linear estimator.
- (c) Do you see evidence of nonlinearity, and/or a change in the relationship at 90%?
- (d) Now estimate a regression of gdp growth on the inflation rate. Comment on what you find.

**Exercise 20.11** We will consider a nonlinear AR(1) model for gdp growth rates

$$\begin{aligned}y_t &= m(y_{t-1}) + e_t \\y_t &= 100 \left( \left( \frac{GDP_t}{GDP_{t-1}} \right)^4 - 1 \right)\end{aligned}$$

- (a) Create GDP growth rates  $y_t$ . Extract the level of real U.S. GDP (*GDPCI*) from the FRED-QD dataset and make the above transformation to growth rates.
- (b) Use Nadaraya-Watson to estimate  $m(x)$ . Plot with 95% confidence intervals.
- (c) Repeat using the Local Linear estimator.
- (d) Do you see evidence of nonlinearity?

# Chapter 21

## Series Regression

### 21.1 Introduction

Chapter 20 studied nonparametric regression by kernel smoothing methods. In this chapter we study an alternative class of nonparametric regression methods known as series regression.

The basic model is identical to that examined in Chapter 20. We assume that there are pairs  $(y_i, x_i)$  such that  $\mathbb{E}(y_i^2) < \infty$  and satisfy the regression model

$$\begin{aligned} y_i &= m(x_i) + e_i \\ \mathbb{E}(e_i | x_i) &= 0 \\ \mathbb{E}(e_i^2 | x_i) &= \sigma^2(x_i). \end{aligned} \tag{21.1}$$

The goal is to estimate the conditional mean function  $m(x)$ . We start with the simple setting where  $x_i$  is scalar and consider more general cases later.

A series regression model is a sequence  $K = 1, 2, \dots$ , of approximating models  $m_K(x)$  with  $K$  parameters. In this chapter we exclusively focus on linear series models, and in particular polynomials and splines. This is because these are simple, convenient, and cover most applications of series methods in applied economics. Other series models include trigonometric polynomials, wavelets, orthogonal wavelets, B-splines, and neural networks. For a detailed review see Chen (2007).

Linear series regression models take the form

$$y_i = \mathbf{x}'_{Ki} \boldsymbol{\beta}_K + e_{Ki} \tag{21.2}$$

where  $\mathbf{x}_{Ki} = \mathbf{x}_K(x_i)$  is a vector of regressors obtained by making transformations of  $x_i$ , and  $\boldsymbol{\beta}_K$  is a coefficient vector. There are multiple possible definitions of the coefficient  $\boldsymbol{\beta}_K$ . We define<sup>1</sup> it by projection

$$\begin{aligned} \boldsymbol{\beta}_K &= \mathbb{E}(\mathbf{x}_{Ki} \mathbf{x}'_{Ki})^{-1} \mathbb{E}(\mathbf{x}_{Ki} y_i) \\ &= \mathbb{E}(\mathbf{x}_{Ki} \mathbf{x}'_{Ki})^{-1} \mathbb{E}(\mathbf{x}_{Ki} m(x_i)). \end{aligned} \tag{21.3}$$

The series regression error  $e_{Ki}$  is defined by (21.2) and (21.3), is distinct from the regression error  $e_i$  in (21.1), and is indexed by  $K$  since it depends on the specific regressors  $\mathbf{x}_{Ki}$ . The series approximation to  $m(x)$  is

$$m_K(x) = \mathbf{x}_K(x)' \boldsymbol{\beta}_K. \tag{21.4}$$

The coefficient is typically<sup>2</sup> estimated by least-squares

$$\hat{\boldsymbol{\beta}}_K = \left( \sum_{i=1}^n \mathbf{x}_{Ki} \mathbf{x}'_{Ki} \right)^{-1} \left( \sum_{i=1}^n \mathbf{x}_{Ki} y_i \right) = (\mathbf{X}'_K \mathbf{X}_K)^{-1} (\mathbf{X}'_K \mathbf{y}). \tag{21.5}$$

---

<sup>1</sup>An alternative is to define  $\boldsymbol{\beta}_K$  as the best uniform approximation as in (21.8). It is not critical so long as we are careful to be consistent with our notation.

<sup>2</sup>Penalized estimators have also been recommended. We do not review these methods here.

The estimator for  $m(x)$  is

$$\hat{m}_K(x) = \mathbf{x}_K(x)' \hat{\boldsymbol{\beta}}_K. \quad (21.6)$$

The difference between specific models arises due to the different choices of transformations  $\mathbf{x}_K(x)$ .

The theoretical issues we will explore in this chapter are: (1) Approximation properties of polynomials and splines; (2) Consistent estimation of  $m(x)$ ; (3) Asymptotic normal approximations; (4) Selection of  $K$ ; (5) Extensions.

For a textbook treatment of series regression see Li and Racine (2007). For an advanced treatment see Chen (2007). Two seminal contributions are Andrews (1991a) and Newey (1997). Two recent important papers are Belloni, Chernozhukov, Chetverikov, and Kato (2015) and Chen and Christensen (2015).

## 21.2 Polynomial Regression

The prototypical series regression model for  $m(x)$  is a  $p^{th}$  order polynomial

$$m_K(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p.$$

We can write it in vector notation as (21.4) where

$$\mathbf{x}_K(x) = \begin{pmatrix} 1 \\ x \\ \vdots \\ x^p \end{pmatrix}.$$

The number of parameters is  $K = p + 1$ . Notice that we index  $\mathbf{x}_K(x)$  and  $\boldsymbol{\beta}_K$  by  $K$  as their dimensions and values vary with  $K$ .

The implied **polynomial regression model** for the random pair  $(y_i, x_i)$  is (21.2) with

$$\mathbf{x}_{Ki} = \mathbf{x}_K(x_i) = \begin{pmatrix} 1 \\ x_i \\ \vdots \\ x_i^p \end{pmatrix}.$$

The degree of flexibility of a polynomial regression is controlled by the polynomial order  $p$ . A larger  $p$  yields a more flexible model, while a smaller  $p$  typically results in a estimator with a smaller variance.

In general, a **linear series regression model** takes the form

$$m_K(x) = \beta_1 \tau_1(x) + \beta_2 \tau_2(x) + \cdots + \beta_K \tau_K(x)$$

where the functions  $\tau_j(x)$  are called the **basis transformations**. The polynomial regression model uses the power basis  $\tau_j(x) = x^{j-1}$ . The model  $m_K(x)$  is called a series regression because it is obtained by sequentially adding the series of variables  $\tau_j(x)$ .

## 21.3 Illustrating Polynomial Regression

Consider the `cps09mar` dataset and a regression of log wages on *experience* for women with a college education (*education*= 16), separately for white women and black women. The classical Mincer model uses a quadratic in experience. Given the large sample sizes (4682 for white women and 517 for black women) we can consider higher order polynomials. In Figure 21.1 we plot least-squares estimates of the conditional mean functions using polynomials of order 2, 4, 8, and 12.

Examine panel (a), which shows the estimates for the sub-sample of white women. The quadratic specification appears mis-specified, with a shape noticeably different from the other estimates. The difference between the polynomials of order 4, 8, and 12 is relatively minor, especially for experience levels below 20.

Now examine panel (b), which shows the estimates for the sub-sample of black women. This panel is quite different from panel (a). The estimates are erratic, and increasingly so as the polynomial order increases. Assuming we are expecting a concave (or nearly concave) experience profile, the only estimate which satisfies this is the quadratic.

Why the difference between panels (a) and (b)? The most likely explanation is the different sample sizes. The sub-sample of black women has much fewer observations so the mean function is much less precisely estimated, giving rise to the erratic plots. This suggests (informally) that it may be preferred to use a smaller polynomial order  $p$  in the second sub-sample, or equivalently to use a larger  $p$  when the sample size  $n$  is larger. The idea that model complexity – the number of coefficients  $K$  – should vary with sample size  $n$  is an important feature of series regression.

The erratic nature of the estimated polynomial regressions in Figure 21.1(b) is a common feature of higher-order estimated polynomial regressions. Better results can sometimes be obtained by a spline regression, which is described in Section 21.5.

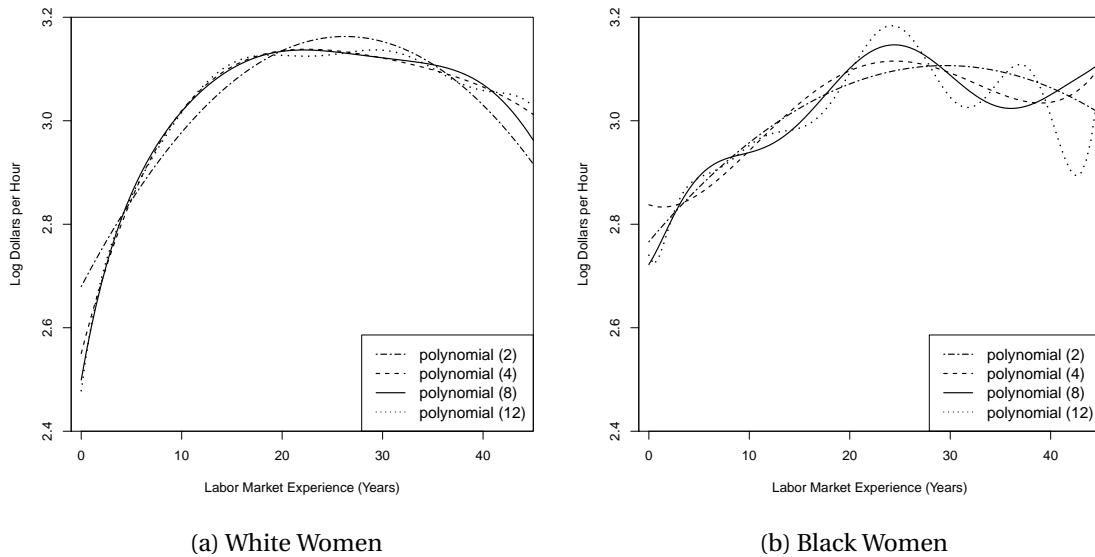


Figure 21.1: Polynomial Estimates of Experience Profile, College-Educated Women

## 21.4 Orthogonal Polynomials

Standard implementation of the least-squares estimator (21.5) of a polynomial regression may return a computational error message when  $p$  is large. (See Section 3.24.) This is because the moments of  $x_i^j$  can be highly heterogeneous across  $j$ , and because the variables  $x_i^j$  can be highly correlated. These two factors imply in practice that the matrix  $\mathbf{X}'_K \mathbf{X}_K$  can be ill-conditioned (the ratio of the largest to smallest eigenvalue can be quite large) and some packages will return error messages rather than compute  $\hat{\beta}_K$ .

In most cases the condition of  $\mathbf{X}'_K \mathbf{X}_K$  can be dramatically improved by rescaling the observations. As discussed in Section 3.24, a simple method for non-negative regressors is to rescale each by its sample mean, e.g. replace  $x_i^j$  with  $x_i^j / (n^{-1} \sum_{i=1}^n x_i^j)$ . Even better conditioning can often be obtained by rescaling  $x_i$  to lie in  $[-1, 1]$  before applying powers. In most applications one of these methods will be sufficient for a well-conditioned regression.

A computationally more robust implementation can be obtained by using orthogonal polynomials. These are linear combinations of the polynomial basis functions, and produce identical regression estimators (21.6). The goal of orthogonal polynomials is to produce regressors which are either orthogonal or close to orthogonal, and have similar variances, so that  $\mathbf{X}'_K \mathbf{X}_K$  is close to diagonal with similar diag-

onal elements. These orthogonalized regressors  $\mathbf{x}_{Ki}^* = \mathbf{A}_K \mathbf{x}_{Ki}$  can be written as linear combinations of the original variables  $\mathbf{x}_{Ki}$ . If the regressors are orthogonalized, then the regression estimator (21.6) is modified by replacing  $\mathbf{x}_K(x)$  with  $\mathbf{x}_K^*(x) = \mathbf{A}_K \mathbf{x}_K(x)$ .

One approach is to use sample orthogonalization. This is done by a sequence of regressions of  $x_i^j$  on the previously orthogonalized variables, and then rescaling. This will result in perfectly orthogonalized variables. This is what is implemented in many statistical packages under the label “orthogonal polynomials”, for example, the function `poly` in R. If this is done then the least-squares coefficients have no meaning outside this specific sample, and it is not convenient for calculation of  $\hat{m}_K(x)$  for values of  $x$  other than sample values. This is the approach used for the examples presented in the previous section.

Another approach is to use an algebraic orthogonal polynomial. This is a polynomial which is orthogonal with respect to a known weight function  $w(x)$ . Specifically, it is a sequence  $p_j(x)$ ,  $j = 0, 1, 2, \dots$ , with the property that  $\int p_j(x) p_\ell(x) w(x) dx = 0$  for  $j \neq \ell$ . This means that if  $w(x) = f(x)$ , the marginal density of  $x_i$ , then the basis transformations  $p_j(x_i)$  will be mutually orthogonal (in expectation). Since we do now know the density of  $x_i$  this is not feasible in practice, but if  $w(x)$  is close to the density of  $x_i$ , then we can expect that the basis transformations will be close to mutually orthogonal. To implement an algebraic orthogonal polynomial, you first should rescale your  $x_i$  variable so that it satisfies the support for the weight function  $w(x)$ .

The following three choices are most relevant for economic applications.

**Legendre Polynomial.** These are orthogonal with respect to the uniform density on  $[-1, 1]$ . (So should be applied to regressors scaled to have support in  $[-1, 1]$ .)

$$p_j(x) = \frac{1}{2^j} \sum_{\ell=0}^j \binom{j}{\ell}^2 (x-1)^{j-\ell} (x+1)^\ell.$$

For example, the first several are  $p_0(x) = 1$ ,  $p_1(x) = x$ ,  $p_2(x) = (3x^2 - 1)/2$ , and  $p_3(x) = (5x^3 - 3x)/2$ . The best computational method is to use the recurrence relationship

$$p_{j+1}(x) = \frac{(2j+1)x p_j(x) - j p_{j-1}(x)}{j+1}.$$

**Laguerre Polynomial.** These are orthogonal with respect to the exponential density  $e^{-x}$  on  $[0, \infty)$ . (So should be applied to non-negative regressors scaled to have unit mean and variance.)

$$p_j(x) = \sum_{\ell=0}^j \binom{j}{\ell} \frac{(-x)^\ell}{\ell!}.$$

For example, the first several are  $p_0(x) = 1$ ,  $p_1(x) = 1-x$ ,  $p_2(x) = (x^2 - 4x + 2)/2$ , and  $p_3(x) = (-x^3 + 9x^2 - 18x + 6)/6$ . The best computational method is to use the recurrence relationship

$$p_{j+1}(x) = \frac{(2j+1-x) p_j(x) - j p_{j-1}(x)}{j+1}.$$

**Hermite Polynomial.** These are orthogonal with respect to the standard normal density on  $(-\infty, \infty)$ . (So should be applied to regressors scaled to have mean zero and variance one.)

$$p_j(x) = j! \sum_{\ell=0}^{\lfloor j/2 \rfloor} \frac{(-1/2)^\ell x^{\ell-2j}}{\ell! (j-2\ell)!}.$$

For example, the first several are  $p_0(x) = 1$ ,  $p_1(x) = x$ ,  $p_2(x) = x^2 - 1$ , and  $p_3(x) = x^3 - 3x$ . The best computational method is to use the recurrence relationship

$$p_{j+1}(x) = x p_j(x) - j p_{j-1}(x).$$

The R package `orthopolynom` provides a convenient set of commands to compute many orthogonal polynomials, including the above.

## 21.5 Splines

A **spline** is a piecewise polynomial. Typically the order of the polynomial is pre-selected to be linear, quadratic, or cubic. The flexibility of the model is determined by the number of polynomial segments. The join points between these segments are called **knots**.

To impose smoothness and parsimony it is common to constrain the spline function to have continuous derivatives up to the order of the spline. Thus a linear spline is constrained to be continuous, a quadratic spline is constrained to have a continuous first derivative, and a cubic spline is constrained to have continuous first and second derivatives.

A simple way to construct a regression spline is as follows. A linear spline with one knot  $\tau$  is

$$m_K(x) = \beta_0 + \beta_1 x + \beta_2 (x - \tau) \mathbf{1}(x \geq \tau).$$

To see that this is a linear spline, observe that for  $x \leq \tau$  the function  $m_K(x) = \beta_0 + \beta_1 x$  is linear with slope  $\beta_1$ ; for  $x \geq \tau$  the function  $m_K(x) = (\beta_0 - \beta_0 \tau) + (\beta_1 + \beta_2)x$  is linear with slope  $\beta_1 + \beta_2$ , and the function is continuous at  $x = \tau$ . Note that  $\beta_2$  is the change in the slope at  $\tau$ . A linear spline with two knots  $\tau_1 < \tau_2$  is

$$m_K(x) = \beta_0 + \beta_1 x + \beta_2 (x - \tau_1) \mathbf{1}(x \geq \tau_2) + \beta_3 (x - \tau_2) \mathbf{1}(x \geq \tau_2).$$

A quadratic spline with one knot is

$$m_K(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 (x - \tau)^2 \mathbf{1}(x \geq \tau).$$

To see that this is a quadratic spline, observe that for  $x \leq \tau$  the function is the quadratic  $\beta_0 + \beta_1 x + \beta_2 x^2$ , for  $x \geq \tau$  it is the quadratic  $\beta_0 + \beta_3 \tau^2 + (\beta_1 - 2\beta_3 \tau)x + (\beta_2 + \beta_3)x^2$ , and the first derivative is  $\beta_1 + 2\beta_2 x$  and continuous at  $x = \tau$ .

In general, a  $p^{th}$ -order spline with  $N$  knots  $\tau_1 < \tau_2 < \dots < \tau_N$  is

$$m_K(x) = \sum_{j=0}^p \beta_j x^j + \sum_{k=1}^N \beta_{p+k} (x - \tau_k)^p \mathbf{1}(x \geq \tau_k)$$

which has  $K = N + p + 1$  coefficients.

The implied **spline regression model** for the random pair  $(y_i, x_i)$  is (21.2) where

$$\mathbf{x}_{Ki} = \mathbf{x}_K(x_i) = \begin{pmatrix} 1 \\ x_i \\ \vdots \\ x_i^p \\ (x_i - \tau_1)^p \mathbf{1}(x_i \geq \tau_1) \\ \vdots \\ (x_i - \tau_N)^p \mathbf{1}(x_i \geq \tau_N) \end{pmatrix}.$$

In practice a spline will depend critically on the choice of the knots  $\tau_k$ . When  $x_i$  is bounded with an approximately uniform distribution it is common to space the knots evenly so all segments have the same length. When the distribution of  $x_i$  is not uniform an alternative is to set the knots at the quantiles  $j/(N+1)$  so that the probability mass is equalized across segments. A third alternative is to set the knots at the points where  $m(x)$  has the greatest change in curvature (see Schumaker (2007), Chapter 7). In all cases the set of knots  $\tau_j$  can change with  $K$ . Therefore a spline is a special case of an approximation of the form

$$m_K(x) = \beta_1 \tau_{1K}(x) + \beta_2 \tau_{2K}(x) + \dots + \beta_K \tau_{KK}(x)$$

where the **basis transformations**  $\tau_{jK}(x)$  depend on both  $j$  and  $K$ . Many authors call such approximations a **sieve** rather than a series, because the basis transformations change with  $K$ . This distinction is not critical to our treatment so for simplicity we refer to splines as series regression models.

## 21.6 Illustrating Spline Regression

In Section 21.3 we illustrated regressions of log wages on *experience* for white and black women with a college education. Now we consider a similar regression for black men with a college education, a sub-sample with 394 observations.

We use a quadratic spline with four knots at experience levels of 10, 20, 30, and 40. This is a regression model with seven coefficients. The estimated regression function is displayed in Figure 21.2. An estimated 6<sup>th</sup> order polynomial regression is also displayed for comparison (a 6<sup>th</sup> order polynomial is an appropriate comparison because it also has seven coefficients).

While the spline is a quadratic over each segment, what you can see is that the first two segments (for experience levels between 0-10 and 10-20 years) are essentially linear. Most of the curvature occurs in the third and fourth segments (20-30 and 30-40 years), where the estimated regression function peaks and twists into a negative slope. The estimated regression function is quite smooth.

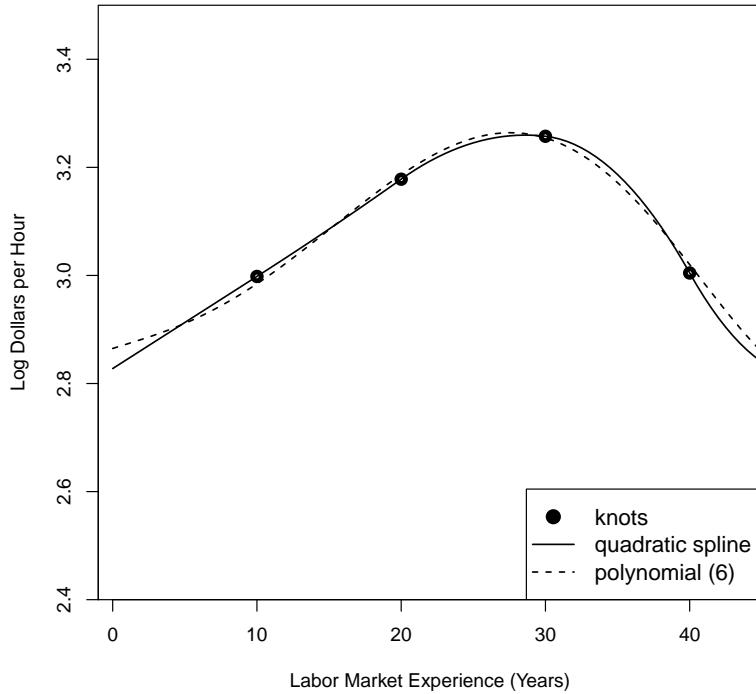


Figure 21.2: Quadratic Spline Estimate of Experience Profile, College-Educated Black Men

A quadratic (or cubic) spline is useful when it is desired to impose smoothness as in Figure 21.2. In contrast, a linear spline is useful when it is desired to allow for sharp changes in slope.

To illustrate we consider the data set CHJ2004 which is a sample of 8684 urban Phillipino households from Cox, Hansen, and Jimenez (2004). This paper studied the crowding-out impact of a family's income on non-governmental (e.g., extended family) income transfers. A model of altruistic transfers predicts that extended families will make gifts (transfers) when the recipient family's income is sufficiently low, but will not make transfers if the recipient family's income exceeds a threshold. A pure altruistic model predicts that the regression of transfers received on family income should be negative with a slope of  $-1$  up to this threshold, and be flat above this threshold. We estimated this regression (including a set of additional controls) using a linear spline with knots at 10000, 20000, 30000, 40000, 50000, 60000, 100000, and 150000 pesos. These knots were selected to give considerable flexibility for low income levels and greater smoothness at higher income levels where there are fewer observations. This model has a total of 26 coefficients.

The estimated regression function (as a function of household income) is displayed in Figure 21.3. For the first two segments (incomes levels below 20000 pesos) the regression function is negatively sloped as predicted, with a slope about  $-0.7$  from 0 to 10000 pesos, and  $-0.3$  from 10000 to 20000 pesos. The estimated regression function is effectively flat for income levels above 20000 pesos. This shape is highly consistent with the pure altruism model. A linear spline model is particularly well suited for this application as it allows for discontinuous changes in slope.

Linear spline models with a single knot have been recently popularized by Card, Lee, Pei, and Weber (2015) with the label **regression kink design**.

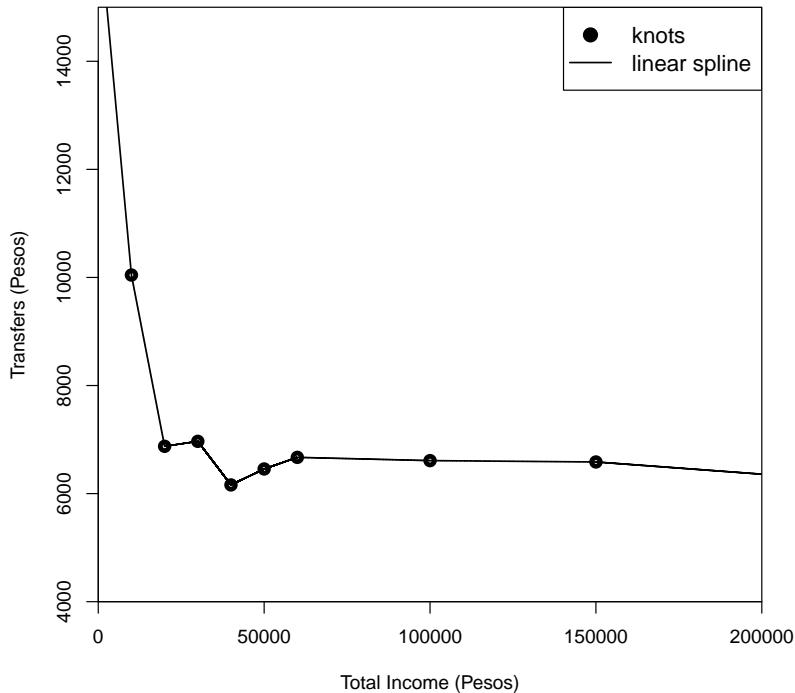


Figure 21.3: Linear Spline Estimate of Effect of Income on Transfers

## 21.7 The Global/Local Nature of Series Regression

Recall from Section 20.18 that we described kernel regression as inherently local in nature. The Nadaraya-Watson, Local Linear, and Local Polynomial estimators of the conditional mean  $m(x)$  are weighted averages of  $y_i$  for observations for which  $x_i$  is close to  $x$ .

In contrast, series regression methods are typically described as global in nature. The estimator  $\hat{m}_K(x) = \mathbf{x}_K(x)' \hat{\beta}_K$  is a function of the entire sample. The coefficients of a fitted polynomial (or spline) are affected by the global shape of the function  $m(x)$ , and thus affect the estimator  $\hat{m}_K(x)$  at any local point  $x$ .

While this description has some merit, it is not a complete description. As we now show, series regression estimators share the local smoothing property of kernel regression. As the number of series terms  $K$  increase a series estimator  $\hat{m}_K(x) = \mathbf{x}_K(x)' \hat{\beta}_K$  also becomes a local weighted average estimator.

To see this, observe that we can write the estimator as

$$\begin{aligned}\hat{m}_K(x) &= \mathbf{x}_K(x)' (\mathbf{X}'_K \mathbf{X}_K)^{-1} (\mathbf{X}'_K \mathbf{y}) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbf{x}_K(x)' \hat{\mathbf{Q}}_K^{-1} \mathbf{x}_K(x_i) y_i \\ &= \frac{1}{n} \sum_{i=1}^n \hat{w}_K(x, x_i) y_i\end{aligned}$$

where  $\hat{\mathbf{Q}}_K = n^{-1} \mathbf{X}'_K \mathbf{X}_K$  and

$$\hat{w}_K(x, u) = \mathbf{x}_K(x)' \hat{\mathbf{Q}}_K^{-1} \mathbf{x}_K(u).$$

Thus  $\hat{m}_K(x)$  is a weighted average of  $y_i$  using the weights  $\hat{w}_K(x, x_i)$ . The weight function  $\hat{w}_K(x, x_i)$  appears to be maximized at  $x_i = x$ , so  $\hat{m}(x)$  puts more weight on observations for which  $x_i$  is close to  $x$ , similarly to kernel regression.

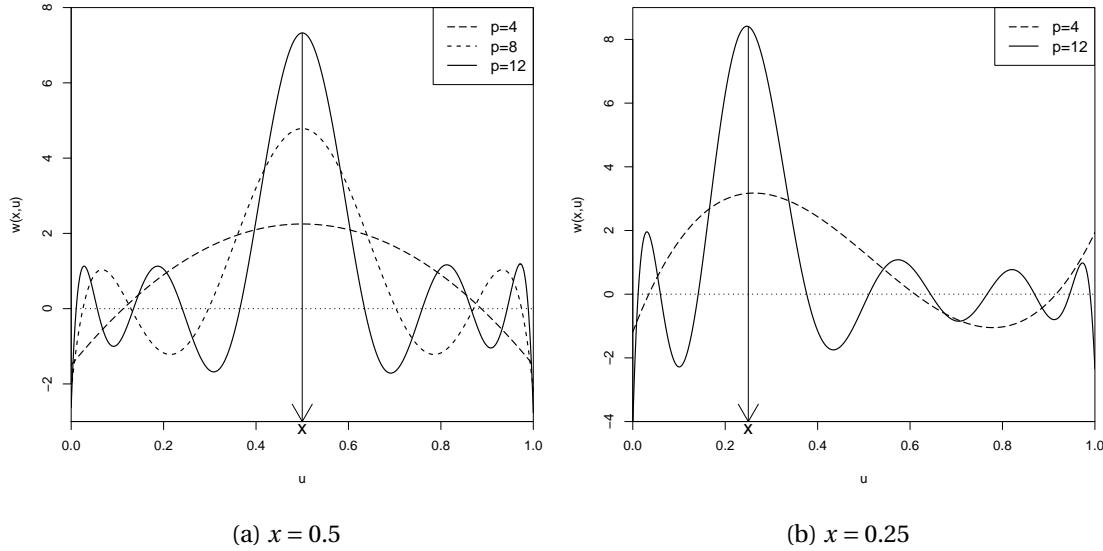


Figure 21.4: Kernel Representation of Polynomial Weight Function

To see this more precisely, observe that since  $\hat{\mathbf{Q}}_K$  will be close (in large samples) to  $\mathbf{Q}_K = \mathbb{E}(\mathbf{x}_{Ki} \mathbf{x}'_{Ki})$ ,  $\hat{w}_K(x, u)$  will be close to the deterministic weight function

$$w_K(x, u) = \mathbf{x}_K(x)' \mathbf{Q}_K^{-1} \mathbf{x}_K(u).$$

Take the case  $x_i \sim U[0, 1]$ . In Figure 21.4 we plot the weight function  $w_K(x, u)$  as a function of  $u$  for  $x = 0.5$  (panel (a)) and  $x = 0.25$  (panel (b)) for  $p = 4, 8, 12$  in panel (a) and  $p = 4, 12$  in panel (b). First, examine panel (a). Here you can see that the weight function  $w(x, u)$  is symmetric in  $u$  about  $x$ . For  $p = 4$  the weight function appears similar to a quadratic in  $u$ , and as  $p$  increases the weight function concentrates its main weight around  $x$ . However, the weight function is not non-negative. It is quite similar in shape to what are known as higher-order (or bias-reducing) kernels, which were not reviewed in the previous chapters but are part of the kernel estimation toolkit. Second, examine panel (b). Again the weight function is maximized at  $x$ , but now it is asymmetric in  $u$  about the point  $x$ . Still, the general features from panel (a) carry over to panel (b). Namely, as  $p$  increases the polynomial estimator puts most weight on observations for which  $x_i$  is close to  $x$  (just as for kernel regression), but is different from conventional kernel regression in that the weight function is not non-negative. Qualitatively similar plots are obtained for spline regression.

There is little formal theory (of which I am aware) which makes a formal link between series regression and kernel regression, so the comments presented here are illustrative<sup>3</sup>. However, the point is that statements of the form “Series regression is a global method; Kernel regression is a local method” may not be complete descriptions. Both are global in nature when  $h$  is large (for kernels) or  $K$  is small (series), and are local in nature when  $h$  is small (for kernels) or  $K$  is large (series).

## 21.8 Stone-Weierstrass and Jackson Approximation Theory

A good series approximation  $m_K(x)$  has the property that it gets close to the true CEF  $m(x)$  as the complexity  $K$  increases. Formal statements can be derived from the mathematical theory of the approximation of functions.

An elegant and famous theorem is the **Stone-Weierstrass Theorem**, (Weierstrass, 1885, Stone, 1948) which states that any continuous function can be uniformly well approximated by a polynomial of sufficiently high order. Specifically, the theorem states that if  $m(x)$  is continuous on a compact set  $S$ , then for any  $\varepsilon > 0$  there is some  $K$  sufficiently large such that

$$\inf_{\boldsymbol{\beta}} \sup_{x \in S} |m(x) - \mathbf{x}_K(x)' \boldsymbol{\beta}| \leq \varepsilon. \quad (21.7)$$

Thus the true unknown  $m(x)$  can be arbitrarily well approximated by selecting a suitable polynomial.

Jackson (1912) strengthened this result to give convergence rates which depend on the smoothness of  $m(x)$ . The basic result has also been extended to spline functions. The following notation will be useful. Define the  $\boldsymbol{\beta}$  which minimizes the left-side of (21.7) as

$$\boldsymbol{\beta}_K^* = \operatorname{argmin}_{\boldsymbol{\beta}} \sup_{x \in S} |m(x) - \mathbf{x}_K(x)' \boldsymbol{\beta}|, \quad (21.8)$$

define the approximation error

$$r_K^*(x) = m(x) - \mathbf{x}_K(x)' \boldsymbol{\beta}_K^* \quad (21.9)$$

and define the minimized value of (21.7)

$$\delta_K^* \stackrel{\text{def}}{=} \inf_{\boldsymbol{\beta}} \sup_{x \in S} |m(x) - \mathbf{x}_K(x)' \boldsymbol{\beta}| = \sup_{x \in S} |m(x) - \mathbf{x}_K(x)' \boldsymbol{\beta}_K^*| = \sup_{x \in S} |r_K^*(x)|. \quad (21.10)$$

**Theorem 21.1** If for some  $\alpha \geq 0$ ,  $m^{(\alpha)}(x)$  is uniformly continuous on a compact set  $S$ , and  $\mathbf{x}_K(x)$  is either a polynomial basis or a spline basis (with uniform knot spacing) of order  $s \geq \alpha$ , then as  $K \rightarrow \infty$

$$\delta_K^* \leq o(K^{-\alpha}). \quad (21.11)$$

Furthermore, if  $m^{(2)}(x)$  is uniformly continuous on  $S$  and  $\mathbf{x}_K(x)$  is a linear spline basis, then  $\delta_K^* \leq O(K^{-2})$ .

For a proof for the polynomial case, see Theorem 4.3 of Lorentz (1986), or Theorem 3.12 of Schumaker (2007) plus his equations (2.119) and (2.121). For the spline case see Theorem 6.27 of Schumaker (2007) plus his equations (2.119) and (2.121). For the linear spline case see Theorem 6.15 of Schumaker, equation (6.28).

<sup>3</sup>Similar connections are made in the appendix of Chen, Liao, and Sun (2012).

Theorem 21.1 is more useful than the classic Stone-Weierstrass Theorem, as it gives an approximation rate which depends on the smoothness order  $\alpha$ . The rate  $o(K^{-\alpha})$  in (21.11) means that the approximation error (21.10) decreases as  $K$  increases, and decreases at a faster rate when  $\alpha$  is large. The standard interpretation is that when  $m(x)$  is smoother it is possible to approximate it with a fewer number of series terms.

It will turn out that for our distributional theory results it will be sufficient to consider the case that  $m^{(2)}(x)$  is uniformly continuous. For this case, Theorem 21.1 shows that polynomials and quadratic/cubic splines achieve the rate  $o(K^{-2})$ , and linear splines achieve the rate  $O(K^{-2})$ . For most of our results the latter bound will be sufficient.

More generally, Theorem 21.1 makes a distinction between polynomials and splines, as polynomials achieve the rate  $o(K^{-\alpha})$  adaptively (without input from the user) while splines achieve the rate  $o(K^{-\alpha})$  only if the spline order  $s$  is appropriately chosen. This is an advantage for polynomials. However, as emphasized by Schumaker (2007), splines simultaneously approximate the derivatives  $m^{(q)}(x)$  for  $q < \alpha$ . Thus, for example, a quadratic spline simultaneously approximates the function  $m(x)$  and its first derivative  $m'(x)$ . There is no comparable result for polynomials. This is an advantage for quadratic and cubic splines. Since economists are often more interested in marginal effects (derivatives) than in levels, this may be a good reason to prefer such splines over polynomials.

Theorem 21.1 is a bound on the best uniform approximation error. The coefficient  $\boldsymbol{\beta}_K^*$  which minimizes (21.11) is not, however, the projection coefficient  $\boldsymbol{\beta}_K$  as defined in (21.3). Thus Theorem 21.1 does not directly inform us concerning the approximation error obtained by series regression. It turns out, however, that the projection error can be easily deduced from (21.11). It is useful to define the projection approximation error

$$r_K(x) = m(x) - \mathbf{x}_K(x)' \boldsymbol{\beta}_K. \quad (21.12)$$

This is similar to (21.9) but evaluated using the projection coefficient rather than the minimizing coefficient  $\boldsymbol{\beta}_K^*$  (21.8). Also define  $r_{Ki} = r_K(x_i)$ . Assuming that  $x_i$  has compact support  $S$ , the expected squared projection error is

$$\begin{aligned} \delta_K &\stackrel{\text{def}}{=} (\mathbb{E}(r_{Ki})^2)^{1/2} \\ &= \left( \int_S (m(x) - \mathbf{x}_K(x)' \boldsymbol{\beta}_K)^2 dF(x) \right)^{1/2} \\ &\leq \left( \int_S (m(x) - \mathbf{x}_K(x)' \boldsymbol{\beta}_K^*)^2 dF(x) \right)^{1/2} \\ &\leq \left( \int_S \delta_K^{*2} dF(x) \right)^{1/2} \\ &= \delta_K^*. \end{aligned} \quad (21.13)$$

The first inequality holds since the projection coefficient  $\boldsymbol{\beta}_K$  minimizes the expected squared projection error (see Section 2.25). The second inequality is the definition of  $\delta_K^*$ . Combined with Theorem 21.1 we have established the following result.

**Theorem 21.2** If  $x_i$  has compact support  $S$ , for some  $\alpha \geq 0$   $m^{(\alpha)}(x)$  is uniformly continuous on  $S$ , and  $\mathbf{x}_K(x)$  is either a polynomial basis or a spline basis of order  $s \geq \alpha$ , then as  $K \rightarrow \infty$

$$\delta_K \leq \delta_K^* \leq o(K^{-\alpha}).$$

Furthermore, if  $m^{(2)}(x)$  is uniformly continuous on  $S$  and  $\mathbf{x}_K(x)$  is a linear spline basis, then  $\delta_K \leq O(K^{-2})$ .

The available theory of the approximation of functions goes beyond the results described here. For example, there is a theory of weighted polynomial approximation (Mhaskar, 1996) which provides an analog of Theorem 21.2 for the unbounded real line when  $x_i$  has a density with exponential tails.

## 21.9 Regressor Bounds

The approximation result in Theorem 21.2 assumes that the regressors  $x_i$  have bounded support  $S$ . This is conventional in series regression theory, as it greatly simplifies the analysis. Bounded support implies that the regressor function  $\mathbf{x}_K(x)$  is bounded. Define

$$\zeta_K(x) = (\mathbf{x}_K(x)' \mathbf{Q}_K^{-1} \mathbf{x}_K(x))^{1/2} \quad (21.14)$$

$$\zeta_K = \sup_x \zeta_K(x) \quad (21.15)$$

where  $\mathbf{Q}_K = \mathbb{E}(\mathbf{x}_{Ki} \mathbf{x}'_{Ki})$  is the population design matrix given the regressors  $\mathbf{x}_{Ki}$ . This implies that for all observations

$$(\mathbf{x}'_{Ki} \mathbf{Q}_K^{-1} \mathbf{x}_{Ki})^{1/2} \leq \zeta_K. \quad (21.16)$$

The constant  $\zeta_K(x)$  is the normalized length of the regressor vector  $\mathbf{z}_K(x)$ . The constant  $\zeta_K$  is the maximum normalized length. Their values are determined by the basis function transformations and the distribution of  $x_i$ . They are invariant, however, to rescaling  $\mathbf{x}_{Ki}$  or linear rotations.

For polynomials and splines we have explicit expressions for the rate at which  $\zeta_K$  grows with  $K$ .

**Theorem 21.3** If  $x_i$  has compact support  $S$  with a strictly positive density  $f(x)$  on  $S$  then

1.  $\zeta_K \leq O(K)$  for polynomials
2.  $\zeta_K \leq O(K^{1/2})$  for splines

For a proof of Theorem 21.3 see Newey (1997, Theorem 4).

Furthermore, when  $x_i$  is uniformly distributed then we can explicitly calculate for polynomials that  $\zeta_K = K$ , so the polynomial bound  $\zeta_K \leq O(K)$  cannot be improved.

To illustrate, we plot in Figure 21.5 the values  $\zeta_K(x)$  for the case  $x_i \sim U[0, 1]$ . We plot  $\zeta_K(x)$  for a polynomial of degree  $p = 9$  and a quadratic spline with  $N = 7$  knots (both satisfy  $K = 10$ ). You can see that the values of  $\zeta_K(x)$  are close to 3 for both basis transformations and most values of  $x$ , but  $\zeta_K(x)$  increases sharply for  $x$  near the boundary. The maximum values are  $\zeta_K = 10$  for the polynomial and  $\zeta_K = 7.4$  for the quadratic spline. While Theorem 21.3 shows the two have different rates for large  $K$ , we see for moderate  $K$  that the differences are relatively minor.

## 21.10 Matrix Convergence

One of the challenges which arise when developing a theory for the least squares estimator is how to describe the large-sample behavior of the sample design matrix

$$\hat{\mathbf{Q}}_K = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_{Ki} \mathbf{x}'_{Ki}$$

as  $K \rightarrow \infty$ . The trouble is that the dimension of  $\hat{\mathbf{Q}}_K$  is increasing with  $K$ , so we cannot apply a standard WLLN.

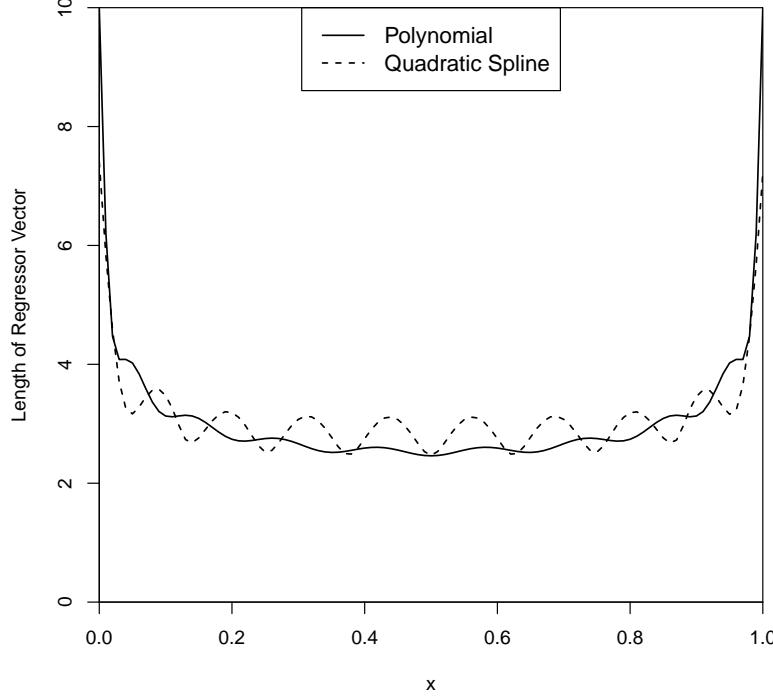


Figure 21.5: Normalized Regressor Lengths  $\zeta_K(x)$ ,  $K = 10$

It turns out to be convenient for the theory if we first rotate the regressor vector so that the elements are orthogonal in expectation. Thus we define the standardized regressors and design matrix as

$$\tilde{\mathbf{x}}_{Ki} = \mathbf{Q}_K^{-1/2} \mathbf{x}_{Ki} \quad (21.17)$$

$$\tilde{\mathbf{Q}}_K = \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{x}}_{Ki} \tilde{\mathbf{x}}'_{Ki}.$$

Note that  $\mathbb{E}(\tilde{\mathbf{x}}_{Ki} \tilde{\mathbf{x}}'_{Ki}) = \mathbf{I}_K$ . The standardized regressors are not used in practice; they are introduced only to simplify the theoretical derivations.

Our convergence theory will require the following fundamental rate bound on the number of coefficients  $K$ .

**Assumption 21.1**

1.  $\lambda_{\min}(\mathbf{Q}_K) \geq \underline{\lambda} > 0$
2.  $\zeta_K^2 \log(K)/n \rightarrow 0$  as  $n, K \rightarrow \infty$

Assumption 21.1.1 ensures that the transformation (21.17) is well defined<sup>4</sup>. Assumption 21.1.2 states that the squared maximum regressor length  $\zeta_K^2$  grows slower than  $n$ . Since  $\zeta_K$  increases with  $K$  this is a bound on the rate at which  $K$  can increase with  $n$ . By Theorem 21.2, the rate in Assumption 21.1.2 holds

---

<sup>4</sup>Technically, what is required is that  $\lambda_{\min}(\mathbf{B}_K \mathbf{Q}_K \mathbf{B}'_K) \geq \underline{\lambda} > 0$  for some  $K \times K$  sequence of matrices  $\mathbf{B}_K$ , or equivalently that Assumption 21.1.1 holds after replacing  $\mathbf{x}_{Ki}$  with  $\mathbf{B}_K \mathbf{x}_{Ki}$ .

for polynomials if  $K^2 \log(K)/n \rightarrow 0$  and for splines if  $K \log(K)/n \rightarrow 0$ . In either case, this means that the number of coefficients  $K$  is growing at a rate slower than  $n$ .

We are now in a position to describe a convergence result for the standardized design matrix. The following is Lemma 6.2 of Belloni, Chernozhukov, Chetverikov, and Kato (2015).

**Theorem 21.4** If Assumption 21.1 holds then

$$\|\tilde{\mathbf{Q}}_K - \mathbf{I}_K\|_2 \xrightarrow{p} 0. \quad (21.18)$$

A simplified proof of Theorem 21.4 can be found in Section 21.31.

The norm in (21.18) is the **spectral norm**

$$\|\mathbf{A}\|_2 = (\lambda_{\max}(\mathbf{A}'\mathbf{A}))^{1/2}$$

where  $\lambda_{\max}(\mathbf{B})$  denotes the largest eigenvalue of the matrix  $\mathbf{B}$ . For a full description see Section A.23. It is a useful norm for matrices which are growing in dimension.

For the least-squares estimator what is particularly important is the inverse of the sample design matrix. Fortunately we can easily deduce consistency of its inverse from (21.18) when the regressors have been orthogonalized as described.

**Theorem 21.5** If Assumption 21.1 holds then

$$\|\tilde{\mathbf{Q}}_K^{-1} - \mathbf{I}_K\|_2 \xrightarrow{p} 0 \quad (21.19)$$

and

$$\lambda_{\max}(\tilde{\mathbf{Q}}_K^{-1}) = 1/\lambda_{\min}(\tilde{\mathbf{Q}}_K) \xrightarrow{p} 1. \quad (21.20)$$

The proof of Theorem 21.5 can be found in Section 21.31.

## 21.11 Consistent Estimation

In this section we give conditions for consistent estimation of  $m(x)$  by the series estimator  $\hat{m}_K(x) = \mathbf{x}_K(x)' \hat{\boldsymbol{\beta}}_K$ .

What we know from standard regression theory is that for any fixed  $K$ ,  $\hat{\boldsymbol{\beta}}_K \xrightarrow{p} \boldsymbol{\beta}_K$  and thus  $\hat{m}_K(x) = \mathbf{x}_K(x)' \hat{\boldsymbol{\beta}}_K \xrightarrow{p} \mathbf{x}_K(x)' \boldsymbol{\beta}_K$  as  $n \rightarrow \infty$ . Furthermore, from the Stone-Weierstrass Theorem we know that  $\mathbf{x}_K(x)' \boldsymbol{\beta}_K \rightarrow m(x)$  as  $K \rightarrow \infty$ . It therefore seems reasonable to expect that  $\hat{m}_K(x) \xrightarrow{p} m(x)$  as both  $n \rightarrow \infty$  and  $K \rightarrow \infty$  together. Making this argument rigorous, however, is technically challenging, in part because the dimensions of  $\hat{\boldsymbol{\beta}}_K$  and its components are changing with  $K$ .

Since  $\hat{m}_K(x)$  and  $m(x)$  are functions, convergence should be defined with respect to an appropriate metric. For kernel regression we focused on pointwise convergence (for each value of  $x$  separately) as that is the simplest to analyze in that context. For series regression it turns out to be simplest to describe convergence with respect to integrated squared error (ISE). We define the latter as

$$\text{ISE}(K) = \int (\hat{m}_K(x) - m(x))^2 dF(x)$$

where  $F$  is the marginal distribution of  $x_i$ .  $\text{ISE}(K)$  is the average squared distance between  $\hat{m}_K(x)$  and  $m(x)$ , weighted by the marginal distribution of  $x_i$ . The  $\text{ISE}$  is random, depends on both sample size  $n$  and model complexity  $K$ , and its distribution is determined by the joint distribution of the observations  $(y_i, x_i)$ .

We can establish the following.

**Theorem 21.6** Under Assumption 21.1 and  $\delta_K = o(1)$ , then as  $n, K \rightarrow \infty$ ,

$$\text{ISE}(K) = o_p(1). \quad (21.21)$$

The proof of Theorem 21.6 can be found in Section 21.31.

Theorem 21.6 shows that the series estimator  $\hat{m}_K(x)$  is consistent in the  $\text{ISE}$  norm under very mild conditions. The assumption  $\delta_K = o(1)$  holds for polynomials and splines if  $K \rightarrow \infty$  and  $m(x)$  is uniformly continuous. This result is analogous to Theorem 20.7 which showed that kernel regression estimator is consistent if  $m(x)$  is continuous.

## 21.12 Convergence Rate

Theorem 21.6 showed that the series regression estimator is consistent in the  $\text{ISE}$  norm. We now give a rate of convergence.

**Theorem 21.7** Under Assumption 21.1 and  $\sigma^2(x) \leq \bar{\sigma}^2 < \infty$ , then as  $n, K \rightarrow \infty$ ,

$$\text{ISE}(K) \leq O_p\left(\delta_K^2 + \frac{K}{n}\right). \quad (21.22)$$

Furthermore, if  $m^{(2)}(x)$  is uniformly continuous then for polynomial or spline basis functions

$$\text{ISE}(K) \leq O_p\left(K^{-4} + \frac{K}{n}\right). \quad (21.23)$$

The proof of Theorem 21.7 can be found in Section 21.31. It is based on Newey (1997).

The bound (21.23) is particularly useful as it gives an explicit rate in terms of  $K$  and  $n$ . The result shows that the integrated squared error is bounded in probability by two terms. The first  $K^{-4}$  is the squared bias. The second  $K/n$  is the estimation variance. This is analogous to the AIMSE for kernel regression (20.5). We can see that increasing the number of series terms  $K$  affects the integrated squared error by decreasing the bias but increasing the variance. The fact that the estimation variance is of order  $K/n$  can be intuitively explained by the fact that the regression model is estimating  $K$  coefficients.

If desired, the bound (21.23) can be written as  $o_p(K^{-4}) + O_p(K/n)$  for polynomials and quadratic splines.

We are interested in the sequence  $K$  which minimizes the trade-off in (21.23). By examining the first-order condition, we find that the sequence which minimizes this bound is  $K \sim n^{1/5}$ . With this choice we obtain the optimal integrated squared error  $\text{ISE}(K) \leq O_p(n^{-4/5})$ . This is the same convergence rate as obtained by kernel regression under similar assumptions.

It is interesting to contrast the optimal rate  $K \sim n^{1/5}$  for series regression with  $h \sim n^{-1/5}$  for kernel regression. Essentially, one can view the rate  $K^{-1}$  in series regression as a “bandwidth” similar to kernel regression, or one can view the rate  $1/h$  in kernel regression as the effective number of coefficients.

The rate  $K \sim n^{1/5}$  means that the optimal  $K$  increases very slowly with the sample size. For example, doubling your sample size implies only a 15% increase in the optimal number of coefficients  $K$ . To obtain a doubling in the optimal number of coefficients, you need to multiply the sample size by 32.

To illustrate, Figure 21.6 displays the ISE rate bounds  $K^{-4} + K/n$  as a function of  $K$  for  $n = 10, 30, 150$ . The filled circles mark the ISE-minimizing  $K$ , which are  $K = 2, 3$ , and  $4$  for the three functions. Notice that the ISE functions are steeply downward sloping for small  $K$ , and nearly flat for large  $K$  (when  $n$  is large). This is because the bias term  $K^{-4}$  dominates for small values of  $K$  while the variance term  $K/n$  dominates for large values of  $K$ , and the latter flattens as  $n$  increases.

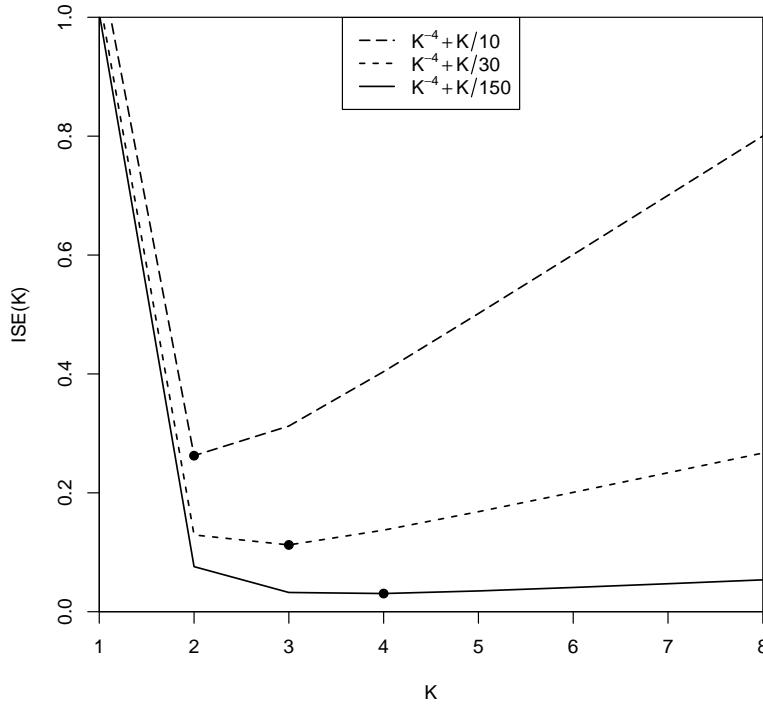


Figure 21.6: Integrated Squared Error

## 21.13 Asymptotic Normality

The theory we present in this section will apply to any linear function of the regression function. That is, we consider parameters of interest which can be written as a real-valued linear function of the regression function:

$$\theta = a(m).$$

This includes the regression function  $m(x)$  at a given point  $x$ , derivatives of  $m(x)$ , and integrals over  $m(x)$ . Given  $\hat{m}_K(x) = \mathbf{x}_K(x)' \hat{\beta}_K$  as an estimator for  $m(x)$ , the estimator for  $\theta$  is

$$\hat{\theta}_K = a(\hat{m}_K) = \mathbf{a}'_K \hat{\beta}_K$$

for some  $K \times 1$  vector of constants  $\mathbf{a}_K \neq \mathbf{0}$ . (The relationship  $a(\hat{m}_K) = \mathbf{a}'_K \hat{\beta}_K$  follows since  $a$  is linear in  $m$  and  $\hat{m}_K$  is linear in  $\hat{\beta}_K$ .)

If  $K$  were fixed as  $n \rightarrow \infty$ , then by standard asymptotic theory we would expect  $\hat{\theta}_K$  to be asymptotically normal with variance

$$V_K = \mathbf{a}'_K \mathbf{Q}_K^{-1} \boldsymbol{\Omega}_K \mathbf{Q}_K^{-1} \mathbf{a}_K$$

where  $\Omega_K = \mathbb{E}(\mathbf{x}_{Ki}\mathbf{x}'_{Ki}e_i^2)$ . The standard justification, however, is not valid in the nonparametric case. This is in part because  $V_K$  may diverge as  $K \rightarrow \infty$ , and in part due to the finite sample bias due to the approximation error. Therefore a new theory is required. Interestingly, it turns out that in the nonparametric case  $\hat{\theta}_K$  is still asymptotically normal, and  $V_K$  is still the appropriate variance for  $\hat{\theta}_K$ . The proof is different than the parametric case as the dimensions of the matrices are increasing with  $K$ , and we need to be attentive to the estimator's bias due to the series approximation.

**Assumption 21.2** In addition to Assumption 21.1

1.  $\lim_{B \rightarrow \infty} \sup_x \mathbb{E}(e_i^2 \mathbf{1}(e_i^2 > B) | x_i = x) = 0$
2.  $\mathbb{E}(e_i^2 | x_i) \geq \underline{\sigma}^2 > 0$
3.  $\zeta_K \delta_K = o(1)$  as  $K \rightarrow \infty$

Assumption 21.2.1 is conditional square integrability. It implies that the conditional variance  $\mathbb{E}(e_i^2 | x_i)$  is bounded. It is used to verify the Lindeberg condition for the CLT.

Assumption 21.2.2 states that the conditional variance is nowhere degenerate. Thus there is no  $x_i$  for which  $y_i$  is perfectly predictable. This is a technical condition used to bound  $V_K$  from below.

Assumption 21.2.3 states that approximation error  $\delta_K$  declines faster than the maximal regressor length  $\zeta_K$ . For polynomials a sufficient condition for this assumption is that  $m^{(2)}(x)$  is uniformly continuous. For splines a sufficient condition is that  $m^{(1)}(x)$  is uniformly continuous.

**Theorem 21.8** Under Assumption 21.2, as  $n \rightarrow \infty$ ,

$$\frac{\sqrt{n}(\hat{\theta}_K - \theta + a(r_K))}{V_K^{1/2}} \xrightarrow{d} N(0, 1). \quad (21.24)$$

The proof of Theorem 21.8 can be found in Section 21.31.

Theorem 21.8 shows that the estimator  $\hat{\theta}_K$  is approximately normal with bias  $-a(r_K)$  and variance  $V_K/n$ . The variance is the same as in the parametric case, but the asymptotic distribution contains an asymptotic bias, similar as is found in kernel regression.

One useful message from Theorem 21.8 is that the classical variance formula  $V_K$  for  $\hat{\theta}_K$  still applies for series regression. This motivates using conventional estimators for  $V_K$ , as will be discussed in Section 21.18.

Theorem 21.8 shows that the estimator  $\hat{\theta}_K$  has a bias term  $a(r_K)$ . What is this? It is the same transformation of the function  $r_K(x)$  as  $\theta = a(m)$  is of the regression function  $m(x)$ . For example, if  $\theta = m(x)$  is the regression at a fixed point  $x$ , then  $a(r_K) = r_K(x)$ , the approximation error at the same point. If  $\theta = \frac{d}{dx}m(x)$  is the regression derivative, then  $a(r_K) = \frac{d}{dx}r_K(x)$  is the derivative of the approximation error.

This means that the bias in the estimator  $\hat{\theta}_K$  for  $\theta$  shown in Theorem 21.8 is simply the approximation error transformed by the functional of interest. If we are estimating the regression function then the bias is the error in approximating the regression function; if we are estimating the regression derivative then the bias is the error in the derivative in the approximation error for the regression function.

## 21.14 Regression Estimation

A special yet important example of a linear estimator is the regression function at a fixed point  $x$ . In the notation of the previous section,  $a(m) = m(x)$  and  $\mathbf{a}_K = \mathbf{x}_K(x)$ . The series estimator of  $m(x)$  is  $\hat{\theta}_K = \hat{m}_K(x) = \mathbf{x}_K(x)' \hat{\beta}_K$ . As this is a key problem of interest, we restate the asymptotic result of Theorems 21.8 for this estimator.

**Theorem 21.9** Under Assumption 21.2, as  $n \rightarrow \infty$ ,

$$\frac{\sqrt{n}(\hat{m}_K(x) - m(x) + r_K(x))}{V_K^{1/2}(x)} \xrightarrow{d} N(0, 1) \quad (21.25)$$

where

$$V_K(x) = \mathbf{x}_K(x)' \mathbf{Q}_K^{-1} \boldsymbol{\Omega}_K \mathbf{Q}_K^{-1} \mathbf{x}_K(x).$$

There are two important features about the asymptotic distribution (21.25).

First, as mentioned in the previous section, it shows that the classical variance formula  $V_K(x)$  applies for the series estimator  $\hat{m}_K(x)$ . Second, (21.25) shows that the estimator has the asymptotic bias  $r_K(x)$ . This is due to the fact that the finite order series is an approximation to the unknown regression function  $m(x)$ , and this results in finite sample bias.

## 21.15 Undersmoothing

An unpleasant aspect about Theorem 21.9 is the bias term. An interesting trick is that this bias term can be made asymptotically negligible if we assume that  $K$  increases with  $n$  at a sufficiently fast rate.

**Theorem 21.10** Under Assumption 21.2, if in addition  $n\delta_K^{*2} \rightarrow 0$  then

$$\frac{\sqrt{n}(\hat{m}_K(x) - m(x))}{V_K^{1/2}(x)} \xrightarrow{d} N(0, 1). \quad (21.26)$$

The condition  $n\delta_K^{*2} \rightarrow 0$  implies that the squared bias converges faster than the estimation variance, so the former is asymptotically negligible. If  $m^{(2)}(x)$  is uniformly continuous, then a sufficient condition for polynomials and quadratic splines is that  $K \sim n^{1/4}$ . For linear splines a sufficient condition is for  $K$  to diverge faster than  $K^{1/4}$ . The rate  $K \sim n^{1/4}$  is somewhat faster than the ISE-optimal rate  $K \sim n^{1/5}$ .

The assumption  $n\delta_K^{*2} \rightarrow 0$  is often stated by authors as an innocuous technical condition. This is misleading as it is a technical trick and should be discussed explicitly. The reason why the assumption eliminates the bias from (21.26) is that the assumption forces the estimation variance to dominate the squared bias so that the latter can be ignored. This means that the estimator itself is inefficient.

Because  $n\delta_K^{*2} \rightarrow 0$  means that  $K$  is larger than optimal, we say that  $\hat{m}_K(x)$  is **undersmoothed** relative to the optimal series estimator.

Many authors like to focus their asymptotic theory on the assumptions in Theorem 21.10 as the distribution (21.26) appears cleaner. However, it is a poor use of asymptotic theory. There are three problems with the assumption  $n\delta_K^{*2} \rightarrow 0$  and the approximation (21.26). First, the estimator  $\hat{m}_K(x)$  is inefficient. Second, while the assumption  $n\delta_K^{*2} \rightarrow 0$  makes the bias of lower order than the variance, it only makes the bias of slightly lower order, meaning that the accuracy of the asymptotic approximation

is poor. Effectively, the estimator is still biased in finite samples. Third,  $n\delta_K^{*2} \rightarrow 0$  is an assumption, not a rule for empirical practice. It is unclear what the statement “Assume  $n\delta_K^{*2} \rightarrow 0$ ” means in a practical application. From this viewpoint the difference between (21.24) and (21.26) is in the assumptions, not in the actual reality nor in the actual empirical practice. Eliminating a nuisance (the asymptotic bias) through an assumption is a trick, not a substantive use of theory. My strong view is that the result (21.24) is more informative than (21.26). It shows that the asymptotic distribution is normal but has a non-trivial finite sample bias.

## 21.16 Residuals and Regression Fit

The fitted regression at  $x = x_i$  is  $\hat{m}_K(x_i) = \mathbf{x}'_{Ki} \hat{\boldsymbol{\beta}}_K$  and the fitted residual is

$$\hat{e}_{Ki} = y_i - \hat{m}_K(x_i).$$

The leave-one-out prediction errors are

$$\begin{aligned}\tilde{e}_{Ki} &= y_i - \hat{m}_{K,-i}(x_i) \\ &= y_i - \mathbf{x}'_{Ki} \hat{\boldsymbol{\beta}}_{K,-i}\end{aligned}$$

where  $\hat{\boldsymbol{\beta}}_{K,-i}$  is the least-squares coefficient with the  $i^{\text{th}}$  observation omitted. Using (3.45) we have the simple computational formula

$$\tilde{e}_{Ki} = \hat{e}_{Ki} (1 - \mathbf{x}'_{Ki} (\mathbf{X}'_K \mathbf{X}_K)^{-1} \mathbf{x}_{Ki})^{-1}. \quad (21.27)$$

As for kernel regression, the prediction errors  $\tilde{e}_{Ki}$  are better estimators of the errors than the fitted residuals  $\hat{e}_{Ki}$ , as the former do not have the tendency to over-fit when the number of series terms is large.

## 21.17 Cross-Validation Model Selection

A common method for selection of the number of series terms  $K$  is cross-validation. The cross-validation criterion is sum<sup>5</sup> of squared prediction errors

$$\text{CV}(K) = \sum_{i=1}^n \tilde{e}_{Ki}^2 = \sum_{i=1}^n \hat{e}_{Ki}^2 (1 - \mathbf{x}'_{Ki} (\mathbf{X}'_K \mathbf{X}_K)^{-1} \mathbf{x}_{Ki})^{-2}. \quad (21.28)$$

The CV-selected value of  $K$  is the integer which minimizes  $\text{CV}(K)$ .

As shown in Theorem 20.6,  $\text{CV}(K)$  is an approximately unbiased estimator of the integrated mean-squared error IMSE, which is the expected integrated squared error (ISE). The proof of the result is the same for all nonparametric estimators (series as well as kernels) so does not need to be repeated here. Therefore, finding the  $K$  which produces the smallest value of  $\text{CV}(K)$  is a good indicator that the estimator  $\hat{m}_K(x)$  has small IMSE.

For practical implementation we first designate a set of models (sets of basis transformations and number of variables  $K$ ) over which to search. (For example, polynomials of order 1 through  $K_{\max}$  for some pre-selected  $K_{\max}$ .) For each, there is a set of regressors  $\mathbf{x}_{Ki}$  which are obtained by transformations of the original variables  $x_i$ . For each set, we estimate the regression by least-squares, calculate the leave-one-out prediction errors and the CV criterion. Since the errors are a linear operation this is a simple calculation. The CV-selected  $K$  is the integer which produces the smallest value of  $\text{CV}(K)$ . Plots of  $\text{CV}(K)$  against  $K$  can aid assessment and interpretation. Since the model order  $K$  is an integer, the CV criterion for series regression is a discrete function, unlike the case of kernel regression.

If it is desired to produce an estimator  $\hat{m}_K(x)$  with reduced bias it may be preferred to select a value of  $K$  slightly higher than that selected by CV alone.

---

<sup>5</sup>Some authors define  $\text{CV}(K)$  as the average squared prediction errors.

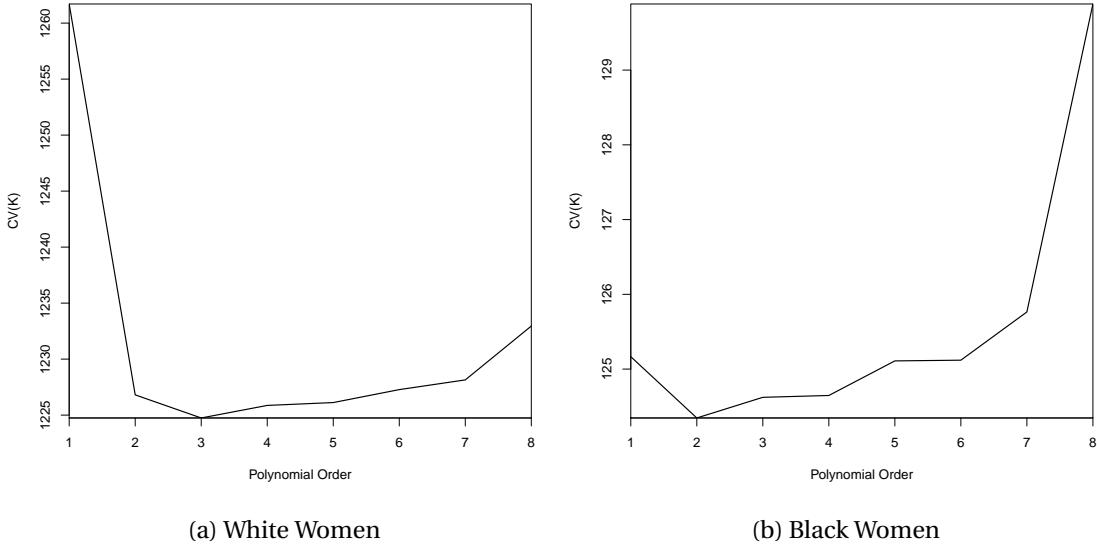


Figure 21.7: Cross-Validation Functions for Polynomial Estimates of Experience Profile, College-Educated Women

To illustrate, in Figure 21.7 we plot the cross-validation functions for the polynomial regression estimates from Figure 21.1. The lowest point marks the polynomial order which minimizes the cross-validation function. In panel (a) we plot the CV function for the sub-sample of white women. Here we see that the CV-selected order is  $p = 3$ , a cubic polynomial. In panel (b) we plot the CV function for the sub-sample of black women, and find that the CV-selected order is  $p = 2$ , a quadratic. As expected from visual examination of Figure 21.1, the selected model is more parsimonious for panel (b), most likely because it has a substantially smaller sample size. What may be surprising is that even for panel (a), which has a large sample and smooth estimates, the CV-selected model is still relatively parsimonious.

A user who desires a reduced bias estimator might increase the polynomial orders to  $p = 4$  or even  $p = 5$  for the subsample of white women, and to  $p = 3$  for the subsample of black women. Both CV functions are relatively similar across these values.

## 21.18 Variance and Standard Error Estimation

The exact conditional variance of the least squares estimator  $\hat{\beta}_K$  under independent sampling is

$$V_{\hat{\beta}} = (\mathbf{X}'_K \mathbf{X}_K)^{-1} \left( \sum_{i=1}^n \mathbf{x}_{Ki} \mathbf{x}'_{Ki} \sigma^2(x) \right) (\mathbf{X}'_K \mathbf{X}_K)^{-1}. \quad (21.29)$$

The exact conditional variance for the conditional mean estimator  $\hat{m}_K(x) = \mathbf{x}_K(x)' \hat{\boldsymbol{\beta}}_K$  is

$$V_K(x) = \mathbf{x}_K(x)' (\mathbf{X}'_K \mathbf{X}_K)^{-1} \left( \sum_{i=1}^n \mathbf{x}_{Ki} \mathbf{x}'_{Ki} \sigma^2(x) \right) (\mathbf{X}'_K \mathbf{X}_K)^{-1} \mathbf{x}_K(x).$$

Using the notation of Section 21.7 this equals

$$\frac{1}{n^2} \sum_{i=1}^n \hat{w}_K(x, x_i)^2 \sigma^2(x).$$

In the case of conditional homoskedasticity the latter simplifies to

$$\frac{1}{n} \widehat{w}_K(x, x) \sigma^2 \simeq \frac{1}{n} \zeta_K(x)^2 \sigma^2.$$

where  $\zeta_K(x)$  is the normalized regressor length defined in (21.14). Under conditional heteroskedasticity but large samples with  $K$  large (so that  $\widehat{w}_K(x, x_i)$  is a local kernel) it approximately equals

$$\frac{1}{n} w_K(x, x) \sigma^2(x) = \frac{1}{n} \zeta_K(x)^2 \sigma^2(x).$$

In either case, we find that the variance is approximately

$$V_K(x) \approx \frac{1}{n} \zeta_K(x)^2 \sigma^2(x).$$

This shows that the variance of the series regression estimator is a scale of  $\zeta_K(x)^2$  and the conditional variance. From the plot of  $\zeta_K(x)$  shown in Figure 21.5 we can deduce that the series regression estimator will be relatively imprecise at the boundary of the support of  $x_i$ .

The estimator of (21.29) recommended by Andrews (1991a) is

$$\widehat{\mathbf{V}}_{\widehat{\boldsymbol{\beta}}} = (\mathbf{X}'_K \mathbf{X}_K)^{-1} \left( \sum_{i=1}^n \mathbf{x}_{Ki} \mathbf{x}'_{Ki} \tilde{e}_{Ki}^2 \right) (\mathbf{X}'_K \mathbf{X}_K)^{-1} \quad (21.30)$$

where  $\tilde{e}_{Ki}$  is the leave-one-out prediction error (21.27). This is the HC3 estimator. An alternative is to replace  $\tilde{e}_{Ki}$  with the least-squares residuals  $\hat{e}_{Ki}$  and then multiply by a degree-of-freedom adjustment, which is the HC1 covariance estimator. These estimators are the same as used in parametric regression.

Given (21.30), a variance estimator for the conditional mean estimator  $\widehat{m}_K(x) = \mathbf{x}_K(x)' \widehat{\boldsymbol{\beta}}_K$  is

$$\widehat{V}_K(x) = \mathbf{x}_K(x)' (\mathbf{X}'_K \mathbf{X}_K)^{-1} \left( \sum_{i=1}^n \mathbf{x}_{Ki} \mathbf{x}'_{Ki} \tilde{e}_{Ki}^2 \right) (\mathbf{X}'_K \mathbf{X}_K)^{-1} \mathbf{x}_K(x). \quad (21.31)$$

A standard error for  $\widehat{m}(x)$  is its square root.

## 21.19 Clustered Observations

Clustered observations take the form  $(y_{ig}, x_{ig})$  for individuals  $i = 1, \dots, n_g$  in cluster  $g = 1, \dots, G$ . The model is

$$\begin{aligned} y_{ig} &= m(x_{ig}) + e_{ig} \\ \mathbb{E}(e_{ig} | \mathbf{X}_g) &= 0 \end{aligned}$$

where  $\mathbf{X}_g$  is the stacked  $x_{ig}$ . Stack  $y_{ig}$  and  $e_{ig}$  into cluster-level variables  $\mathbf{y}_g$  and  $\mathbf{e}_g$ .

The series regression model using cluster-level notation is

$$\mathbf{y}_g = \mathbf{X}_g \boldsymbol{\beta}_K + \mathbf{e}_K g.$$

We can write the series estimator as

$$\widehat{\boldsymbol{\beta}}_K = \left( \sum_{g=1}^G \mathbf{X}'_g \mathbf{X}_g \right)^{-1} \left( \sum_{g=1}^G \mathbf{X}'_g \mathbf{y}_g \right).$$

The cluster-level residual vector is  $\widehat{\mathbf{e}}_g = \mathbf{y}_g - \mathbf{X}_g \widehat{\boldsymbol{\beta}}_K$ .

As for parametric regression with clustered observations, the standard assumption is that the clusters are mutually independent, but dependence within each cluster is unstructured. We therefore use the same variance formulae as used for parametric regression. The standard estimator is

$$\widehat{\mathbf{V}}_{\widehat{\boldsymbol{\beta}}}^{\text{CR1}} = \left( \frac{G}{G-1} \right) (\mathbf{X}'_K \mathbf{X}_K)^{-1} \left( \sum_{g=1}^G \mathbf{X}'_g \widehat{\mathbf{e}}_g \widehat{\mathbf{e}}'_g \mathbf{X}_g \right) (\mathbf{X}'_K \mathbf{X}_K)^{-1}.$$

An alternative is to use the delete-cluster prediction error

$$\begin{aligned}\tilde{\mathbf{e}}_g &= \mathbf{y}_g - \mathbf{X}_g \tilde{\boldsymbol{\beta}}_{K,-g} \\ \tilde{\boldsymbol{\beta}}_{K,-g} &= \left( \sum_{j \neq g} \mathbf{X}'_j \mathbf{X}_j \right)^{-1} \left( \sum_{j \neq g} \mathbf{X}'_j \mathbf{y}_j \right)\end{aligned}$$

leading to the estimator

$$\hat{V}_{\boldsymbol{\beta}}^{\text{CR3}} = (\mathbf{X}'_K \mathbf{X}_K)^{-1} \left( \sum_{g=1}^G \mathbf{X}'_g \tilde{\mathbf{e}}_g \tilde{\mathbf{e}}'_g \mathbf{X}_g \right) (\mathbf{X}'_K \mathbf{X}_K)^{-1}.$$

There is no current theory on how to select the number of series terms  $K$  for clustered observations. A reasonable choice is the delete-cluster cross-validation criterion, which is

$$CV(K) = \sum_{g=1}^G \tilde{\mathbf{e}}'_g \tilde{\mathbf{e}}_g.$$

The delete-cluster choice for  $K$  is the value which minimizes  $CV(K)$ .

## 21.20 Confidence Bands

When displaying nonparametric estimators such as  $\hat{m}_K(x)$  it is customary to display confidence intervals. An asymptotic pointwise 95% confidence interval for  $m(x)$  is

$$\hat{m}_K(x) \pm 1.96 \hat{V}_K^{1/2}(x).$$

These confidence intervals can be plotted along with  $\hat{m}_K(x)$ .

To illustrate, Figure 21.8 plots polynomial estimates of the regression of log(wage) on experience using the selected estimates from Figure 21.1, plus 95% confidence bands. Panel (a) plots the estimate for the subsample of white women using  $p = 5$ . Panel (b) plots the estimate for the subsample of black women using  $p = 3$ . The standard errors are calculated using the formula (21.31). You can see that the confidence bands widen at the boundaries. The confidence bands are tight for the larger subsample of white women, and significantly wider for the smaller subsample of black women. Regardless, both plots indicate that the average wage rises for experience levels up to about 20 years, and then flattens for experience levels above 20 years.

There are two deficiencies with these confidence bands. First, they do not take into account the bias  $r_K(x)$  of the series estimator. Consequently, we should interpret the confidence bounds as valid for the pseudo-true regression (the best finite  $K$  approximation) rather than the true regression function  $m(x)$ . Second, the above confidence intervals are based on a pointwise (in  $x$ ) asymptotic distribution theory. Consequently we should interpret there coverage as having pointwise validity, and be cautious about interpreting global shapes from the confidence bands.

## 21.21 Uniform Approximations

Since  $\hat{m}_K(x)$  is a function it is desirable to have a distribution theory which applies to the entire function, not just the estimator at a point. This can be used, for example, to construct confidence bands with uniform (in  $x$ ) coverage properties.

For those familiar with empirical process theory, it might be hoped that the stochastic process

$$\eta_K(x) = \frac{\sqrt{n}(\hat{m}_K(x) - m(x))}{V_K^{1/2}(x)}$$

might converge to a stochastic (Gaussian) process, but this is not the case. Effectively, the process  $\eta_K(x)$  is not asymptotically stochastically equicontinuous so conventional empirical process theory does not apply.

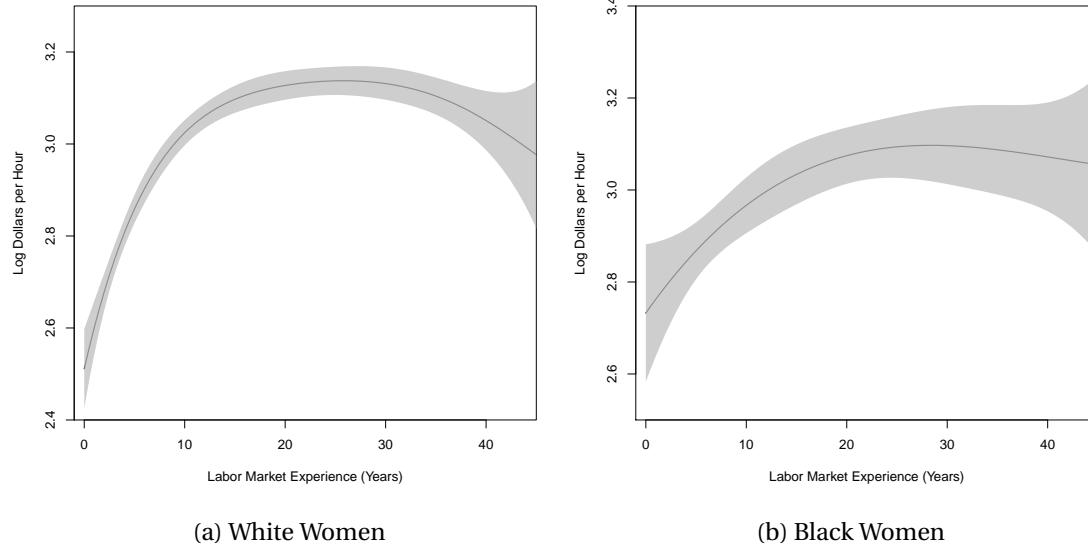


Figure 21.8: Polynomial Estimates with 95% Confidence Bands, College-Educated Women

To develop a uniform theory, Belloni, Chernozhukov, Chetverikov, and Kato (2015) have introduced what are known as strong approximations. Their method shows that  $\eta_K(x)$  is equal in distribution to a sequence of Gaussian processes plus a negligible error. Their theory (Theorem 4.4) takes the following form. Under stronger conditions than Assumption 21.2

$$\eta_K(x) =_d \frac{\mathbf{x}_K(x)' (\mathbf{Q}_K^{-1} \boldsymbol{\Omega}_K \mathbf{Q}_K^{-1})^{1/2}}{V_K^{1/2}(x)} G_K + o_p(1)$$

uniformly in  $x$ , where “ $=_d$ ” means “equality in distribution and  $G_K \sim N(0, \mathbf{I}_K)$ .

This shows the distributional result in Theorem 21.10 can be interpreted as holding uniformly in  $x$ . It can also be used to develop confidence bands (different from those from the previous section) with asymptotic uniform coverage.

## 21.22 Partially Linear Model

A common use of a series regression is to allow  $m(x)$  to be nonparametric with respect to one variable, yet linear in the other variables. This allows flexibility in a particular variable of interest. A partially linear model with vector-valued regressor  $\mathbf{x}_1$  and real-valued continuous  $x_2$  takes the form

$$m(\mathbf{x}_1, x_2) = \mathbf{x}_1' \boldsymbol{\beta}_1 + m_2(x_2).$$

This model is commonly used when  $\mathbf{x}_1$  are discrete (e.g. binary variables) and  $x_2$  is continuously distributed.

Series methods are particularly convenient for estimation of partially linear models, as we can replace the unknown function  $m_2(x_2)$  with a series expansion to obtain

$$\begin{aligned} m(\mathbf{x}) &\simeq m_K(\mathbf{x}) \\ &= \mathbf{x}_1' \boldsymbol{\beta}_1 + \mathbf{x}_{2K}'(\mathbf{x})' \boldsymbol{\beta}_{2K} \\ &= \mathbf{x}_K' \boldsymbol{\beta}_K \end{aligned}$$

where  $\mathbf{x}_{2K} = \mathbf{x}_{2K}(x_2)$  are the basis transformations of  $x_2$  (typically polynomials or splines) and  $\boldsymbol{\beta}_{2K}$  are coefficients. After transformation the regressors are  $\mathbf{x}_K = (\mathbf{x}_1', \mathbf{x}_{2K}')$ , and the coefficients are  $\boldsymbol{\beta}_K = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_{2K}')'$ .

## 21.23 Panel Fixed Effects

The one-way error components nonparametric regression model is

$$y_{it} = m(x_{it}) + u_i + \varepsilon_{it}$$

for  $i = 1, \dots, N$  and  $t = 1, \dots, T$ . It is standard to treat the individual effect  $u_i$  as a fixed effect. This model can be interpreted as a special case of the partially linear model from the previous section, though the dimension of  $u_i$  is increasing with  $N$ .

A series estimator approximates the function  $m(x)$  with  $m_K(x) = \mathbf{x}_K(x)' \boldsymbol{\beta}_K$  as in (21.4). This leads to the series regression model

$$y_{it} = \mathbf{x}'_{Kit} \boldsymbol{\beta}_K + u_i + \varepsilon_{Kit}$$

where  $\mathbf{x}_{Kit} = \mathbf{x}_K(x_{it})$ .

The fixed effects estimator is the same as in linear panel data regression. First, the within transformation is applied to  $y_{it}$  and the elements of the basis transformations  $\mathbf{x}_{Kit}$ . This is

$$\begin{aligned}\dot{y}_{it} &= y_{it} - \bar{y}_i \\ \dot{\mathbf{x}}_{Kit} &= \mathbf{x}_{Kit} - \bar{\mathbf{x}}_{Kit}.\end{aligned}$$

The transformed regression equation is

$$\dot{y}_{it} = \dot{\mathbf{x}}'_{Kit} \boldsymbol{\beta}_K + \dot{\varepsilon}_{Kit}.$$

What is important about the within transformation for the regressors is that it is applied to the transformed variables  $\mathbf{x}_{Kit}$ , not the original regressor  $x_{it}$ . For example, in a polynomial regression the within transformation is applied to the powers  $x_{it}^j$ . It is inappropriate to first apply the within transformation to  $x_{it}$  and then construct the basis transformations.

The coefficient is estimated by least-squares on the within transformed variables

$$\hat{\boldsymbol{\beta}}_K = \left( \sum_{i=1}^n \sum_{t=1}^T \dot{\mathbf{x}}_{Kit} \dot{\mathbf{x}}'_{Kit} \right)^{-1} \left( \sum_{i=1}^n \sum_{t=1}^T \dot{\mathbf{x}}_{Kit} \dot{y}_{it} \right).$$

Variance estimators should be calculated using the clustered variance formulas, clustered at the level of the individual  $i$ , as described in Section 21.19.

For selection of the number of series terms  $K$  there is no current theory. A reasonable method is to use delete-cluster cross-validation as described in Section 21.19.

## 21.24 Multiple Regressors

Suppose  $\mathbf{x} \in \mathbb{R}^d$  is vector-valued and continuously distributed. A multivariate series approximation can be obtained as follows. Construct a set of basis transformations for each variable separately. Then take their tensor cross-products. Use these as regressors. For example, a  $p^{th}$ -order polynomial is

$$m_K(\mathbf{x}) = \beta_0 + \sum_{j_1=1}^p \cdots \sum_{j_d=1}^p x_1^{j_1} \cdots x_d^{j_d} \beta_{j_1, \dots, j_d K}.$$

This includes all powers and cross-products. The coefficient vector has dimension  $K = 1 + p^d$ .

The inclusion of cross-products greatly increases the number of coefficients relative to the univariate case. Consequently series applications with multiple regressors typically require large sample sizes.

## 21.25 Additively Separable Models

As discussed in the previous section, when  $\mathbf{x} \in \mathbb{R}^d$  a full series expansion requires a large number of coefficients, which means that estimation precision will be low unless the sample size is quite large. A common simplification is to treat the regression function  $m(\mathbf{x})$  as additively separable in the individual regressors. This means that

$$m(\mathbf{x}) = m_1(x_1) + m_2(x_2) + \cdots + m_d(x_d).$$

We then apply series expansions (polynomials or splines) separately for each component  $m_j(x_j)$ . Essentially, this is the same as the expansions discussed in the previous section, but omitted all the interaction terms.

The advantage of additive separability is the reduction in dimensionality. While an unconstrained  $p^{th}$  order polynomial has  $1 + p^d$  coefficients, an additively separable polynomial model has only  $1 + dp$  coefficients. This is a major reduction.

The disadvantage of additive separability is that the interaction effects have been eliminated. This is a substantive restriction on  $m(\mathbf{x})$ .

The decision to impose additive separability can be based on an economic model which suggests the absence of interaction effects, or can be a model selection decision similar to the selection of the number of series terms.

## 21.26 Nonparametric Instrumental Variables Regression

The basic nonparametric instrumental variables (NPIV) model takes the form

$$\begin{aligned} y_i &= m(x_i) + e_i \\ \mathbb{E}(e_i | z_i) &= 0 \end{aligned} \tag{21.32}$$

where  $y_i$ ,  $x_i$  and  $z_i$  are real valued. Here,  $z_i$  is an instrumental variable and  $x_i$  is an endogenous regressor.

In recent years there have been many papers in the econometrics literature examining the NPIV model, exploring identification, estimation, and inference. Many of these papers are mathematically advanced. Two important and accessible contributions are Newey and Powell (1993) and Horowitz (2011). Here we describe some of the primary results.

A series estimator approximates the function  $m(x)$  with  $m_K(x) = \mathbf{x}_K(x)' \boldsymbol{\beta}_K$  as in (21.4). This leads to the series structural equation

$$y_i = \mathbf{x}'_{Ki} \boldsymbol{\beta}_K + e_{Ki} \tag{21.33}$$

where  $\mathbf{x}_{Ki} = \mathbf{x}_K(x_i)$ . For example, if a polynomial basis is used then  $\mathbf{x}_{Ki} = (1, x_i, \dots, x_i^{K-1})$ .

Since  $x_i$  is endogenous so is the entire vector  $\mathbf{x}_{Ki}$ . Thus we need at least  $K$  instrumental variables. It is useful to consider the reduced form equation for  $x_i$ . A nonparametric specification is

$$\begin{aligned} x_i &= g(z_i) + u_i \\ \mathbb{E}(u_i | z_i) &= 0. \end{aligned}$$

We can approximate  $g(z)$  by the series expansion

$$g(z) \approx g_L(z) = \mathbf{z}_L(z)' \boldsymbol{\gamma}_L$$

where  $\mathbf{z}_L(z)$  is an  $L \times 1$  vector of basis transformations and  $\boldsymbol{\gamma}_L$  is an  $L \times 1$  coefficient vector. For example, if a polynomial basis is used then  $\mathbf{z}_{Li} = (1, z_i, \dots, z_i^{L-1})$ . Most of the literature for simplicity focuses on the case  $L = K$ , but this is not essential to the method.

If  $L \geq K$  we can then use  $\mathbf{z}_{Li} = \mathbf{z}_L(z_i)$  as instruments for  $\mathbf{x}_{Ki}$ . The 2SLS estimator  $\hat{\boldsymbol{\beta}}_{K,L}$  of  $\boldsymbol{\beta}_K$  is

$$\hat{\boldsymbol{\beta}}_{K,L} = \left( \mathbf{X}'_K \mathbf{Z}_L (\mathbf{Z}'_L \mathbf{Z}_L)^{-1} \mathbf{Z}'_L \mathbf{X}_K \right)^{-1} \left( \mathbf{X}'_K \mathbf{Z}_L (\mathbf{Z}'_L \mathbf{Z}_L)^{-1} \mathbf{Z}'_L \mathbf{y} \right).$$

The estimator of  $m(x)$  is  $\hat{m}_K(x) = \mathbf{x}_K(x)' \hat{\beta}_{K,L}$ . If  $L > K$  the linear GMM estimator can be similarly defined.

One way to think about the choice of instruments is to realize that we are actually estimating reduced form equations for each element of  $\mathbf{x}_{Ki}$ . Thus the reduced form system is

$$\begin{aligned}\mathbf{x}_{Ki} &= \boldsymbol{\Gamma}'_K \mathbf{z}_{Li} + \mathbf{u}_{Ki} \\ \boldsymbol{\Gamma}_K &= \mathbb{E}(\mathbf{z}_{Li} \mathbf{z}_{Li}')^{-1} \mathbb{E}(\mathbf{z}_{Li} \mathbf{x}'_{Ki}).\end{aligned}$$

For example, suppose we use a polynomial basis with  $K = L = 3$ . Then the reduced form system (ignoring intercepts) is

$$\begin{bmatrix} x_i \\ x_i^2 \\ x_i^3 \end{bmatrix} = \begin{bmatrix} \Gamma_{11} & \Gamma_{21} & \Gamma_{31} \\ \Gamma_{12} & \Gamma_{22} & \Gamma_{32} \\ \Gamma_{13} & \Gamma_{13} & \Gamma_{23} \end{bmatrix} \begin{bmatrix} z_i \\ z_i^2 \\ z_i^3 \end{bmatrix} + \begin{bmatrix} u_{1i} \\ u_{2i} \\ u_{3i} \end{bmatrix}. \quad (21.34)$$

This is modeling the conditional mean of  $x_i$ ,  $x_i^2$  and  $x_i^3$  as linear functions of  $z_i$ ,  $z_i^2$  and  $z_i^3$ .

To understand if the coefficient  $\boldsymbol{\beta}_K$  is identified, it is useful to consider the simple reduced form equation  $x_i = \gamma_0 + \gamma_1 z_i + u_i$ . Assume that  $\gamma_1 \neq 0$  so that the equation is strongly identified and assume for simplicity that  $u_i$  is independent of  $z_i$  with mean zero and variance  $\sigma_u^2$ . The identification properties of the reduced form are invariant to rescaling and recentering  $x_i$  and  $z_i$  so without loss of generality we can set  $\gamma_0 = 0$  and  $\gamma_1 = 1$ . Then we can calculate that the coefficient matrix in (21.34) is

$$\begin{bmatrix} \Gamma_{11} & \Gamma_{21} & \Gamma_{31} \\ \Gamma_{12} & \Gamma_{22} & \Gamma_{32} \\ \Gamma_{13} & \Gamma_{13} & \Gamma_{23} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3\sigma_u^2 & 0 & 1 \end{bmatrix}.$$

Notice that this is lower triangular and full rank. It turns out that this property holds for any values of  $K = L$  so the coefficient matrix in (21.34) is full rank for any choice of  $K = L$ . This means that identification of the coefficient  $\boldsymbol{\beta}_K$  is strong if the reduced form equation for  $x_i$  is strong. Thus to check the identification condition for  $\boldsymbol{\beta}_K$  it is sufficient to check the reduced form equation for  $x_i$ . A critically important caveat, however, as discussed in the following section, is that identification of  $\boldsymbol{\beta}_K$  does not mean that the structural function  $m(x)$  is identified.

A simple method for pointwise inference is to use conventional methods to estimate  $V_{K,L} = \text{var}(\hat{\beta}_{K,L})$  and then estimate  $\text{var}(\hat{m}_K(x))$  by  $\mathbf{x}_K(x)' \hat{V}_{K,L} \mathbf{x}_K(x)$  as in series regression. Bootstrap methods are typically advocated to achieve better coverage. See Horowitz (2011) for details. For state-of-the-art inference methods see Chen and Pouzo (2015) and Chen and Christensen (2018).

## 21.27 NPIV Identification

In the previous section we discussed identification of the pseudo-true coefficient  $\boldsymbol{\beta}_K$ . In this section we discuss identification of the structural function  $m(x)$ . This is considerably more challenging.

To understand how the function  $m(x)$  is determined, apply the expectation operator  $\mathbb{E}(\cdot | z_i = z)$  to (21.32). We find

$$\mathbb{E}(y_i | z_i = z) = \mathbb{E}(m(x_i) | z_i = z)$$

with the remainder equal to zero because  $\mathbb{E}(e_i | z_i) = 0$ . We can write this equation as

$$\mu(z) = \int m(x) f(x|z) dx \quad (21.35)$$

where  $\mu(z) = \mathbb{E}(y_i | z_i = z)$  is the conditional mean of  $y_i$  given  $z_i = z$  and  $f(x|z)$  is the conditional density of  $x_i$  given  $z_i$ . These two functions are identified<sup>6</sup> from the joint distribution of  $(y_i, x_i, z_i)$ . This means that the unknown function  $m(x)$  is the solution to the **integral equation** (21.35). Conceptually, you can imagine estimating  $\mu(z)$  and  $f(x|z)$  using standard techniques, and then finding the solution  $m(x)$ . In

<sup>6</sup>Technically, if  $\mathbb{E}|y_i| < \infty$ , the joint density of  $(z_i, x_i)$  exists, and the marginal density of  $z_i$  is positive.

essence, this is how  $m(x)$  is defined, and is the nonparametric analog of the classical relationship between the structural and reduced forms.

Unfortunately the solution  $m(x)$  may not be unique, even in situations where a linear IV model is strongly identified. It is related to what is known as the **ill-posed inverse problem**. The latter means that the solution  $m(x)$  is not necessarily a continuous function of  $\mu(z)$ . Identification requires restricting the class of allowable functions  $f(x|z)$ . This is analogous to the linear IV model, where identification requires restrictions on the reduced form equations, but specifying and understanding the needed restrictions is more subtle than in the linear case.

The function  $m(x)$  is identified if it is the unique solution to (21.35). Equivalently,  $m(x)$  is not identified if we can replace  $m(x)$  in (21.35) with  $m(x) + \delta(x)$  for some non-trivial function  $\delta(x)$  yet the solution does not change. The latter occurs when

$$\int \delta(x) f(x|z) dx = 0 \quad (21.36)$$

for all  $z$ . Equivalently,  $m(x)$  is identified if (and only if) (21.36) holds only for the trivial function  $\delta(x) = 0$ .

Newey and Powell (1993) defined this fundamental condition as **completeness**.

**Proposition 21.1** (Completeness)  $m(x)$  is identified if (and only if) the completeness condition holds: (21.36) for all  $z$  implies  $\delta(x) = 0$ .

Completeness is a property of the reduced form conditional density  $f(x|z)$ . It is unaffected by the structural equation  $m(x)$ . This is analogous to the linear IV model, where identification is a property of the reduced form equations, not a property of the structural equation.

As we stated above, completeness may not be satisfied even if the reduced form relationship is strong. This may be easiest to see by a constructed example<sup>7</sup>. Suppose that the reduced form is

$$x_i = z_i + u_i,$$

$\text{var}(z_i) = 1$ ,  $u_i$  is independent of  $z_i$ , and  $u_i$  is distributed  $U[-1, 1]$ . This reduced form equation has  $R^2 = 0.75$  so is strong. The reduced form conditional density is  $f(x|z) = 1/2$  on  $[-1+z, 1+z]$ . Consider  $\delta(x) = \sin(x/\pi)$ . We calculate that

$$\int \delta(x) f(x|z) dx = \int_{-1+z}^{1+z} \sin(x/\pi) dx = 0$$

for every  $z$ , since  $\sin(x/\pi)$  is periodic on intervals of length 2 and integrates to zero over  $[-1, 1]$ . This means that equation (21.35) holds<sup>8</sup> for  $m(x) + \sin(x/\pi)$ . Thus  $m(x)$  is not identified. This is despite the fact that the reduced form equation is strong.

While identification fails for some conditional distributions  $f(x|z)$ , it does not fail for all distributions. Andrews (2017) provides classes of distributions which satisfy the completeness condition and shows that these distribution classes are quite general.

What does this mean in practice? If completeness fails, then the structural equation is not identified and cannot be consistently estimated. Furthermore, by analogy with the weak instruments literature, we expect that if the conditional distribution is close to incomplete, then the structural equation will be poorly identified and our estimators will be imprecise. Since whether or not the conditional distribution is complete is unknown (and more difficult to assess than in the linear model) this is very troubling for empirical research. Effectively, in any given application we do not know whether or not the structural function  $m(x)$  is identified.

<sup>7</sup>This example was suggested by Joachim Freyberger.

<sup>8</sup>In fact, (21.36) holds for  $m(x) + \delta(x)$  for any function  $\delta(x)$  which is periodic on intervals of length 2 and integrates to zero on  $[-1, 1]$ .

A partial answer is provided by Freyberger (2017). He shows that while the hypothesis of incompleteness cannot be tested, the joint hypothesis of incompleteness and small asymptotic bias can be tested. By applying the test proposed in Freyberger (2017), a user can obtain evidence that their NPIV estimator is well-behaved in the sense of having low bias. Unlike Stock and Yogo (2005), however, Freyberger's result does not address inference.

## 21.28 NPIV Convergence Rate

As described in Horowitz (2011), the convergence rate of  $\hat{m}_K(x)$  for  $m(x)$  is

$$|\hat{m}_K(x) - m(x)| = O_p \left( K^{-s} + K^r \left( \frac{K}{n} \right)^{1/2} \right) \quad (21.37)$$

where  $s$  is the smoothness<sup>9</sup> of  $m(x)$  and  $r$  is the smoothness of the joint density  $f_{xz}(x, z)$  of  $(x_i, z_i)$ . The first term  $K^{-s}$  is the bias due to the approximation of  $m(x)$  by  $m_K(x)$  and takes the same form as for series regression. The second term  $K^r (K/n)^{1/2}$  is the standard deviation of  $\hat{m}_K(x)$ . The component  $(K/n)^{1/2}$  is the same as for series regression. The extra component  $K^r$  is due to the ill-posed inverse problem (see the previous section).

From the rate (21.37) we can calculate that the optimal number of series terms is  $K \sim n^{1/(2r+2s+1)}$ . Given this rate the best possible convergence rate in (21.37) is  $O_p(n^{-s/(2r+2s+1)})$ . For  $r > 0$  these rates are slower than for series regression. If we consider the case  $s = 2$ , these rates are  $K \sim n^{1/(2r+5)}$  and  $O_p(n^{-2/(2r+5)})$ , which are slower than the  $K \sim n^{1/5}$  and  $O_p(n^{-2/5})$  rates obtained by series regression.

A very unusual aspect of the rate (21.37) is that smoothness of  $f_{xz}(x, z)$  adversely affects the convergence rate. Larger  $r$  means a slower rate of convergence. The limiting case as  $r \rightarrow \infty$  (for example, joint normality of  $x$  and  $z$ ) results in a logarithmic convergence rate. This seems very strange. The reason is that when the density  $f_{xz}(x, z)$  is very smooth the data contain little information about the function  $m(x)$ . This is not intuitive, and requires a deeper mathematical treatment.

A practical implication of the convergence rate (21.37) is that the number of series terms  $K$  should be much smaller than for regression estimation. Estimation variance increases quickly as  $K$  increases. Therefore  $K$  should not be taken to be too large. In practice, however, it is unclear how to select the series order  $K$  as standard cross-validation methods do not apply.

## 21.29 Nonparametric vs Parametric Identification

One of the insights from the nonparametric identification literature is that it is important to understand which features of a model are nonparametrically identified, meaning that they are identified without functional form assumptions, and which are only identified based on functional form assumptions. Since functional form assumptions are dubious in most economic applications, the strong implication is that researchers should strive to work only with models which are nonparametrically identified.

Even if a model is determined to be nonparametrically identified a researcher may estimate a linear (or another simple parametric) model. This is valid because it can be viewed as an approximation to the nonparametric structure. If, however, the model is identified only under the parametric assumption, then it cannot be viewed as an approximation, and it is unclear how to interpret the model more broadly.

For example, in the regression model

$$\begin{aligned} y_i &= m(x_i) + e_i \\ \mathbb{E}(e_i | x_i) &= 0 \end{aligned}$$

the conditional mean is nonparametrically identified by Theorem 2.14. This means that researchers who estimate linear regressions (or other low-dimensional regressions) can interpret their estimated model as an approximation to the underlying conditional mean function.

---

<sup>9</sup>The number of bounded derivatives.

As another example, in the NPIV model

$$\begin{aligned} y_i &= m(x_i) + e_i \\ \mathbb{E}(e_i | z_i) &= 0 \end{aligned}$$

the structural function  $m(x)$  is identified under the completeness condition. This means that researchers who estimate linear 2SLS regressions can interpret their estimated model as an approximation to  $m(x)$  (subject to the caveat that it is difficult to know if completeness holds).

But the analysis can also point out simple yet subtle mistakes. Take the simple IV model with one exogenous regressor  $x_{1i}$  and one endogenous regressor  $x_{2i}$

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \\ \mathbb{E}(e_i | x_{1i}) &= 0 \end{aligned} \tag{21.38}$$

with no additional instruments. Suppose that an enterprising researcher suggests using the instrument  $x_{1i}^2$  for  $x_{2i}$ , using the reasoning that the assumptions imply that  $\mathbb{E}(x_{1i}^2 e_i) = 0$  so  $x_{1i}^2$  is a valid instrument. The trouble is that the basic model is not nonparametrically identified. If we write (21.38) as a partially linear nonparametric IV problem

$$\begin{aligned} y_i &= m(x_{1i}) + \beta_2 x_{2i} + e_i \\ \mathbb{E}(e_i | x_{1i}) &= 0 \end{aligned} \tag{21.39}$$

then we can see that this model is not identified. We need a valid excluded instrument  $z_i$ . Since (21.39) is not identified, then (21.38) cannot be viewed as a valid approximation. The apparent identification of (21.38) critically rests on the (unknown) truth of the linearity in (21.38).

The point of this example is that (21.38) should never be estimated by 2SLS using the instrument  $x_{1i}^2$  for  $x_{2i}$ , fundamentally because the nonparametric model (21.39) is not identified.

Another way to describe the mistake is to observe that  $x_{1i}^2$  is a valid instrument in (21.38) only if it is a valid exclusion restriction from the structural equation (21.38). Viewed in the context of (21.39) we can see that this is a functional form restriction. As stated above, identification based on functional form restrictions alone is highly undesirable since functional form assumptions are dubious.

### 21.30 Example: Angrist and Lavy (1999)

To illustrate nonparametric instrumental variables in practice, we follow Horowitz (2011) by extending the empirical work reported in Angrist and Lavy (1999). Their paper is concerned with measuring the causal effect of the number of students in an elementary school classroom on academic achievement. They address this using a sample of 4067 Israeli 4<sup>th</sup> and 5<sup>th</sup> grade classrooms. The dependent variable is the classroom average score on an achievement test. Here we consider the reading score *avgverb*, and consider the mathematics score in Exercise 21.17. The explanatory variables are the number of students in the classroom (*classize*), the number of students in the grade at the school (*enrollment*), and a school-level index of students' socioeconomic status that the authors call percent *disadvantaged*. The variables *enrollment* and *disadvantaged* are treated as exogenous, but *classize* is treated as endogenous since wealthier schools may be able to offer smaller class sizes.

The authors suggest the following instrumental variable for class size. Israeli regulations specify that class sizes must be capped at 40. This means that *classize* should be perfectly predictable from *enrollment*. If the regulation is followed, a school with up to 40 students will have one classroom in the grade, schools with 41-80 students will have two classrooms, etc. The precise prediction is that class size should be

$$p = \frac{\text{enrollment}}{1 + [1 + \text{enrollment}/40]} \tag{21.40}$$

where  $[a]$  is the integer part of  $a$ . Angrist and Lavy suggest using  $p$  as an instrumental variable for *classize*.

They estimate several specifications. We focus on equation (6) from their Table VII, which specifies *avgverb* as a linear function of *classize*, *disadvantaged*, *enrollment*, *Grade4*, and the interaction of *classize* and *disadvantaged*, where *Grade4* is a dummy indicator for 4<sup>th</sup> grade classrooms. The equation is estimated by instrumental variables, using *p* and *p \* disadvantaged* as instruments. The observations are treated as clustered at the level of the school. Their estimates show a negative and statistically significant impact of class size on reading test scores.

We are interested in a nonparametric version of their equation. To keep the specification reasonably parsimonious yet flexible we use the following equation.

$$\begin{aligned} \text{avgverb} = & \beta_1 \left( \frac{\text{classize}}{40} \right) + \beta_2 \left( \frac{\text{classize}}{40} \right)^2 + \beta_3 \left( \frac{\text{classize}}{40} \right)^3 \\ & + \beta_4 \left( \frac{\text{disadvantaged}}{14} \right) + \beta_5 \left( \frac{\text{disadvantaged}}{14} \right)^2 + \beta_6 \left( \frac{\text{disadvantaged}}{14} \right)^3 \\ & + \beta_7 \left( \frac{\text{classize}}{40} \right) \left( \frac{\text{disadvantaged}}{14} \right) + \beta_8 \text{enrollment} + \beta_9 \text{Grade4} + \beta_{10} + e. \end{aligned}$$

This is a cubic equation in *classize* and *disadvantaged*, with a single interaction term, and linear in *enrollment* and *Grade4*. The cubic in *disadvantaged* was selected by a delete-cluster cross-validation regression without *classize*. The cubic in *classize* was selected to allow for a minimal degree of nonparametric flexibility without overparameterization. The variables *classize* and *disadvantaged* were scaled by 40 and 14, respectively, so that the regression is well conditioned. The scaling for *classize* was selected so that the variable essentially falls in [0, 1], and the scaling for *disadvantaged* was selected so that its mean is 1.

Table 21.1: Nonparametric Instrumental Variable Regression for Reading Test Score

classize/40	34.2
	(33.4)
(classize/40) <sup>2</sup>	-61.2
	(53.0)
(classize/40) <sup>3</sup>	29.0
	(26.8)
disadvantaged/14	-12.4
	(1.7)
(disadvantaged/14) <sup>2</sup>	3.33
	(0.54)
(disadvantaged/14) <sup>3</sup>	-0.377
	(0.078)
(classize/40)(disadvantaged/14)	0.81
	(1.77)
enrollment	0.015
	(0.007)
Grade 4	-1.96
	(0.16)
Intercept	77.0
	(6.9)

The equation is estimated by just-identified 2SLS, using  $(p/40)$ ,  $(p/40)^2$ ,  $(p/40)^3$  and  $(p/40)*(disadvantaged/14)$  as instruments for the four variables involving *classize*. The parameter estimates are reported in Table 21.1. The standard errors are clustered at the level of the school. Most of the individual coefficients do not have interpretable meaning, except the positive coefficient on *enrollment* shows that larger schools achieve slightly higher testscores, and the negative coefficient on *Grade4* shows that 4<sup>th</sup> grade students have somewhat lower testscores than 5<sup>th</sup> grade students.

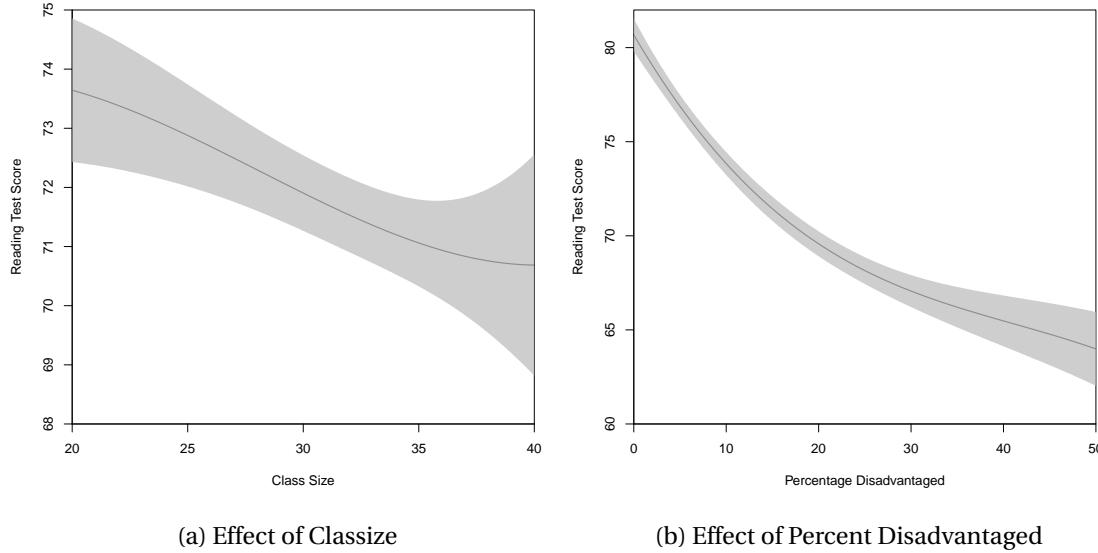


Figure 21.9: Nonparametric Instrumental Variables Estimates of the Effect of Classize and Disadvantaged on Reading Test Scores

To obtain a better interpretation of the results we display the estimated regression functions in Figure 21.9. Panel (a) displays the estimated effect of classize on reading test scores. Panel (b) displays the estimated effect of percent disadvantaged. In both figures the other variables are set at their sample means<sup>10</sup>.

In panel (a) we can see that increasing class size decreases the average test score. This is consistent with the results from the linear model estimated by Angrist and Lavy (1999). The estimated effect is remarkably close to linear. However, the relationship is not precisely estimated, as the pointwise confidence bands are wide.

In panel (b) we can see that increasing the percentage of disadvantaged students greatly decreases the average test score. This effect is substantially greater in magnitude than the effect of class size. The effect also appears to be nonlinear. The effect is quite precisely estimated, with tight pointwise confidence bands.

We can also use the estimated model for hypothesis testing. The question addressed by Angrist and Lavy was whether or not class size has an effect on test scores. Within the nonparametric model estimated here, this hypothesis holds under the linear restriction  $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_7 = 0$ . Examining the individual coefficient estimates and standard errors, it is unclear if this is a significant effect as none of these four coefficient estimates is statistically different from zero. This hypothesis is better tested by a Wald test (using cluster-robust variance estimates). This statistic is 12.7 which has an asymptotic p-value of 0.013. This appears to support the hypothesis that class size has negative effect on student performance.

We can also use the model to quantify the impact of class size on test scores. Consider the impact of increasing a class from 20 to 40 students. In the above model the predicted impact on test scores is

$$\theta = \frac{1}{2}\beta_1 + \frac{3}{4}\beta_2 + \frac{7}{8}\beta_3 + \frac{1}{2}\beta_4.$$

This is a linear function of the coefficients. The point estimate is  $\hat{\theta} = -2.96$  with a standard error of 1.21. (The point estimate is identical to the difference between the endpoints of the estimated function shown in panel (a).) This is a small but substantive impact.

<sup>10</sup>If they are set at other values it does not change the qualitative nature of the plots.

### 21.31 Technical Proofs\*

**Proof of Theorem 21.4.** We provide a proof under the stronger rate assumption  $\zeta_K^2 K/n \rightarrow 0$ . (The proof presented by Belloni, Chernozhukov, Chetverikov, and Kato (2015) requires a more advanced treatment.) Let  $\|\mathbf{A}\|_F$  denote the Frobenius norm (see Section A.23), and write the  $j^{th}$  element of  $\tilde{\mathbf{x}}_{Ki}$  as  $\tilde{x}_{jKi}$ . Using (A.17),

$$\|\tilde{\mathbf{Q}}_K - \mathbf{I}_K\|_2^2 \leq \|\tilde{\mathbf{Q}}_K - \mathbf{I}_K\|_F^2 = \sum_{j=1}^K \sum_{\ell=1}^K \left( \frac{1}{n} \sum_{i=1}^n (\tilde{x}_{jKi} \tilde{x}_{\ell Ki} - \mathbb{E}(\tilde{x}_{jKi} \tilde{x}_{\ell Ki})) \right)^2.$$

Then

$$\begin{aligned} \mathbb{E}(\|\tilde{\mathbf{Q}}_K - \mathbf{I}_K\|_2^2) &\leq \sum_{j=1}^K \sum_{\ell=1}^K \text{var}\left(\frac{1}{n} \sum_{i=1}^n \tilde{x}_{jKi} \tilde{x}_{\ell Ki}\right) \\ &= \frac{1}{n} \sum_{j=1}^K \sum_{\ell=1}^K \text{var}(\tilde{x}_{jKi} \tilde{x}_{\ell Ki}) \\ &\leq \frac{1}{n} \mathbb{E}\left(\sum_{j=1}^K \tilde{x}_{jKi}^2 \sum_{\ell=1}^K \tilde{x}_{\ell Ki}^2\right) \\ &= \frac{1}{n} \mathbb{E}\left((\tilde{\mathbf{x}}'_{Ki} \tilde{\mathbf{x}}_{Ki})^2\right) \\ &\leq \frac{\zeta_K^2}{n} \mathbb{E}(\tilde{\mathbf{x}}'_{Ki} \tilde{\mathbf{x}}_{Ki}) \\ &= \frac{\zeta_K^2 K}{n} \\ &\longrightarrow 0 \end{aligned}$$

where final three lines use (21.16),  $\mathbb{E}(\tilde{\mathbf{x}}'_{Ki} \tilde{\mathbf{x}}_{Ki}) = K$ , and  $\zeta_K^2 K/n \rightarrow 0$ . Markov's inequality implies (21.18). ■

**Proof of Theorem 21.5.** By the spectral decomposition we can write  $\tilde{\mathbf{Q}}_K = \mathbf{H}' \Lambda \mathbf{H}$  where  $\mathbf{H}' \mathbf{H} = \mathbf{I}_K$  and  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_K)$  are the eigenvalues. Then

$$\|\tilde{\mathbf{Q}}_K - \mathbf{I}_K\|_2 = \|\mathbf{H}' (\Lambda - \mathbf{I}_K) \mathbf{H}\|_2 = \|\Lambda - \mathbf{I}_K\|_2 = \max_{j \leq K} |\lambda_j - 1| \xrightarrow{p} 0$$

by Theorem 21.4. This implies

$$\min_{j \leq K} |\lambda_j| \xrightarrow{p} 1$$

which is (21.20). Similarly

$$\begin{aligned} \|\tilde{\mathbf{Q}}_K^{-1} - \mathbf{I}_K\|_2 &= \|\mathbf{H}' (\Lambda^{-1} - \mathbf{I}_K) \mathbf{H}\|_2 \\ &= \|\Lambda^{-1} - \mathbf{I}_K\|_2 \\ &= \max_{j \leq K} |\lambda_j^{-1} - 1| \\ &\leq \frac{\max_{j \leq K} |1 - \lambda_j|}{\min_{j \leq K} |\lambda_j|} \\ &\xrightarrow{p} 0. \end{aligned}$$

**Proof of Theorem 21.6.** Using (21.12) we can write

$$\hat{m}_K(x) - m(x) = \mathbf{x}_K(x)' (\hat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K) - r_K(x). \quad (21.41)$$

Since  $e_{Ki} = r_{Ki} + e_i$  is a projection error, it satisfies  $\mathbb{E}(x_{Ki}e_{Ki}) = 0$ . Since  $e_i$  is a regression error it satisfies  $\mathbb{E}(x_{Ki}e_i) = 0$ . We deduce  $\mathbb{E}(x_{Ki}r_{Ki}) = 0$ . Hence  $\int x_K(x)r_K(x)f(x)dx = \mathbb{E}(x_{Ki}r_{Ki}) = 0$ . Also observe that  $\int x_K(x)x_K(x)'dF(x) = Q_K$  and  $\int r_K(x)^2dF(x) = \mathbb{E}(r_{Ki}^2) = \delta_K^2$ . Then

$$\begin{aligned} ISE(K) &= \int (x_K(x)'(\hat{\beta}_K - \beta_K) - r_K(x))^2 dF(x) \\ &= (\hat{\beta}_K - \beta_K)' \left( \int x_K(x)x_K(x)'dF(x) \right) (\hat{\beta}_K - \beta_K) \\ &\quad - 2(\hat{\beta}_K - \beta_K)' \left( \int x_K(x)r_K(x)dF(x) \right) + \int r_K(x)^2 dF(x) \\ &= (\hat{\beta}_K - \beta_K)' Q_K (\hat{\beta}_K - \beta_K) + \delta_K^2. \end{aligned} \quad (21.42)$$

We calculate that

$$\begin{aligned} (\hat{\beta}_K - \beta_K)' Q_K (\hat{\beta}_K - \beta_K) &= (e'_K X_K)(X'_K X_K)^{-1} Q_K (X'_K X_K)^{-1} (X'_K e_K) \\ &= (e'_K \tilde{X}_K) (\tilde{X}'_K \tilde{X}_K)^{-1} (\tilde{X}'_K \tilde{X}_K)^{-1} (\tilde{X}'_K e_K) \\ &= n^{-2} (e'_K \tilde{X}_K) \tilde{Q}_K^{-1} \tilde{Q}_K^{-1} (\tilde{X}'_K e_K) \\ &\leq \left( \lambda_{\max}(\tilde{Q}_K^{-1}) \right)^2 \left( n^{-2} e'_K \tilde{X}_K \tilde{X}'_K e_K \right) \\ &\leq O_p(1) (n^{-2} e'_K X_K Q_K^{-1} X'_K e_K) \end{aligned} \quad (21.43)$$

where  $\tilde{X}_K$  and  $\tilde{Q}_K$  are the orthogonalized regressors as defined in (21.17). The first inequality is the Quadratic Inequality (B.18), the second is (21.20).

Using the fact that  $x_{Ki}e_{Ki}$  are mean zero and uncorrelated, (21.16),  $\mathbb{E}(e_{Ki}^2) \leq \mathbb{E}(y_i^2) < \infty$  and Assumption 21.1.2

$$\begin{aligned} \mathbb{E}(n^{-2} e'_K X_K Q_K^{-1} X'_K e_K) &= n^{-1} \mathbb{E}(x'_{Ki} Q_K^{-1} x_{Ki} e_{Ki}^2) \\ &\leq \frac{\zeta_K^2}{n} \mathbb{E}(e_{Ki}^2) \\ &\leq o(1). \end{aligned} \quad (21.44)$$

This shows that (21.43) is  $o_p(1)$ . Combined with (21.42) we find  $ISE(K) = o_p(1)$  as claimed. ■

**Proof of Theorem 21.7.** The assumption  $\sigma^2(x) \leq \bar{\sigma}^2$  implies that

$$\mathbb{E}(e_{Ki}^2 | x_i) = \mathbb{E}((r_{Ki} + e_i)^2 | x_i) = r_{Ki}^2 + \sigma^2(x_i) \leq r_{Ki}^2 + \bar{\sigma}^2.$$

Thus (21.44) is bounded by

$$\begin{aligned} n^{-1} \mathbb{E}(x'_{Ki} Q_K^{-1} x_{Ki} r_{Ki}^2) + n^{-1} \mathbb{E}(x'_{Ki} Q_K^{-1} x_{Ki}) \bar{\sigma}^2 &\leq \frac{\zeta_K^2}{n} \mathbb{E}(r_{Ki}^2) + n^{-1} \mathbb{E} \text{tr}(Q_K^{-1} x_{Ki} x'_{Ki}) \bar{\sigma}^2 \\ &= \frac{\zeta_K^2}{n} \delta_K^2 + n^{-1} \mathbb{E} \text{tr}(I_K) \bar{\sigma}^2 \\ &\leq o(\delta_K^2) + \frac{K}{n} \bar{\sigma}^2 \end{aligned}$$

where the inequality is Assumption 21.1.2. This implies (21.43) is  $o_p(\delta_K^2) + O_p(K/n)$ . Combined with (21.42) we find  $ISE(K) = O_p(\delta_K^2 + K/n)$  as claimed. ■

**Proof of Theorem 21.8.** Using (21.12) and linearity

$$\begin{aligned} \theta &= a(m) \\ &= a(z_K(x)' \beta_K) + a(r_K) \\ &= a'_K \beta_K + a(r_K). \end{aligned}$$

Thus

$$\begin{aligned} \sqrt{\frac{n}{V_K}} (\hat{\theta}_K - \theta + a(r_K)) &= \sqrt{\frac{n}{V_K}} \mathbf{a}'_K (\hat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K) \\ &= \sqrt{\frac{1}{nV_K}} \mathbf{a}'_K \hat{\mathbf{Q}}_K^{-1} \mathbf{X}'_K \mathbf{e}_K \\ &= \frac{1}{\sqrt{nV_K}} \mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{X}'_K \mathbf{e} \end{aligned} \quad (21.45)$$

$$+ \frac{1}{\sqrt{nV_K}} \mathbf{a}'_K (\hat{\mathbf{Q}}_K^{-1} - \mathbf{Q}_K^{-1}) \mathbf{X}'_K \mathbf{e} \quad (21.46)$$

$$+ \frac{1}{\sqrt{nV_K}} \mathbf{a}'_K \hat{\mathbf{Q}}_K^{-1} \mathbf{X}'_K \mathbf{r}_K. \quad (21.47)$$

where we have used  $\mathbf{e}_K = \mathbf{e} + \mathbf{r}_K$ . We now take the terms in (21.45)-(21.47) separately. We show that (21.45) is asymptotically normal and (21.46)-(21.47) are asymptotically negligible.

First, take (21.45). We can write

$$\frac{1}{\sqrt{nV_K}} \mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{X}'_K \mathbf{e} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{\sqrt{V_K}} \mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{x}_{Ki} e_i. \quad (21.48)$$

Observe that  $\mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{x}_{Ki} e_i / \sqrt{V_K}$  are independent across  $i$ , mean zero, and have variance 1. We will apply Theorem 6.12, for which it is sufficient to verify Lindeberg's condition: for all  $\varepsilon > 0$

$$\mathbb{E} \left( \frac{(\mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{x}_{Ki} e_i)^2}{V_K} \mathbf{1} \left( \frac{(\mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{x}_{Ki} e_i)^2}{V_K} \geq n\varepsilon \right) \right) \rightarrow 0. \quad (21.49)$$

Pick  $\eta > 0$ . Set  $B$  sufficiently large so that  $\mathbb{E}(e_i^2 \mathbf{1}(e_i^2 > B) | x_i) \leq \underline{\sigma}^2 \eta$  which is feasible by Assumption 21.2.1. Pick  $n$  sufficiently large so that  $\zeta_K^2/n \leq \varepsilon \underline{\sigma}^2/B$ , which is feasible under Assumption 21.1.2.

By Assumption 21.2.2

$$\begin{aligned} V_K &= \mathbb{E} \left( (\mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{x}_{Ki})^2 e_i^2 \right) \\ &= \mathbb{E} \left( (\mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{x}_{Ki})^2 \sigma(x_i^2) \right) \\ &\geq \mathbb{E} \left( (\mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{x}_{Ki})^2 \underline{\sigma}^2 \right) \\ &= \mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbb{E}(\mathbf{x}_{Ki} \mathbf{x}'_{Ki}) \mathbf{Q}_K^{-1} \mathbf{a}_K \underline{\sigma}^2 \\ &= \mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{a}_K \underline{\sigma}^2. \end{aligned} \quad (21.50)$$

Then by the Schwarz Inequality, (21.16), (21.50), and  $\zeta_K^2/n \leq \varepsilon \underline{\sigma}^2/B$

$$\frac{(\mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{x}_{Ki})^2}{V_K} \leq \frac{(\mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{a}_K)(\mathbf{x}'_{Ki} \mathbf{Q}_K^{-1} \mathbf{x}_{Ki})}{V_K} \leq \frac{\zeta_K^2}{\underline{\sigma}^2} \leq \frac{\varepsilon}{B} n.$$

Then the left-side of (21.49) is smaller than

$$\begin{aligned} \mathbb{E} \left( \frac{(\mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{x}_{Ki})^2}{V_K} e_i^2 \mathbf{1}(e_i^2 \geq B) \right) &= \mathbb{E} \left( \frac{(\mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{x}_{Ki})^2}{V_K} \mathbb{E}(e_i^2 \mathbf{1}(e_i^2 \geq B) | x_i) \right) \\ &\leq \mathbb{E} \left( \frac{(\mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{x}_{Ki})^2}{V_K} \right) \underline{\sigma}^2 \eta \\ &\leq \frac{\mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{a}_K}{V_K} \underline{\sigma}^2 \eta \\ &\leq \eta \end{aligned}$$

the final inequality by (21.50). Since  $\eta$  is arbitrary this verifies (21.49) and we conclude

$$\frac{1}{\sqrt{nV_K}} \mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{X}'_K \mathbf{e} \xrightarrow{d} N(0, 1). \quad (21.51)$$

Second, take (21.46). Assumption 21.2 implies  $\mathbb{E}(e_i^2 | x_i) \leq \bar{\sigma}^2 < \infty$ . Since  $\mathbb{E}(\mathbf{e} | \mathbf{X}) = 0$ , then applying  $\mathbb{E}(e_i^2 | x_i) \leq \bar{\sigma}^2$ , the Schwarz and Norm Inequalities, (21.50), Theorems 21.4 and 21.5,

$$\begin{aligned} & \mathbb{E}\left(\left(\frac{1}{\sqrt{nV_K}} \mathbf{a}'_K (\widehat{\mathbf{Q}}_K^{-1} - \mathbf{Q}_K^{-1}) \mathbf{X}'_K \mathbf{e}\right)^2 | \mathbf{X}\right) \\ &= \frac{1}{nV_K} \mathbf{a}'_K (\widehat{\mathbf{Q}}_K^{-1} - \mathbf{Q}_K^{-1}) \mathbf{X}'_K \mathbb{E}(\mathbf{e} \mathbf{e}' | \mathbf{X}) \mathbf{X}_K (\widehat{\mathbf{Q}}_K^{-1} - \mathbf{Q}_K^{-1}) \mathbf{a}_K \\ &\leq \frac{\bar{\sigma}^2}{V_K} \mathbf{a}'_K (\widehat{\mathbf{Q}}_K^{-1} - \mathbf{Q}_K^{-1}) \widehat{\mathbf{Q}}_K (\widehat{\mathbf{Q}}_K^{-1} - \mathbf{Q}_K^{-1}) \mathbf{a}_K \\ &\leq \frac{\bar{\sigma}^2 \mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{a}_K}{V_K} \|(\widehat{\mathbf{Q}}_K^{-1} - \mathbf{Q}_K^{-1}) \widehat{\mathbf{Q}}_K (\widehat{\mathbf{Q}}_K^{-1} - \mathbf{Q}_K^{-1})\| \\ &= \frac{\bar{\sigma}^2 \mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{a}_K}{V_K} \|(\mathbf{I}_K - \widetilde{\mathbf{Q}}_K)(\widetilde{\mathbf{Q}}_K^{-1} - \mathbf{I}_K)\| \\ &\leq \frac{\bar{\sigma}^2}{\underline{\sigma}^2} \|\mathbf{I}_K - \widetilde{\mathbf{Q}}_K\| \|\widetilde{\mathbf{Q}}_K^{-1} - \mathbf{I}_K\| \\ &\leq \frac{\bar{\sigma}^2}{\underline{\sigma}^2} o_p(1). \end{aligned}$$

This establishes that (21.46) is  $o_p(1)$ .

Third, take (21.47). By the Cauchy-Schwarz inequality, the Quadratic Inequality, (21.50), and (21.20),

$$\begin{aligned} & \left(\frac{1}{\sqrt{nV_K}} \mathbf{a}'_K \widehat{\mathbf{Q}}_K^{-1} \mathbf{X}'_K \mathbf{r}_K\right)^2 \\ &\leq \frac{\mathbf{a}'_K \mathbf{Q}_K^{-1} \mathbf{a}_K}{nV_K} \mathbf{r}'_K \mathbf{X}_K \widehat{\mathbf{Q}}_K^{-1} \mathbf{Q}_K \widehat{\mathbf{Q}}_K^{-1} \mathbf{X}'_K \mathbf{r}_K \\ &\leq \frac{1}{\underline{\sigma}^2} \left(\lambda_{\max} \widetilde{\mathbf{Q}}_K^{-1}\right)^2 \frac{1}{n} \mathbf{r}'_K \mathbf{X}_K \mathbf{Q}_K^{-1} \mathbf{X}'_K \mathbf{r}_K \\ &\leq O_p(1) \frac{1}{n} \mathbf{r}'_K \mathbf{X}_K \mathbf{Q}_K^{-1} \mathbf{X}'_K \mathbf{r}_K. \end{aligned} \quad (21.52)$$

Observe that since the observations are independent,  $\mathbb{E}(\mathbf{x}_{Ki} r_{Ki}) = 0$ ,  $\mathbf{x}'_{Ki} \mathbf{Q}_K^{-1} \mathbf{x}_{Ki} \leq \zeta_K^2$ , and  $\mathbb{E}(r_{Ki}^2) = \delta_K^2$ ,

$$\begin{aligned} \mathbb{E}\left(\frac{1}{n} \mathbf{r}'_K \mathbf{X}_K \mathbf{Q}_K^{-1} \mathbf{X}'_K \mathbf{r}_K\right) &= \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n r_{Ki} \mathbf{x}'_{Ki} \mathbf{Q}_K^{-1} \sum_{j=1}^n \mathbf{x}_{Kj} r_{Kj}\right) \\ &= \mathbb{E}(\mathbf{x}'_{Ki} \mathbf{Q}_K^{-1} \mathbf{x}_{Ki} r_{Ki}^2) \\ &\leq \zeta_K^2 \mathbb{E}(r_{Ki}^2) \\ &= \zeta_K^2 \delta_K^2 \\ &= o(1) \end{aligned}$$

under Assumption 21.2.3. Thus  $\frac{1}{n} \mathbf{r}'_K \mathbf{X}_K \mathbf{Q}_K^{-1} \mathbf{X}'_K \mathbf{r}_K = o_p(1)$ , (21.52) is  $o_p(1)$  and (21.47) is  $o_p(1)$ .

Together, we have shown that

$$\sqrt{\frac{n}{V_K}} (\widehat{\theta}_K - \theta_K + \alpha(r_K)) \xrightarrow{d} N(0, 1)$$

as claimed. ■

**Proof of Theorem 21.10.** It is sufficient to show that

$$\frac{\sqrt{n}}{V_K^{1/2}(x)} r_K(x) = o(1). \quad (21.53)$$

Notice that by Assumption 21.2.2

$$\begin{aligned} V_K(x) &= \mathbf{x}_K(x)' \mathbf{Q}_K^{-1} \boldsymbol{\Omega}_K \mathbf{Q}_K^{-1} \mathbf{x}_K(x) \\ &= \mathbb{E} \left( (\mathbf{x}_K(x)' \mathbf{Q}_K^{-1} \mathbf{x}_{Ki})^2 e_i^2 \right) \\ &= \mathbb{E} \left( (\mathbf{x}_K(x)' \mathbf{Q}_K^{-1} \mathbf{x}_{Ki})^2 \sigma^2(x_i) \right) \\ &\geq \mathbb{E} \left( (\mathbf{x}_K(x)' \mathbf{Q}_K^{-1} \mathbf{x}_{Ki})^2 \right) \underline{\sigma}^2 \\ &= \mathbf{x}_K(x)' \mathbf{Q}_K^{-1} \mathbb{E}(\mathbf{x}_{Ki} \mathbf{x}'_{Ki}) \mathbf{Q}_K^{-1} \mathbf{x}_K(x) \underline{\sigma}^2 \\ &= \mathbf{x}_K(x)' \mathbf{Q}_K^{-1} \mathbf{x}_K(x) \underline{\sigma}^2 \\ &= \zeta_K(x)^2 \underline{\sigma}^2. \end{aligned} \quad (21.54)$$

Using the definitions  $\boldsymbol{\beta}_K^*$ ,  $r_K^*(x)$  and  $\delta_K^*$  from Section 21.8, note that

$$r_K(x) = m(x) - \mathbf{x}'_K(x) \boldsymbol{\beta}_K = r_K^*(x) + \mathbf{x}'_K(x) (\boldsymbol{\beta}_K^* - \boldsymbol{\beta}_K).$$

By the Triangle Inequality, the definition (21.10), the Schwarz Inequality, and definition (21.14)

$$\begin{aligned} |r_K(x)| &\leq |r_K^*(x)| + |\mathbf{x}'_K(x) (\boldsymbol{\beta}_K^* - \boldsymbol{\beta}_K)| \\ &\leq \delta_K^* + |\mathbf{x}'_K(x) \mathbf{Q}_K^{-1} \mathbf{x}'_K(x)|^{1/2} \left| (\boldsymbol{\beta}_K^* - \boldsymbol{\beta}_K)' \mathbf{Q}_K (\boldsymbol{\beta}_K^* - \boldsymbol{\beta}_K) \right|^{1/2} \\ &= \delta_K^* + \zeta_K(x) \left| (\boldsymbol{\beta}_K^* - \boldsymbol{\beta}_K)' \mathbf{Q}_K (\boldsymbol{\beta}_K^* - \boldsymbol{\beta}_K) \right|^{1/2}. \end{aligned}$$

The coefficients satisfy the relationship

$$\boldsymbol{\beta}_K = \mathbb{E}(\mathbf{x}_{Ki} \mathbf{x}'_{Ki})^{-1} \mathbb{E}(\mathbf{x}_{Ki} m(x_i)) = \boldsymbol{\beta}_K^* + \mathbb{E}(\mathbf{x}_{Ki} \mathbf{x}'_{Ki})^{-1} \mathbb{E}(\mathbf{x}_{Ki} r_{Ki}^*).$$

Thus

$$\begin{aligned} (\boldsymbol{\beta}_K^* - \boldsymbol{\beta}_K)' \mathbf{Q}_K (\boldsymbol{\beta}_K^* - \boldsymbol{\beta}_K) &= \mathbb{E}(r_{Ki}^* \mathbf{x}'_{Ki}) \mathbb{E}(\mathbf{x}_{Ki} \mathbf{x}'_{Ki})^{-1} \mathbb{E}(\mathbf{x}_{Ki} r_{Ki}^*) \\ &\leq \mathbb{E}(r_{Ki}^{2*}) \\ &\leq \delta_K^{*2}. \end{aligned}$$

The first inequality is because  $\mathbb{E}(r_{Ki}^* \mathbf{x}'_{Ki}) \mathbb{E}(\mathbf{x}_{Ki} \mathbf{x}'_{Ki})^{-1} \mathbb{E}(\mathbf{x}_{Ki} r_{Ki}^*)$  is a projection. The second inequality follows from the definition (21.10). We deduce that

$$|r_K(x)| \leq (1 + \zeta_K(x)) \delta_K^* \leq 2\zeta_K(x) \delta_K^*. \quad (21.55)$$

Equations (21.54), (21.55), and  $n\delta_K^{*2} = o(1)$  together imply that

$$\frac{n}{V_K(x)} r_K^2(x) \leq \frac{4}{\underline{\sigma}^2} n\delta_K^{*2} = o(1)$$

which is (21.53), as required. ■

## Exercises

**Exercise 21.1** Take the linear spline with three knots

$$m_K(x) = \beta_0 + \beta_1 x + \beta_2 (x - \tau_1) \mathbf{1}(x \geq \tau_1) + \beta_3 (x - \tau_2) \mathbf{1}(x \geq \tau_2) + \beta_4 (x - \tau_3) \mathbf{1}(x \geq \tau_3).$$

Find the (inequality) restrictions on the coefficients  $\beta_j$  so that  $m_K(x)$  is non-decreasing.

**Exercise 21.2** Take the linear spline from the previous question. Find the (inequality) restrictions on the coefficients  $\beta_j$  so that  $m_K(x)$  is concave.

**Exercise 21.3** Take the quadratic spline with three knots

$$m_K(x) = \beta_0 + \beta_1 x + \beta_2 x^3 + \beta_3 (x - \tau_1)^2 \mathbf{1}(x \geq \tau_1) + \beta_4 (x - \tau_2)^2 \mathbf{1}(x \geq \tau_2) + \beta_5 (x - \tau_3)^2 \mathbf{1}(x \geq \tau_3).$$

Find the (inequality) restrictions on the coefficients  $\beta_j$  so that  $m_K(x)$  is concave.

**Exercise 21.4** Consider spline estimation with one knot  $\tau$ . Explain why the knot  $\tau$  must be within the sample support of  $x_i$ . [Explain what happens if you estimate the regression with the knot placed outside the support of  $x_i$ ].

**Exercise 21.5** You estimate the polynomial regression model:

$$\hat{m}_K(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \cdots + \hat{\beta}_p x^p.$$

You are interested in the regression derivative  $m'(x)$  at  $x$ .

- (a) Write out the estimator  $\hat{m}'_K(x)$  of  $m'(x)$ .
- (b) Is  $\hat{m}'_K(x)$  a linear function of the coefficient estimates?
- (c) Use Theorem 21.8 to obtain the asymptotic distribution of  $\hat{m}'_K(x)$ .
- (d) Show how to construct standard errors and confidence intervals for  $\hat{m}'_K(x)$ .

**Exercise 21.6** Does rescaling  $y_i$  or  $x_i$  (multiplying by a constant) affect the CV( $K$ ) function? The  $K$  which minimizes it?

**Exercise 21.7** Take the NPIV approximating equation (21.33) and error  $e_{Ki}$ .

- (a) Does it satisfy  $\mathbb{E}(e_{Ki} | z_i) = 0$ ?
- (b) If  $L = K$ , can you define  $\beta_K$  so that  $\mathbb{E}(z_{Ki} e_{Ki}) = 0$ ?
- (c) If  $L > K$ , does  $\mathbb{E}(z_{Ki} e_{Ki}) = 0$ ?

**Exercise 21.8** Take the cps09mar dataset (full sample).

- (a) Estimate a 6<sup>th</sup> order polynomial regression of  $\log(wage)$  on *experience*. To reduce the ill-conditioned problem, first rescale *experience* to lie in the interval [0, 1] before estimating the regression.
- (b) Plot the estimated regression function along with 95% pointwise confidence intervals.
- (c) Interpret the findings. How do you interpret the estimated function for experience levels exceeding 65?

**Exercise 21.9** Continuing the previous exercise, compute the cross-validation function (or alternatively the AIC) for polynomial orders 1 through 8.

- (a) Which order minimizes the function?
- (b) Plot the estimated regression function along with 95% pointwise confidence intervals.

**Exercise 21.10** Take the `cps09mar` dataset (full sample).

- (a) Estimate a 6<sup>th</sup> order polynomial regression of  $\log(wage)$  on  $education$ . To reduce the ill-conditioned problem, first rescale  $education$  to lie in the interval [0, 1].
- (b) Plot the estimated regression function along with 95% pointwise confidence intervals.

**Exercise 21.11** Continuing the previous exercise, compute the cross-validation function (or alternatively the AIC) for polynomial orders 1 through 8.

- (a) Which order minimizes the function?
- (b) Plot the estimated regression function along with 95% pointwise confidence intervals.

**Exercise 21.12** Take the `cps09mar` dataset (full sample).

- (a) Estimate quadratic spline regressions of  $\log(wage)$  on  $experience$ . Estimate four models: (1) no knots (a quadratic); (2) one knot at 20 years; (3) two knots at 20 and 40; (4) four knots at 10, 20, 30 & 40. Plot the four estimates. Interpret your findings.
- (b) Compare the four splines models using either cross-validation or AIC. Which is the preferred specification?
- (c) For your selected specification, plot the estimated regression function along with 95% pointwise confidence intervals. Interpret your findings.
- (d) If you also estimated a polynomial specification, do you prefer the polynomial or the quadratic spline estimates?

**Exercise 21.13** Take the `cps09mar` dataset (full sample).

- (a) Estimate quadratic spline regressions of  $\log(wage)$  on  $education$ . Estimate four models: (1) no knots (a quadratic); (2) one knot at 10 years; (3) three knots at 5, 10 and 15; (4) four knots at 4, 8, 12, & 16. Plot the four estimates. Interpret your findings.
- (b) Compare the four splines models using either cross-validation or AIC. Which is the preferred specification?
- (c) For your selected specification, plot the estimated regression function along with 95% pointwise confidence intervals. Interpret your findings.
- (d) If you also estimated a polynomial specification, do you prefer the polynomial or the quadratic spline estimates?

**Exercise 21.14** The RR2010 dataset is from Reinhart and Rogoff (2010). It contains observations on annual U.S. GDP growth rates, inflation rates, and the debt/gdp ratio for the long time span 1791-2009. The paper made the strong claim that gdp growth slows as debt/gdp increases, and in particular that this relationship is nonlinear with debt negatively affecting growth for debt ratios exceeding 90%. Their full dataset includes 44 countries, our extract only includes the United States. Let  $y_t$  denote GDP growth and let  $d_t$  denote debt/gdp. We will estimate the partial linear specificaiton

$$y_t = \alpha y_{t-1} + m(d_{t-1}) + e_t$$

using a linear spline for  $m(d)$ .

- (a) Estimate (1) linear model; (2) linear spline with one knot at  $d_{t-1} = 60$ ; (3) linear spline with two knots at 40 and 80. Plot the three estimates.
- (b) For the model with one knot, plot with 95% confidence intervals.
- (c) Compare the three splines models using either cross-validation or AIC. Which is the preferred specification?
- (d) Interpret the findings.

**Exercise 21.15** Take the DDK2011 dataset (full sample). Use a quadratic spline to estimate the regression of *testscores* on *percentile*.

- (a) Estimate five models: (1) no knots (a quadratic); (2) one knot at 50; (3) two knots at 33 and 66; (4) three knots at 25, 50 & 75; (5) knots at 20, 40, 60, & 80. Plot the five estimates. Interpret your findings.
- (b) Select a model. Consider using leave-cluster-one CV.
- (c) For your selected specification, plot the estimated regression function along with 95% pointwise confidence intervals. [Use cluster-robust standard errors.] Interpret your findings.

**Exercise 21.16** The CHJ2004 dataset is from Cox, Hansen and Jimenez (2004). As described in Section 21.6 it contains a sample of 8684 urban Phillipino households. This paper studied the crowding-out impact of a family's *income* on non-governmental *transfers*. Estimate an analog of Figure 21.3 using polynomial regression. Regress *transfers* on the regression controls (variables 2 through 16 in the dataset) and a high-order polynomial in *income*. Ideally, select the polynomial order by cross-validation. You will need to rescale the variable *income* before taking polynomial powers. Plot the estimated function along with 95% pointwise confidence intervals. Comment on the similarities and differences with Figure 21.3.

**Exercise 21.17** The AL1999 dataset is from Angrist and Lavy (1999). It contains 4067 observations on classroom test scores and explanatory variables, including those described in Section 21.30. In Section 21.30 we report a nonparametric instrumental variables regression of reading test scores (*avgverb*) on *classize*, *disadvantaged*, *enrollment* and *Grade4*, using the Angrist-Levy variable (21.40) as an instrument. Repeat the analysis, but instead of reading test scores (*avgverb*) use math test scores (*avgmath*) as the dependent variable. Comment on the similarities and differences with the results for reading test scores.

## **Chapter 22**

# **Regression Discontinuity**

To be written.

# Chapter 23

## Nonlinear Econometric Models

### 23.1 Introduction

This chapter surveys a set of core econometric methods which require nonlinear estimation. This chapter is preliminary.

For more detailed textbook treatments see Maddala (1983), Cameron and Trivedi (1998), Gourieroux (2000), Cameron and Trivedi (2005), Wooldridge (2010), and Greene (2017).

### 23.2 Nonlinear Least Squares

In some cases we might use a parametric regression function  $m(\mathbf{x}, \boldsymbol{\theta}) = \mathbb{E}(y_i | \mathbf{x}_i = \mathbf{x})$  which is a nonlinear function of the parameters  $\boldsymbol{\theta}$ . We describe this setting as **nonlinear regression**.

**Example 23.1** Exponential Link Regression

$$m(\mathbf{x}, \boldsymbol{\theta}) = \exp(\mathbf{x}'\boldsymbol{\theta})$$

The exponential link function is strictly positive, so this choice can be useful when it is desired to constrain the mean to be strictly positive.

**Example 23.2** Logistic Link Regression

$$m(\mathbf{x}, \boldsymbol{\theta}) = \Lambda(\mathbf{x}'\boldsymbol{\theta})$$

where

$$\Lambda(u) = (1 + \exp(-u))^{-1} \quad (23.1)$$

is the Logistic distribution function. Since the logistic link function lies in  $[0, 1]$ , this choice can be useful when the conditional mean is bounded between 0 and 1.

**Example 23.3** Exponentially Transformed Regressors

$$m(\mathbf{x}, \boldsymbol{\theta}) = \theta_1 + \theta_2 \exp(\theta_3 x)$$

**Example 23.4** Power Transformation

$$m(\mathbf{x}, \boldsymbol{\theta}) = \theta_1 + \theta_2 x^{\theta_3}$$

with  $x > 0$ .

**Example 23.5** Box-Cox Transformed Regressors

$$m(x, \boldsymbol{\theta}) = \theta_1 + \theta_2 x^{(\theta_3)}$$

where

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \text{if } \lambda > 0 \\ \log(x), & \text{if } \lambda = 0 \end{cases} \quad (23.2)$$

and  $x > 0$ . The function (23.2) is called the Box-Cox Transformation and was introduced by Box and Cox (1964). The function nests linearity ( $\lambda = 1$ ) and logarithmic ( $\lambda = 0$ ) transformations continuously.

**Example 23.6** Continuous Threshold Regression

$$m(x, \boldsymbol{\theta}) = \theta_1 + \theta_2 x + \theta_3 (x - \theta_4) \mathbf{1}(x > \theta_4)$$

**Example 23.7** Threshold Regression

$$m(\mathbf{x}, \boldsymbol{\theta}) = (\theta'_1 \mathbf{x}_1) \mathbf{1}(x_2 < \theta_3) + (\theta'_2 \mathbf{x}_1) \mathbf{1}(x_2 \geq \theta_3)$$

**Example 23.8** Smooth Transition

$$m(\mathbf{x}, \boldsymbol{\theta}) = \theta'_1 \mathbf{x}_1 + (\theta'_2 \mathbf{x}_1) \Lambda\left(\frac{x_2 - \theta_3}{\theta_4}\right)$$

where  $\Lambda(u)$  is the logit function (23.1).

What differentiates these examples from the linear regression model is that the conditional mean cannot be written as a linear function of the parameter vector  $\boldsymbol{\theta}$ .

Nonlinear regression is sometimes adopted because the functional form  $m(\mathbf{x}, \boldsymbol{\theta})$  is suggested by an economic model. In other cases, it is adopted as a flexible approximation to an unknown regression function.

The least squares estimator  $\hat{\boldsymbol{\theta}}$  minimizes the normalized sum-of-squared-errors

$$\hat{S}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n (y_i - m(\mathbf{x}_i, \boldsymbol{\theta}))^2.$$

When the regression function is nonlinear, we call  $\hat{\boldsymbol{\theta}}$  the **nonlinear least squares** (NLLS) estimator. The NLLS residuals are  $\hat{e}_i = y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$ .

One motivation for the choice of NLLS as the estimation method is that the parameter  $\boldsymbol{\theta}$  is the solution to the population problem  $\min_{\boldsymbol{\theta}} \mathbb{E}(y_i - m(\mathbf{x}_i, \boldsymbol{\theta}))^2$

Since the criterion  $\hat{S}(\boldsymbol{\theta})$  is not quadratic,  $\hat{\boldsymbol{\theta}}$  must be found by numerical methods. See Appendix E. When  $m(\mathbf{x}, \boldsymbol{\theta})$  is differentiable, then the FOC for minimization are

$$\mathbf{0} = \sum_{i=1}^n \mathbf{m}_{\boldsymbol{\theta}}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) \hat{e}_i \quad (23.3)$$

where

$$\mathbf{m}_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} m(\mathbf{x}, \boldsymbol{\theta}).$$

**Theorem 23.1 Asymptotic Distribution of NLLS Estimator**

If the model is identified and  $m(\mathbf{x}, \boldsymbol{\theta})$  is differentiable with respect to  $\boldsymbol{\theta}$ ,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}_{\boldsymbol{\theta}})$$

$$\mathbf{V}_{\boldsymbol{\theta}} = (\mathbb{E}(\mathbf{m}_{\boldsymbol{\theta}} \mathbf{m}_{\boldsymbol{\theta}}'))^{-1} (\mathbb{E}(\mathbf{m}_{\boldsymbol{\theta}} \mathbf{m}_{\boldsymbol{\theta}}' e_i^2)) (\mathbb{E}(\mathbf{m}_{\boldsymbol{\theta}} \mathbf{m}_{\boldsymbol{\theta}}'))^{-1}$$

where  $\mathbf{m}_{\boldsymbol{\theta}} = \mathbf{m}_{\boldsymbol{\theta}}(\mathbf{x}_i, \boldsymbol{\theta}_0)$ .

Based on Theorem 23.1, an estimate of the asymptotic variance  $V_{\theta}$  is

$$\widehat{V}_{\theta} = \left( \frac{1}{n} \sum_{i=1}^n \widehat{\mathbf{m}}_{\theta i} \widehat{\mathbf{m}}'_{\theta i} \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \widehat{\mathbf{m}}_{\theta i} \widehat{\mathbf{m}}'_{\theta i} \widehat{e}_i^2 \right) \left( \frac{1}{n} \sum_{i=1}^n \widehat{\mathbf{m}}_{\theta i} \widehat{\mathbf{m}}'_{\theta i} \right)^{-1}$$

where  $\widehat{\mathbf{m}}_{\theta i} = \mathbf{m}_{\theta}(\mathbf{x}_i, \widehat{\theta})$  and  $\widehat{e}_i = y_i - m(\mathbf{x}_i, \widehat{\theta})$ .

Identification is often tricky in nonlinear regression models. Suppose that

$$m(\mathbf{x}_i, \theta) = \beta'_1 \mathbf{z}_i + \beta'_2 \mathbf{x}_i(\gamma)$$

where  $\mathbf{x}_i(\gamma)$  is a function of  $\mathbf{x}_i$  and the unknown parameter  $\gamma$ . Examples include  $x_i(\gamma) = x_i^\gamma$ ,  $x_i(\gamma) = \exp(\gamma x_i)$ , and  $x_i(\gamma) = x_i 1(g(x_i) > \gamma)$ . The model is linear when  $\beta_2 = \mathbf{0}$ , and this is often a useful hypothesis (sub-model) to consider. Thus we want to test

$$\mathbb{H}_0 : \beta_2 = \mathbf{0}.$$

However, under  $\mathbb{H}_0$ , the model is

$$y_i = \beta'_1 \mathbf{z}_i + e_i$$

and both  $\beta_2$  and  $\gamma$  have dropped out. This means that under  $\mathbb{H}_0$ ,  $\gamma$  is not identified. This renders the distribution theory presented in the previous section invalid. Thus when the truth is that  $\beta_2 = \mathbf{0}$ , the parameter estimates are not asymptotically normally distributed. Furthermore, tests of  $\mathbb{H}_0$  do not have asymptotic normal or chi-square distributions.

The asymptotic theory of such tests have been worked out by Andrews and Ploberger (1994) and B. E. Hansen (1996). In particular, Hansen shows how to use simulation (similar to the bootstrap) to construct the asymptotic critical values (or p-values) in a given application.

**Proof of Theorem 23.1 (Sketch).** NLLS estimation falls in the class of optimization estimators. For this theory, it is useful to denote the true value of the parameter  $\theta$  as  $\theta_0$ .

The first step is to show that  $\widehat{\theta} \xrightarrow{P} \theta_0$ . Proving that nonlinear estimators are consistent is more challenging than for linear estimators. We sketch the main argument. The idea is that  $\widehat{\theta}$  minimizes the sample criterion function  $\widehat{S}(\theta)$ , which (for any  $\theta$ ) converges in probability to the mean-squared error function  $\mathbb{E}((y_i - m(\mathbf{x}_i, \theta))^2)$ . Thus it seems reasonable that the minimizer  $\widehat{\theta}$  will converge in probability to  $\theta_0$ , the minimizer of  $\mathbb{E}((y_i - m(\mathbf{x}_i, \theta))^2)$ . It turns out that to show this rigorously, we need to show that  $\widehat{S}(\theta)$  converges *uniformly* to its expectation  $\mathbb{E}((y_i - m(\mathbf{x}_i, \theta))^2)$ , which means that the maximum discrepancy must converge in probability to zero, to exclude the possibility that  $\widehat{S}(\theta)$  is excessively wiggly in  $\theta$ . Proving uniform convergence is technically challenging, but it can be shown to hold broadly for relevant nonlinear regression models, especially if the regression function  $m(\mathbf{x}_i, \theta)$  is differentiable in  $\theta$ . For a complete treatment of the theory of optimization estimators see Newey and McFadden (1994).

Since  $\widehat{\theta} \xrightarrow{P} \theta_0$ ,  $\widehat{\theta}$  is close to  $\theta_0$  for  $n$  large, so the minimization of  $\widehat{S}(\theta)$  only needs to be examined for  $\theta$  close to  $\theta_0$ . Let

$$y_i^0 = e_i + \mathbf{m}'_{\theta i} \theta_0.$$

For  $\theta$  close to the true value  $\theta_0$ , by a first-order Taylor series approximation,

$$m(\mathbf{x}_i, \theta) \approx m(\mathbf{x}_i, \theta_0) + \mathbf{m}'_{\theta i} (\theta - \theta_0).$$

Thus

$$\begin{aligned} y_i - m(\mathbf{x}_i, \theta) &\approx (e_i + m(\mathbf{x}_i, \theta_0)) - (m(\mathbf{x}_i, \theta_0) + \mathbf{m}'_{\theta i} (\theta - \theta_0)) \\ &= e_i - \mathbf{m}'_{\theta i} (\theta - \theta_0) \\ &= y_i^0 - \mathbf{m}'_{\theta i} \theta. \end{aligned}$$

Hence the normalized sum of squared errors function is

$$\widehat{S}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n (y_i - m(\mathbf{x}_i, \boldsymbol{\theta}))^2 \simeq \frac{1}{n} \sum_{i=1}^n (y_i^0 - \mathbf{m}'_{\boldsymbol{\theta} i} \boldsymbol{\theta})^2$$

and the right-hand-side is the criterion function for a linear regression of  $y_i^0$  on  $\mathbf{m}_{\boldsymbol{\theta} i}$ . Thus the NLLS estimator  $\widehat{\boldsymbol{\theta}}$  has the same asymptotic distribution as the (infeasible) OLS regression of  $y_i^0$  on  $\mathbf{m}_{\boldsymbol{\theta} i}$ , which is that stated in the theorem.

---

### 23.3 Least Absolute Deviations

We stated that a conventional goal in econometrics is estimation of impact of variation in  $\mathbf{x}_i$  on the central tendency of  $y_i$ . We have discussed projections and conditional means, but these are not the only measures of central tendency. An alternative good measure is the conditional median.

To recall the definition and properties of the median, let  $y$  be a continuous random variable. The median  $\theta = \text{med}(y)$  is the value such that  $\mathbb{P}(y \leq \theta) = \mathbb{P}(y \geq \theta) = 0.5$ . Two useful facts about the median are that

$$\theta = \operatorname{argmin}_{\theta} \mathbb{E}|y - \theta| \quad (23.4)$$

and

$$\mathbb{E}(\operatorname{sgn}(y - \theta)) = 0$$

where

$$\operatorname{sgn}(u) = \begin{cases} 1 & \text{if } u \geq 0 \\ -1 & \text{if } u < 0 \end{cases}$$

is the sign function.

These facts and definitions motivate three estimators of  $\theta$ . The first definition is the 50th empirical quantile. The second is the value which minimizes  $\frac{1}{n} \sum_{i=1}^n |y_i - \theta|$ , and the third definition is the solution to the moment equation  $\frac{1}{n} \sum_{i=1}^n \operatorname{sgn}(y_i - \theta)$ . These distinctions are illusory, however, as these estimators are indeed identical.

Now let's consider the conditional median of  $y$  given a random vector  $\mathbf{x}$ . Let  $m(\mathbf{x}) = \text{med}(y | \mathbf{x})$  denote the conditional median of  $y$  given  $\mathbf{x}$ . The linear median regression model takes the form

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \text{med}(e_i | \mathbf{x}_i) &= 0 \end{aligned}$$

In this model, the linear function  $\text{med}(y_i | \mathbf{x}_i = \mathbf{x}) = \mathbf{x}' \boldsymbol{\beta}$  is the conditional median function, and the substantive assumption is that the median function is linear in  $\mathbf{x}$ .

Conditional analogs of the facts about the median are

- $\mathbb{P}(y_i \leq \mathbf{x}' \boldsymbol{\beta} | \mathbf{x}_i = \mathbf{x}) = \mathbb{P}(y_i > \mathbf{x}' \boldsymbol{\beta} | \mathbf{x}_i = \mathbf{x}) = .5$
- $\mathbb{E}(\operatorname{sgn}(e_i) | \mathbf{x}_i) = 0$
- $\mathbb{E}(\mathbf{x}'_i \operatorname{sgn}(e_i)) = 0$
- $\boldsymbol{\beta} = \min_{\boldsymbol{\beta}} \mathbb{E}|y_i - \mathbf{x}'_i \boldsymbol{\beta}|$

These facts motivate the following estimator. Let

$$LAD(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n |y_i - \mathbf{x}'_i \boldsymbol{\beta}|$$

be the average of absolute deviations. The **least absolute deviations** (LAD) estimator of  $\beta$  minimizes this function

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} LAD(\beta)$$

Equivalently, it is a solution to the moment condition

$$\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \operatorname{sgn}(y_i - \mathbf{x}'_i \hat{\beta}) = 0. \quad (23.5)$$

The LAD estimator has an asymptotic normal distribution.

**Theorem 23.2 Asymptotic Distribution of LAD Estimator**

When the conditional median is linear in  $\mathbf{x}$

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(\mathbf{0}, V)$$

where

$$V = \frac{1}{4} (\mathbb{E}(\mathbf{x}_i \mathbf{x}'_i f(0 | \mathbf{x}_i)))^{-1} (\mathbb{E}(\mathbf{x}_i \mathbf{x}'_i)) (\mathbb{E}(\mathbf{x}_i \mathbf{x}'_i f(0 | \mathbf{x}_i)))^{-1}$$

and  $f(e | \mathbf{x})$  is the conditional density of  $e_i$  given  $\mathbf{x}_i = \mathbf{x}$ .

The variance of the asymptotic distribution inversely depends on  $f(0 | \mathbf{x})$ , the conditional density of the error at its median. When  $f(0 | \mathbf{x})$  is large, then there are many innovations near to the median, and this improves estimation of the median. In the special case where the error is independent of  $\mathbf{x}_i$ , then  $f(0 | \mathbf{x}) = f(0)$  and the asymptotic variance simplifies

$$V = \frac{(\mathbb{E}(\mathbf{x}_i \mathbf{x}'_i))^{-1}}{4f(0)^2} \quad (23.6)$$

This simplification is similar to the simplification of the asymptotic covariance of the OLS estimator under homoskedasticity.

Computation of standard error for LAD estimates typically is based on equation (23.6). The main difficulty is the estimation of  $f(0)$ , the height of the error density at its median. This can be done with kernel estimation techniques. See Chapter 19. While a complete proof of Theorem 23.2 is advanced, we provide a sketch here for completeness.

---

**Proof of Theorem 23.2:** Similar to NLLS, LAD is an optimization estimator. Let  $\beta_0$  denote the true value of  $\beta$ .

The first step is to show that  $\hat{\beta} \xrightarrow{p} \beta_0$ . The general nature of the proof is similar to that for the NLLS estimator, and is sketched here. For any fixed  $\beta$ , by the WLLN,  $LAD(\beta) \xrightarrow{p} \mathbb{E}|y_i - \mathbf{x}'_i \beta|$ . Furthermore, it can be shown that this convergence is uniform in  $\beta$ . (Proving uniform convergence is more challenging than for the NLLS criterion since the LAD criterion is not differentiable in  $\beta$ .) It follows that  $\hat{\beta}$ , the minimizer of  $LAD(\beta)$ , converges in probability to  $\beta_0$ , the minimizer of  $\mathbb{E}|y_i - \mathbf{x}'_i \beta|$ .

Since  $\operatorname{sgn}(a) = 1 - 2 \cdot \mathbf{1}(a \leq 0)$ , (23.5) is equivalent to  $\bar{\mathbf{g}}_n(\hat{\beta}) = 0$ , where  $\bar{\mathbf{g}}_n(\beta) = n^{-1} \sum_{i=1}^n \mathbf{g}_i(\beta)$  and  $\mathbf{g}_i(\beta) = \mathbf{x}_i (1 - 2 \cdot \mathbf{1}(y_i \leq \mathbf{x}'_i \beta))$ . Let  $\mathbf{g}(\beta) = \mathbb{E}(\mathbf{g}_i(\beta))$ . We need three preliminary results. First, since  $\mathbb{E}(\mathbf{g}_i(\beta_0)) = 0$  and  $\mathbb{E}(\mathbf{g}_i(\beta_0) \mathbf{g}_i(\beta_0)') = \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i)$ , we can apply the central limit theorem (Theorem 6.11) and find that

$$\sqrt{n} \bar{\mathbf{g}}_n(\beta_0) = n^{-1/2} \sum_{i=1}^n \mathbf{g}_i(\beta_0) \xrightarrow{d} N(\mathbf{0}, \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i)).$$

Second using the law of iterated expectations and the chain rule of differentiation,

$$\begin{aligned}\frac{\partial}{\partial \beta'} \mathbf{g}(\boldsymbol{\beta}) &= \frac{\partial}{\partial \beta'} \mathbb{E}_{\mathbf{x}_i} (1 - 2 \cdot \mathbf{1}(y_i \leq \mathbf{x}'_i \boldsymbol{\beta})) \\ &= -2 \frac{\partial}{\partial \beta'} \mathbb{E}(\mathbf{x}_i \mathbb{E}(1(e_i \leq \mathbf{x}'_i \boldsymbol{\beta} - \mathbf{x}'_i \boldsymbol{\beta}_0) | \mathbf{x}_i)) \\ &= -2 \frac{\partial}{\partial \beta'} \mathbb{E}\left(\mathbf{x}_i \int_{-\infty}^{\mathbf{x}'_i \boldsymbol{\beta} - \mathbf{x}'_i \boldsymbol{\beta}_0} f(e | \mathbf{x}_i) de\right) \\ &= -2 \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i f(\mathbf{x}'_i \boldsymbol{\beta} - \mathbf{x}'_i \boldsymbol{\beta}_0 | \mathbf{x}_i))\end{aligned}$$

so

$$\frac{\partial}{\partial \beta'} \mathbf{g}(\boldsymbol{\beta}) = -2 \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i f(0 | \mathbf{x}_i)).$$

Third, by a Taylor series expansion and the fact  $\mathbf{g}(\boldsymbol{\beta}) = 0$

$$\mathbf{g}(\hat{\boldsymbol{\beta}}) \approx \frac{\partial}{\partial \beta'} \mathbf{g}(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

Together

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &\approx \left(\frac{\partial}{\partial \beta'} \mathbf{g}(\boldsymbol{\beta}_0)\right)^{-1} \sqrt{n} \mathbf{g}(\hat{\boldsymbol{\beta}}) \\ &= (-2 \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i f(0 | \mathbf{x}_i)))^{-1} \sqrt{n} (\mathbf{g}(\hat{\boldsymbol{\beta}}) - \bar{\mathbf{g}}_n(\hat{\boldsymbol{\beta}})) \\ &\approx \frac{1}{2} (\mathbb{E}(\mathbf{x}_i \mathbf{x}'_i f(0 | \mathbf{x}_i)))^{-1} \sqrt{n} (\bar{\mathbf{g}}_n(\boldsymbol{\beta}_0) - \mathbf{g}(\boldsymbol{\beta}_0)) \\ &\xrightarrow{d} \frac{1}{2} (\mathbb{E}[\mathbf{x}_i \mathbf{x}'_i f(0 | \mathbf{x}_i)])^{-1} \mathcal{N}(\mathbf{0}, \mathbb{E}(\mathbf{x}_i \mathbf{x}'_i)) \\ &= \mathcal{N}(\mathbf{0}, \mathbf{V}).\end{aligned}$$

The third line follows from an asymptotic empirical process argument and the fact that  $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ .

---

## 23.4 Quantile Regression

Quantile regression has become quite popular in recent econometric practice. For  $\tau \in [0, 1]$  the  $\tau^{th}$  quantile  $Q_\tau$  of a random variable with distribution function  $F(u)$  is defined as

$$Q_\tau = \inf\{u : F(u) \geq \tau\}$$

When  $F(u)$  is continuous and strictly monotonic, then  $F(Q_\tau) = \tau$ , so you can think of the quantile as the inverse of the distribution function. The quantile  $Q_\tau$  is the value such that  $\tau$  (percent) of the mass of the distribution is less than  $Q_\tau$ . The median is the special case  $\tau = .5$ .

The following alternative representation is useful. If the random variable  $U$  has  $\tau^{th}$  quantile  $Q_\tau$ , then

$$Q_\tau = \operatorname{argmin}_\theta \mathbb{E}(\rho_\tau(U - \theta)). \quad (23.7)$$

where  $\rho_\tau(q)$  is the piecewise linear function

$$\begin{aligned}\rho_\tau(q) &= \begin{cases} -q(1 - \tau) & q < 0 \\ q\tau & q \geq 0 \end{cases} \\ &= q(\tau - \mathbf{1}(q < 0)).\end{aligned} \quad (23.8)$$

This generalizes representation (23.4) for the median to all quantiles.

For the random variables  $(y_i, \mathbf{x}_i)$  with conditional distribution function  $F(y | \mathbf{x})$  the conditional quantile function  $q_\tau(\mathbf{x})$  is

$$Q_\tau(\mathbf{x}) = \inf \{y : F(y | \mathbf{x}) \geq \tau\}.$$

Again, when  $F(y | \mathbf{x})$  is continuous and strictly monotonic in  $y$ , then  $F(Q_\tau(\mathbf{x}) | \mathbf{x}) = \tau$ . For fixed  $\tau$ , the quantile regression function  $q_\tau(\mathbf{x})$  describes how the  $\tau^{th}$  quantile of the conditional distribution varies with the regressors.

As functions of  $\mathbf{x}$ , the quantile regression functions can take any shape. However for computational convenience it is typical to assume that they are (approximately) linear in  $\mathbf{x}$  (after suitable transformations). This linear specification assumes that  $Q_\tau(\mathbf{x}) = \boldsymbol{\beta}'_\tau \mathbf{x}$  where the coefficients  $\boldsymbol{\beta}_\tau$  vary across the quantiles  $\tau$ . We then have the linear quantile regression model

$$y_i = \mathbf{x}'_i \boldsymbol{\beta}_\tau + e_i$$

where  $e_i$  is the error defined to be the difference between  $y_i$  and its  $\tau^{th}$  conditional quantile  $\mathbf{x}'_i \boldsymbol{\beta}_\tau$ . By construction, the  $\tau^{th}$  conditional quantile of  $e_i$  is zero, otherwise its properties are unspecified without further restrictions.

Given the representation (23.7), the quantile regression estimator  $\hat{\boldsymbol{\beta}}_\tau$  for  $\boldsymbol{\beta}_\tau$  solves the minimization problem

$$\hat{\boldsymbol{\beta}}_\tau = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} S^\tau(\boldsymbol{\beta})$$

where

$$S^\tau(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}'_i \boldsymbol{\beta})$$

and  $\rho_\tau(q)$  is defined in (23.8).

Since the quantile regression criterion function  $S^\tau(\boldsymbol{\beta})$  does not have an algebraic solution, numerical methods are necessary for its minimization. Furthermore, since it has discontinuous derivatives, conventional Newton-type optimization methods are inappropriate. Fortunately, fast linear programming methods have been developed for this problem, and are widely available.

An asymptotic distribution theory for the quantile regression estimator can be derived using similar arguments as those for the LAD estimator in Theorem 23.2.

**Theorem 23.3 Asymptotic Distribution of the Quantile Regression Estimator**

When the  $\tau^{th}$  conditional quantile is linear in  $\mathbf{x}$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_\tau),$$

where

$$\mathbf{V}_\tau = \tau(1-\tau)(\mathbb{E}(\mathbf{x}_i \mathbf{x}'_i f(0 | \mathbf{x}_i)))^{-1} (\mathbb{E}(\mathbf{x}_i \mathbf{x}'_i)) (\mathbb{E}(\mathbf{x}_i \mathbf{x}'_i f(0 | \mathbf{x}_i)))^{-1}$$

and  $f(e | \mathbf{x})$  is the conditional density of  $e_i$  given  $\mathbf{x}_i = \mathbf{x}$ .

In general, the asymptotic variance depends on the conditional density of the quantile regression error. When the error  $e_i$  is independent of  $\mathbf{x}_i$ , then  $f(0 | \mathbf{x}_i) = f(0)$ , the unconditional density of  $e_i$  at 0, and we have the simplification

$$\mathbf{V}_\tau = \frac{\tau(1-\tau)}{f(0)^2} (\mathbb{E}(\mathbf{x}_i \mathbf{x}'_i))^{-1}.$$

An excellent monograph on quantile regression is Koenker (2005).

## 23.5 Limited Dependent Variables

$y$  is a **limited dependent variable** if it takes values in a strict subset of  $\mathbb{R}$ . The most common cases are

- Binary:  $y \in \{0, 1\}$
- Multinomial:  $y \in \{0, 1, 2, \dots, k\}$
- Integer:  $y \in \{0, 1, 2, \dots\}$
- Censored:  $y \in \mathbb{R}^+$

The traditional approach to the estimation of limited dependent variable (LDV) models is parametric maximum likelihood. A parametric model is constructed, allowing the construction of the likelihood function. A more modern approach is semi-parametric, eliminating the dependence on a parametric distributional assumption. We will discuss only the first (parametric) approach, due to time constraints. They still constitute the majority of LDV applications. If, however, you were to write a thesis involving LDV estimation, you would be advised to consider employing a semi-parametric estimation approach.

For the parametric approach, estimation is by MLE. A major practical issue is construction of the likelihood function.

## 23.6 Binary Choice

The dependent variable  $y_i \in \{0, 1\}$ . This represents a Yes/No outcome. Given some regressors  $\mathbf{x}_i$ , the goal is to describe  $\mathbb{P}(y_i = 1 | \mathbf{x}_i)$ , as this is the full conditional distribution.

The linear probability model specifies that

$$\mathbb{P}(y_i = 1 | \mathbf{x}_i) = \mathbf{x}'_i \boldsymbol{\beta}.$$

As  $\mathbb{P}(y_i = 1 | \mathbf{x}_i) = \mathbb{E}(y_i | \mathbf{x}_i)$ , this yields the regression:  $y_i = \mathbf{x}'_i \boldsymbol{\beta} + e_i$  which can be estimated by OLS. However, the linear probability model does not impose the restriction that  $0 \leq \mathbb{P}(y_i | \mathbf{x}_i) \leq 1$ . Even so estimation of a linear probability model is a useful starting point for subsequent analysis.

The standard alternative is to use a function of the form

$$\mathbb{P}(y_i = 1 | \mathbf{x}_i) = F(\mathbf{x}'_i \boldsymbol{\beta})$$

where  $F(\cdot)$  is a known CDF, typically assumed to be symmetric about zero, so that  $F(u) = 1 - F(-u)$ . The two standard choices for  $F$  are

- Logistic:  $F(u) = (1 + e^{-u})^{-1}$ .
- Normal:  $F(u) = \Phi(u)$ .

If  $F$  is logistic, we call this the **logit** model, and if  $F$  is normal, we call this the **probit** model.

This model is identical to the latent variable model

$$\begin{aligned} y_i^* &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ e_i &\sim F(\cdot) \\ y_i &= \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases}. \end{aligned}$$

For then

$$\begin{aligned}\mathbb{P}(y_i = 1 \mid \mathbf{x}_i) &= \mathbb{P}(y_i^* > 0 \mid \mathbf{x}_i) \\ &= \mathbb{P}(\mathbf{x}'_i \boldsymbol{\beta} + e_i > 0 \mid \mathbf{x}_i) \\ &= \mathbb{P}(e_i > -\mathbf{x}'_i \boldsymbol{\beta} \mid \mathbf{x}_i) \\ &= 1 - F(-\mathbf{x}'_i \boldsymbol{\beta}) \\ &= F(\mathbf{x}'_i \boldsymbol{\beta}).\end{aligned}$$

Estimation is by maximum likelihood. To construct the likelihood, we need the conditional distribution of an individual observation. Recall that if  $y$  is Bernoulli, such that  $\mathbb{P}(y = 1) = p$  and  $\mathbb{P}(y = 0) = 1 - p$ , then we can write the density of  $y$  as

$$f(y) = p^y(1-p)^{1-y}, \quad y = 0, 1.$$

In the Binary choice model,  $y_i$  is conditionally Bernoulli with  $\mathbb{P}(y_i = 1 \mid \mathbf{x}_i) = p_i = F(\mathbf{x}'_i \boldsymbol{\beta})$ . Thus the conditional density is

$$\begin{aligned}f(y_i \mid \mathbf{x}_i) &= p_i^{y_i}(1-p_i)^{1-y_i} \\ &= F(\mathbf{x}'_i \boldsymbol{\beta})^{y_i}(1-F(\mathbf{x}'_i \boldsymbol{\beta}))^{1-y_i}.\end{aligned}$$

Hence the log-likelihood function is

$$\begin{aligned}\log L(\boldsymbol{\beta}) &= \sum_{i=1}^n \log f(y_i \mid \mathbf{x}_i) \\ &= \sum_{i=1}^n \log(F(\mathbf{x}'_i \boldsymbol{\beta})^{y_i}(1-F(\mathbf{x}'_i \boldsymbol{\beta}))^{1-y_i}) \\ &= \sum_{i=1}^n [y_i \log F(\mathbf{x}'_i \boldsymbol{\beta}) + (1-y_i) \log(1-F(\mathbf{x}'_i \boldsymbol{\beta}))] \\ &= \sum_{y_i=1} \log F(\mathbf{x}'_i \boldsymbol{\beta}) + \sum_{y_i=0} \log(1-F(\mathbf{x}'_i \boldsymbol{\beta})).\end{aligned}$$

The MLE  $\hat{\boldsymbol{\beta}}$  is the value of  $\boldsymbol{\beta}$  which maximizes  $\log L(\boldsymbol{\beta})$ . Standard errors and test statistics are computed by asymptotic approximations. Details of such calculations are left to more advanced courses.

## 23.7 Count Data

If  $y \in \{0, 1, 2, \dots\}$ , a typical approach is to employ **Poisson regression**. This model specifies that

$$\begin{aligned}\mathbb{P}(y_i = k \mid \mathbf{x}_i) &= \frac{\exp(-\lambda_i) \lambda_i^k}{k!}, \quad k = 0, 1, 2, \dots \\ \lambda_i &= \exp(\mathbf{x}'_i \boldsymbol{\beta}).\end{aligned}$$

The conditional density is the Poisson with parameter  $\lambda_i$ . The functional form for  $\lambda_i$  has been picked to ensure that  $\lambda_i > 0$ .

The log-likelihood function is

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^n \log f(y_i \mid \mathbf{x}_i) = \sum_{i=1}^n (-\exp(\mathbf{x}'_i \boldsymbol{\beta}) + y_i \mathbf{x}'_i \boldsymbol{\beta} - \log(y_i!)).$$

The MLE is the value  $\hat{\boldsymbol{\beta}}$  which maximizes  $\log L(\boldsymbol{\beta})$ .

Since

$$\mathbb{E}(y_i \mid \mathbf{x}_i) = \lambda_i = \exp(\mathbf{x}'_i \boldsymbol{\beta})$$

is the conditional mean, this motivates the label Poisson “regression.”

Also observe that the model implies that

$$\text{var}(y_i | \mathbf{x}_i) = \lambda_i = \exp(\mathbf{x}'_i \boldsymbol{\beta}),$$

so the model imposes the restriction that the conditional mean and variance of  $y_i$  are the same. This may be considered restrictive. A generalization is the negative binomial.

## 23.8 Censored Data

The idea of **censoring** is that some data above or below a threshold are mis-reported at the threshold. Thus the model is that there is some latent process  $y_i^*$  with unbounded support, but we observe only

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* \geq 0 \\ 0 & \text{if } y_i^* < 0 \end{cases}. \quad (23.9)$$

(This is written for the case of the threshold being zero, any known value can substitute.) The observed data  $y_i$  therefore come from a mixed continuous/discrete distribution.

Censored models are typically applied when the data set has a meaningful proportion (say 5% or higher) of data at the boundary of the sample support. The censoring process may be explicit in data collection, or it may be a by-product of economic constraints.

An example of a data collection censoring is top-coding of income. In surveys, incomes above a threshold are typically reported at the threshold.

The first censored regression model was developed by Tobin (1958) to explain consumption of durable goods. Tobin observed that for many households, the consumption level (purchases) in a particular period was zero. He proposed the latent variable model

$$\begin{aligned} y_i^* &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ e_i &\stackrel{iid}{\sim} N(0, \sigma^2) \end{aligned}$$

with the observed variable  $y_i$  generated by the censoring equation (23.9). This model (now called the Tobit) specifies that the latent (or ideal) value of consumption may be negative (the household would prefer to sell than buy). All that is reported is that the household purchased zero units of the good.

The naive approach to estimate  $\boldsymbol{\beta}$  is to regress  $y_i$  on  $\mathbf{x}_i$ . This does not work because regression estimates  $\mathbb{E}(y_i | \mathbf{x}_i)$ , not  $\mathbb{E}(y_i^* | \mathbf{x}_i) = \mathbf{x}'_i \boldsymbol{\beta}$ , and the latter is of interest. Thus OLS will be biased for the parameter of interest  $\boldsymbol{\beta}$ .

[Note: it is still possible to estimate  $\mathbb{E}(y_i | \mathbf{x}_i)$  by LS techniques. The Tobit framework postulates that this is not inherently interesting, that the parameter of  $\boldsymbol{\beta}$  is defined by an alternative statistical structure.]

Consistent estimation will be achieved by the MLE. To construct the likelihood, observe that the probability of being censored is

$$\begin{aligned} \mathbb{P}(y_i = 0 | \mathbf{x}_i) &= \mathbb{P}(y_i^* < 0 | \mathbf{x}_i) \\ &= \mathbb{P}(\mathbf{x}'_i \boldsymbol{\beta} + e_i < 0 | \mathbf{x}_i) \\ &= \mathbb{P}\left(\frac{e_i}{\sigma} < -\frac{\mathbf{x}'_i \boldsymbol{\beta}}{\sigma} | \mathbf{x}_i\right) \\ &= \Phi\left(-\frac{\mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right). \end{aligned}$$

The conditional density function above zero is normal:

$$\sigma^{-1} \phi\left(\frac{y - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right), \quad y > 0.$$

Therefore, the density function for  $y \geq 0$  can be written as

$$f(y | \mathbf{x}_i) = \Phi\left(-\frac{\mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right)^{1(y=0)} \left[\sigma^{-1} \phi\left(\frac{z - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right)\right]^{1(y>0)},$$

where  $1(\cdot)$  is the indicator function.

Hence the log-likelihood is a mixture of the probit and the normal:

$$\begin{aligned} \log L(\boldsymbol{\beta}) &= \sum_{i=1}^n \log f(y_i | \mathbf{x}_i) \\ &= \sum_{y_i=0} \log \Phi\left(-\frac{\mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right) + \sum_{y_i>0} \log \left[\sigma^{-1} \phi\left(\frac{y_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right)\right]. \end{aligned}$$

The MLE is the value  $\hat{\boldsymbol{\beta}}$  which maximizes  $\log L(\boldsymbol{\beta})$ .

## 23.9 Sample Selection

The problem of sample selection arises when the sample is a non-random selection of potential observations. This occurs when the observed data is systematically different from the population of interest. For example, if you ask for volunteers for an experiment, and they wish to extrapolate the effects of the experiment on a general population, you should worry that the people who volunteer may be systematically different from the general population. This has great relevance for the evaluation of anti-poverty and job-training programs, where the goal is to assess the effect of “training” on the general population, not just on the volunteers.

A simple sample selection model can be written as the latent model

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_{1i} \\ T_i &= \mathbf{1}(\mathbf{z}'_i \boldsymbol{\gamma} + e_{0i} > 0). \end{aligned}$$

The dependent variable  $y_i$  is observed if (and only if)  $T_i = 1$ . Else it is unobserved.

For example,  $y_i$  could be a wage, which can be observed only if a person is employed. The equation for  $T_i$  is an equation specifying the probability that the person is employed.

The model is often completed by specifying that the errors are jointly normal

$$\begin{pmatrix} e_{0i} \\ e_{1i} \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & \sigma^2 \end{pmatrix}\right).$$

It is presumed that we observe  $\{\mathbf{x}_i, \mathbf{z}_i, T_i\}$  for all observations.

Under the normality assumption,

$$e_{1i} = \rho e_{0i} + v_i,$$

where  $v_i$  is independent of  $e_{0i} \sim N(0, 1)$ . A useful fact about the standard normal distribution is that

$$\mathbb{E}(e_{0i} | e_{0i} > -x) = \lambda(x) = \frac{\phi(x)}{\Phi(x)},$$

and the function  $\lambda(x)$  is called the inverse Mills ratio.

The naive estimator of  $\boldsymbol{\beta}$  is OLS regression of  $y_i$  on  $\mathbf{x}_i$  for those observations for which  $y_i$  is available. The problem is that this is equivalent to conditioning on the event  $\{T_i = 1\}$ . However,

$$\begin{aligned} \mathbb{E}(e_{1i} | T_i = 1, \mathbf{z}_i) &= \mathbb{E}(e_{1i} | \{e_{0i} > -\mathbf{z}'_i \boldsymbol{\gamma}\}, \mathbf{z}_i) \\ &= \rho \mathbb{E}(e_{0i} | \{e_{0i} > -\mathbf{z}'_i \boldsymbol{\gamma}\}, \mathbf{z}_i) + \mathbb{E}(v_i | \{e_{0i} > -\mathbf{z}'_i \boldsymbol{\gamma}\}, \mathbf{z}_i) \\ &= \rho \lambda(\mathbf{z}'_i \boldsymbol{\gamma}), \end{aligned}$$

which is non-zero. Thus

$$e_{1i} = \rho \lambda(\mathbf{z}'_i \boldsymbol{\gamma}) + u_i,$$

where

$$\mathbb{E}(u_i | T_i = 1, \mathbf{z}_i) = 0.$$

Hence

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \rho \lambda(\mathbf{z}'_i \boldsymbol{\gamma}) + u_i \quad (23.10)$$

is a valid regression equation for the observations for which  $T_i = 1$ .

Heckman (1979) observed that we could consistently estimate  $\boldsymbol{\beta}$  and  $\rho$  from this equation, if  $\boldsymbol{\gamma}$  were known. It is unknown, but also can be consistently estimated by a Probit model for selection. The “Heckit” estimator is thus calculated as follows

- Estimate  $\hat{\boldsymbol{\gamma}}$  from a Probit, using regressors  $\mathbf{z}_i$ . The binary dependent variable is  $T_i$ .
- Estimate  $(\hat{\boldsymbol{\beta}}, \hat{\rho})$  from OLS of  $y_i$  on  $\mathbf{x}_i$  and  $\lambda(\mathbf{z}'_i \hat{\boldsymbol{\gamma}})$ .
- The OLS standard errors will be incorrect, as this is a two-step estimator. They can be corrected using a more complicated formula. Or, alternatively, by viewing the Probit/OLS estimation equations as a large joint GMM problem.

The Heckit estimator is frequently used to deal with problems of sample selection. However, the estimator is built on the assumption of normality, and the estimator can be quite sensitive to this assumption. Some modern econometric research is exploring how to relax the normality assumption.

The estimator can also work quite poorly if  $\lambda(\mathbf{z}'_i \hat{\boldsymbol{\gamma}})$  does not have much in-sample variation. This can happen if the Probit equation does not “explain” much about the selection choice. Another potential problem is that if  $\mathbf{z}_i = \mathbf{x}_i$ , then  $\lambda(\mathbf{z}'_i \hat{\boldsymbol{\gamma}})$  can be highly collinear with  $\mathbf{x}_i$ , so the second step OLS estimator will not be able to precisely estimate  $\boldsymbol{\beta}$ . Based this observation, it is typically recommended to find a valid exclusion restriction: a variable should be in  $\mathbf{z}_i$  which is not in  $\mathbf{x}_i$ . If this is valid, it will ensure that  $\lambda(\mathbf{z}'_i \hat{\boldsymbol{\gamma}})$  is not collinear with  $\mathbf{x}_i$ , and hence improve the second stage estimator’s precision.

## Exercises

**Exercise 23.1** Suppose that  $y_i = g(\mathbf{x}_i, \boldsymbol{\theta}) + e_i$  with  $\mathbb{E}(e_i | \mathbf{x}_i) = 0$ ,  $\hat{\boldsymbol{\theta}}$  is the NLLS estimator, and  $\hat{V}$  is the estimate of  $\text{var}(\hat{\boldsymbol{\theta}})$ . You are interested in the conditional mean function  $\mathbb{E}(y_i | \mathbf{x}_i = \mathbf{x}) = g(\mathbf{x})$  at some  $\mathbf{x}$ . Find an asymptotic 95% confidence interval for  $g(\mathbf{x})$ .

**Exercise 23.2** In Exercise 9.26, you estimated a cost function on a cross-section of electric companies. The equation you estimated was

$$\log TC_i = \beta_1 + \beta_2 \log Q_i + \beta_3 \log PL_i + \beta_4 \log PK_i + \beta_5 \log PF_i + e_i. \quad (23.11)$$

- (a) Following Nerlove, add the variable  $(\log Q_i)^2$  to the regression. Do so. Assess the merits of this new specification using a hypothesis test. Do you agree with this modification?
- (b) Now try a non-linear specification. Consider model (23.11) plus the extra term  $\beta_6 z_i$ , where

$$z_i = \log Q_i (1 + \exp(-(\log Q_i - \beta_7)))^{-1}.$$

In addition, impose the restriction  $\beta_3 + \beta_4 + \beta_5 = 1$ . This model is called a smooth threshold model. For values of  $\log Q_i$  much below  $\beta_7$ , the variable  $\log Q_i$  has a regression slope of  $\beta_2$ . For values much above  $\beta_7$ , the regression slope is  $\beta_2 + \beta_6$ , and the model imposes a smooth transition between these regimes. The model is non-linear because of the parameter  $\beta_7$ .

The model works best when  $\beta_7$  is selected so that several values (in this example, at least 10 to 15) of  $\log Q_i$  are both below and above  $\beta_7$ . Examine the data and pick an appropriate range for  $\beta_7$ .

- (c) Estimate the model by non-linear least squares. I recommend the concentration method: Pick 10 (or more if you like) values of  $\beta_7$  in this range. For each value of  $\beta_7$ , calculate  $z_i$  and estimate the model by OLS. Record the sum of squared errors, and find the value of  $\beta_7$  for which the sum of squared errors is minimized.
- (d) Calculate standard errors for all the parameters  $(\beta_1, \dots, \beta_7)$ .

**Exercise 23.3** For any predictor  $g(\mathbf{x}_i)$  for  $y_i$ , the mean absolute error (MAE) is

$$\mathbb{E}|y_i - g(\mathbf{x}_i)|.$$

Show that the function  $g(\mathbf{x})$  which minimizes the MAE is the conditional median  $m(\mathbf{x}) = \text{med}(y_i | \mathbf{x}_i)$ .

**Exercise 23.4** Define

$$g(u) = \tau - 1(u < 0)$$

where  $1(\cdot)$  is the indicator function (takes the value 1 if the argument is true, else equals zero). Let  $\theta$  satisfy  $\mathbb{E}(g(y_i - \theta)) = 0$ . Is  $\theta$  a quantile of the distribution of  $y_i$ ?

**Exercise 23.5** Verify equation (23.7)

**Exercise 23.6** You are interested in estimating the equation  $y_i = \mathbf{x}'_i \boldsymbol{\beta} + e_i$ . You believe the regressors are exogenous, but you are uncertain about the properties of the error. You estimate the equation both by least absolute deviations (LAD) and OLS. A colleague suggests that you should prefer the OLS estimate, because it produces a higher  $R^2$  than the LAD estimate. Is your colleague correct?

**Exercise 23.7** Your model is

$$\begin{aligned} y_i^* &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(e_i | \mathbf{x}_i) &= 0. \end{aligned}$$

However,  $y_i^*$  is not observed. Instead only a capped version is reported. That is, the dataset contains the variable

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* \leq \tau \\ \tau & \text{if } y_i^* > \tau \end{cases}$$

Suppose you regress  $y_i$  on  $x_i$  using OLS. Is OLS consistent for  $\beta$ ? Describe the nature of the effect of the mis-measured observation on the OLS estimate.

**Exercise 23.8** Take the model

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ \mathbb{E}(e_i | \mathbf{x}_i) &= 0. \end{aligned}$$

Let  $\hat{\boldsymbol{\beta}}$  denote the OLS estimator for  $\boldsymbol{\beta}$  based on an available sample.

- (a) Suppose that the  $i^{th}$  observation is in the sample only if  $x_{1i} > 0$ , where  $x_{1i}$  is an element of  $\mathbf{x}_i$ . Assume  $\mathbb{P}(x_{1i} < 0) > 0$ .

i Is  $\hat{\boldsymbol{\beta}}$  consistent for  $\boldsymbol{\beta}$ ?

ii If not, can you obtain an expression for its probability limit?

(For this, you may assume that  $e_i$  is independent of  $\mathbf{x}_i$  and  $N(0, \sigma^2)$ .)

- (b) Suppose that the  $i^{th}$  observation is in the sample only if  $y_i > 0$ .

i Is  $\hat{\boldsymbol{\beta}}$  consistent for  $\boldsymbol{\beta}$ ?

ii If not, can you obtain an expression for its probability limit?

(For this, you may assume that  $e_i$  is independent of  $\mathbf{x}_i$  and  $N(0, \sigma^2)$ .)

**Exercise 23.9** The Tobit model is

$$\begin{aligned} y_i^* &= \mathbf{x}'_i \boldsymbol{\beta} + e_i \\ e_i &\sim N(0, \sigma^2) \\ y_i &= y_i^* \mathbf{1}(y_i^* \geq 0) \end{aligned}$$

where  $\mathbf{1}(\cdot)$  is the indicator function.

- (a) Find  $\mathbb{E}(y_i | \mathbf{x}_i)$ .

Note: You may use the fact that since  $e_i \sim N(0, \sigma^2)$ ,

$$\mathbb{E}(e_i \mathbf{1}(e_i \geq -u)) = \sigma \lambda(u/\sigma) = \sigma \phi(u/\sigma)/\Phi(u/\sigma).$$

- (b) Use the result from part (a) to suggest a NLLS estimator for the parameter  $\beta$  given a sample  $\{y_i, \mathbf{x}_i\}$ .

**Exercise 23.10** A latent variable  $y_i^*$  is generated by

$$y_i^* = x_i \beta + e_i$$

The distribution of  $e_i$ , conditional on  $x_i$ , is  $N(0, \sigma_i^2)$ , where  $\sigma_i^2 = \gamma_0 + x_i^2 \gamma_1$  with  $\gamma_0 > 0$  and  $\gamma_1 > 0$ . The binary variable  $y_i$  equals 1 if  $y_i^* \geq 0$ , else  $y_i = 0$ . Find the log-likelihood function for the conditional distribution of  $y_i$  given  $x_i$  (the parameters are  $\beta, \gamma_0, \gamma_1$ ).

# Chapter 24

# Machine Learning

## 24.1 Introduction

This chapter reviews machine learning methods for econometrics. The term “machine learning” is a new and somewhat vague term, but typically is taken to mean procedures which are primarily used for point prediction in settings with unknown structure. Machine learning methods generally allow for large sample sizes, large number of variables, and unknown structural form.

The chapter reviews methods for model selection, James-Stein shrinkage, model averaging, ridge regression, LASSO, elastic net, regression trees, bagging, random forests, and ensembling. The chapter is preliminary, with the latter material incomplete and only briefly sketched.

Model selection is a tool for selecting one model (or estimator) out of a set of models. Different model selection methods are distinguished by the criteria used to rank and compare models.

Model averaging is a generalization of model selection. Models and estimators are averaged using data-dependent weights.

James-Stein shrinkage modifies classical estimators by shrinking towards a reasonable target. Shrinking reduces mean squared error.

Penalization methods add a parameterization penalty to a traditional criterion such as the sum of squared errors. The resulting estimators can have characteristics similar to model selection and shrinkage estimators. Penalization techniques can be applied even when the number of parameters is much greater than the sample size. A quadratic penalty produces Ridge regression. An  $L_1$  penalty produces the Lasso.

Two excellent monographs on model selection and averaging are Burnham and Anderson (1998) and Claeskens and Hjort (2008). James-Stein shrinkage theory is thoroughly covered in Lehmann and Casella (1998). See also Efron (2010) and Wasserman (2006). For penalization methods a classic reference is Hastie, Tibshirani, and Friedman (2008). Introductory treatments include James, Witten, Hastie, and Tibshirani (2013) and Efron and Hastie (2017).

This chapter is preliminary.

## 24.2 Model Selection

In the course of an applied project, an economist will routinely estimate multiple models. Indeed, most applied papers include tables displaying the results from different specifications. The question arises: Which model is best? Which should be used in practice? How can we select the best choice? This is the question of model selection.

Take, for example, a wage regression. Suppose we want a regression model which conditions on education, experience, region, and marital status. How should we proceed? Should we estimate a simple linear model plus a quadratic in experience? Should education enter linearly, a simple spline as in section ?, or with separate dummies for each education level? Should marital status enter as a simple dummy

(married or not) or allowing for all recorded categories? Should interactions be included? Which? How many? That is, we need to select the specific regressors to include in the regression model.

Model selection may be mis-named. It would be more appropriate to call the issue “estimator selection”. When we examine a table containing the results from multiple regressions we are comparing multiple estimates of the same regression. One estimator may include fewer variables than another; that is a restricted estimator. One may be estimated by least squares and another by 2SLS. Another could be nonparametric. The underlying model is the same; the difference is the estimator. Regardless, the literature has adopted the term “model selection” and we will adhere to this convention.

To gain some basic understanding it may be helpful to start with a stylized example. Suppose that we have a  $K \times 1$  estimator  $\hat{\theta}$  which has mean  $\theta$  and variance matrix  $V$ . An alternative feasible estimator is  $\tilde{\theta} = \mathbf{0}$ . The latter may seem like a silly estimator, but it captures the feature that model selection typically takes the form of exclusion restrictions set coefficients to 0. In this context we can compare the accuracy of the two estimators by their weighted mean-squared error (WMSE). For a given weight matrix  $W$  define

$$\text{wmse}(\hat{\theta}) = \text{tr}\left(\mathbb{E}\left((\hat{\theta} - \theta)(\hat{\theta} - \theta)'\right) W\right) = \mathbb{E}\left((\hat{\theta} - \theta)' W (\hat{\theta} - \theta)\right).$$

The calculations simplify by setting  $W = V^{-1}$ , which we do for our remaining calculations.

For our two estimators we calculate that

$$\text{wmse}(\hat{\theta}) = K \quad (24.1)$$

$$\text{wmse}(\tilde{\theta}) = \theta' V^{-1} \theta \stackrel{def}{=} \lambda. \quad (24.2)$$

(See Exercise 24.1) The WMSE of  $\hat{\theta}$  is smaller if  $K > \lambda$  and the WMSE of  $\tilde{\theta}$  is smaller if  $K < \lambda$ . We can visualize this distinction through Figure 24.1, which plots WMSE/K as a function of  $\lambda/K$ . For the smaller values of  $\lambda$  (left part of the figure) the simple estimator  $\tilde{\theta}$  has lower WMSE, while for the larger values of  $\lambda$  (the right part of the figure) the regular estimator  $\hat{\theta}$  has lower WMSE. One insight from this simple analysis is that we should prefer smaller (simpler) models when potentially omitted variables have small coefficients relative to estimation variance, and should prefer larger (more complicated) models when these variables have large coefficients relative to estimation variance.

Now consider a somewhat broader comparison. Suppose  $\hat{\theta}$  is  $\bar{K} \times 1$  with mean  $\theta$  and variance matrix  $V$ . For some  $\bar{K} \times (\bar{K} - K)$  full-rank matrix  $R$  consider

$$\tilde{\theta} = \hat{\theta} - VR(R'VR)^{-1}R'\hat{\theta}.$$

This is the standard restricted estimator under the assumption  $R'\theta = \mathbf{0}$ . You can calculate (see Exercise 24.1) that the weighted MSE of  $\tilde{\theta}$  is

$$\begin{aligned} \text{wmse}(\tilde{\theta}) &= \mathbb{E}\left((\tilde{\theta} - \theta)' V^{-1} (\tilde{\theta} - \theta)\right) \\ &= \theta' R (R'VR)^{-1} R' \theta + K. \end{aligned} \quad (24.3)$$

The first term is the squared bias, the second is the weighted variance. This simple expression illustrates the basic bias-variance trade-off. Increasing  $K$  increases the estimation variance but decreases the squared bias, the latter by decreasing the rank of  $R$ .

The bias can be estimated by replacing  $\hat{\theta}$  with  $\theta$ . This squared bias estimate is biased since

$$\mathbb{E}\left(\hat{\theta}' R (R'VR)^{-1} R' \hat{\theta}\right) = \theta' R (R'VR)^{-1} R' \theta + \bar{K} - K. \quad (24.4)$$

Putting these calculations together we see that an unbiased estimator for the weighted MSE is

$$\begin{aligned} M_K &= \hat{\theta}' R (R'VR)^{-1} R' \hat{\theta} + 2K - \bar{K} \\ &= (\hat{\theta} - \tilde{\theta})' V^{-1} (\hat{\theta} - \tilde{\theta}) + 2K - \bar{K}. \end{aligned}$$

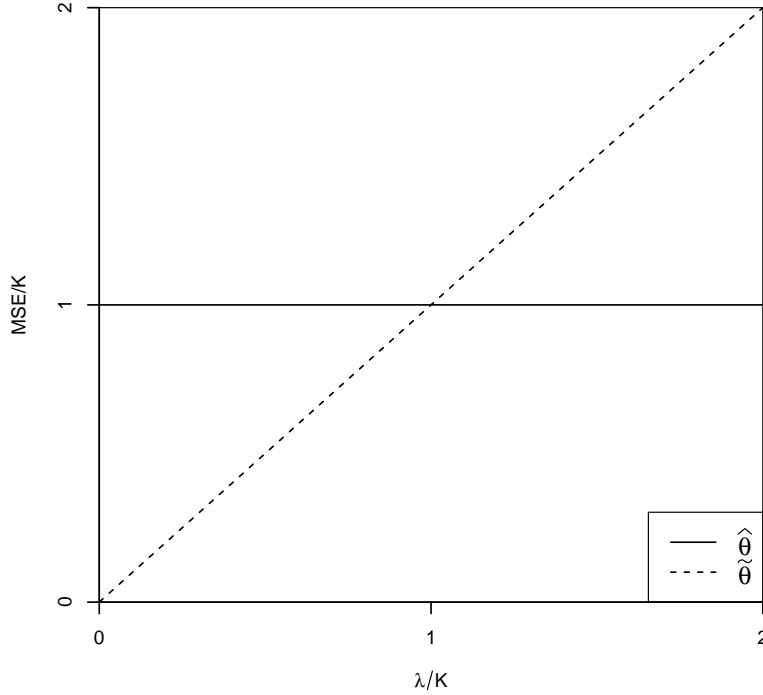


Figure 24.1: MSE Comparison

**Theorem 24.1** If  $\hat{\theta}$  has mean  $\theta$  and variance  $V$  and  $\tilde{\theta} = \hat{\theta} - VR(R'VR)^{-1}R'\hat{\theta}$  then

$$\mathbb{E}(M_K) = \text{wmse}(\hat{\theta}) - \text{wmse}(\tilde{\theta}).$$

(See Exercise 24.2 for the proof.)

The factor  $\bar{K}$  in  $M_K$  is constant across models so can be omitted for the purposes of model comparison.

In practice  $V$  is unknown. It can be replaced with a consistent estimator and we arrive at the **MSE Selection Criterion**

$$\begin{aligned} M_K &= \hat{\theta}'R(R'\hat{V}R)^{-1}R'\hat{\theta} + 2K \\ &= (\hat{\theta} - \tilde{\theta})' \hat{V}^{-1} (\hat{\theta} - \tilde{\theta}) + 2K. \end{aligned}$$

MSE selection picks the model for which the estimated WMSE  $M_K$  is the smallest. For implementation, a set of models are estimated,  $M_K$  calculated, and the model with the smallest  $M_K$  selected.

In practice the relative magnitudes of the coefficients and estimation variance are unknown. Model selection procedures address this uncertainty by using sample information to estimate a specific definition of model fit. Many selection procedures take the form of a penalized estimation criterion, where the penalty depends on the number of estimated parameters.

The MSE selection criterion described here is not a common model selection tool, but we have presented it as it is the simplest to derive and understand. Furthermore, it turns out to be quite similar to several popular methods, as we show later.

A large number of model selection criteria have been proposed. We list here those most frequently used in applied econometrics.

We first list selection criteria for the linear regression model  $y_i = \mathbf{x}'_i \boldsymbol{\beta} + e_i$  with  $\sigma^2 = \mathbb{E}(e_i^2)$  and a  $k \times 1$  coefficient vector  $\boldsymbol{\beta}$ . Let  $\hat{\boldsymbol{\beta}}$  be the least squares estimator,  $\hat{e}_i$  the least squares residual, and  $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n \hat{e}_i^2$  be the variance estimator. The number of estimated parameters ( $\boldsymbol{\beta}$  and  $\sigma^2$ ) is  $K = k + 1$ .

### Bayesian Information Criterion

$$\text{BIC} = n + n \log(2\pi\hat{\sigma}^2) + K \log(n). \quad (24.5)$$

### Akaike Information Criterion

$$\text{AIC} = n + n \log(2\pi\hat{\sigma}^2) + 2K. \quad (24.6)$$

### Cross-Validation

$$\text{CV} = \sum_{i=1}^n \tilde{e}_i^2 \quad (24.7)$$

where  $\tilde{e}_i$  are the least squares leave-one-out prediction errors.

As we show later, there is a close connection between the AIC, CV and the MSE selection criterion. The AIC and BIC are similar in form, but have quite different performance in practice.

We next list two commonly-used selection criteria for likelihood-based estimation. Let  $f(\mathbf{y}, \boldsymbol{\theta})$  be a parametric density with a  $K \times 1$  parameter  $\boldsymbol{\theta}$ . The likelihood  $L(\boldsymbol{\theta}) = f(\mathbf{y}, \boldsymbol{\theta}) = \prod_{i=1}^n f(y_i, \boldsymbol{\theta})$  is the density evaluated at the observations. The maximum likelihood estimator  $\hat{\boldsymbol{\theta}}$  maximizes  $L(\boldsymbol{\theta})$ .

### Bayesian Information Criterion

$$\text{BIC} = -2 \log L(\hat{\boldsymbol{\theta}}) + K \log(n). \quad (24.8)$$

### Akaike Information Criterion

$$\text{AIC} = -2 \log L(\hat{\boldsymbol{\theta}}) + 2K. \quad (24.9)$$

In the following sections we derive and discuss these and other model selection criteria.

## 24.3 Bayesian Information Criterion

The **Bayesian Information Criterion (BIC)**, also known as the **Schwarz Criterion**, was introduced by Schwarz (1978). It is appropriate for parametric models estimated by maximum likelihood, and is used to select the model with the highest approximate probability of being the true model.

Suppose that  $f(\mathbf{y}, \boldsymbol{\theta})$  is a parametric density. The likelihood  $L(\boldsymbol{\theta}) = f(\mathbf{y}, \boldsymbol{\theta}) = \prod_{i=1}^n f(y_i, \boldsymbol{\theta})$  is the density evaluated at the observations, and the maximum likelihood estimator  $\hat{\boldsymbol{\theta}}$  maximizes  $L(\boldsymbol{\theta})$ . Let  $\pi(\boldsymbol{\theta})$  be a prior density for  $\boldsymbol{\theta}$ . The joint density of  $\mathbf{y}$  and  $\boldsymbol{\theta}$  is  $f(\mathbf{y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta})$ . The marginal density of  $\mathbf{y}$  is

$$p(\mathbf{y}) = \int f(\mathbf{y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

The marginal density  $p(\mathbf{y})$  evaluated at the observations is known as the **marginal likelihood**.

Schwarz (1978) established the following approximation.

**Theorem 24.2** (Schwarz) If the model  $f(\mathbf{y}, \boldsymbol{\theta})$  satisfies standard regularity conditions and the prior  $\pi(\boldsymbol{\theta})$  is diffuse, then

$$-2 \log p(\mathbf{y}) = -2 \log L(\hat{\boldsymbol{\theta}}) + K \log(n) + O(1)$$

where the  $O(1)$  term is bounded as  $n \rightarrow \infty$ .

A heuristic proof for normal linear regression is given in Section 24.40. A “diffuse” prior is one which distributes weight uniformly over the parameter space.

Schwarz’s theorem shows that the marginal likelihood approximately equals the maximized likelihood, multiplied by an adjustment depending on the number of estimated parameters and the sample size. The approximation is commonly called the **Bayesian Information Criterion** or **BIC**:

$$\text{BIC} = -2\log L(\hat{\boldsymbol{\theta}}) + K \log(n).$$

The BIC is a **penalized log likelihood**. The term  $K \log(n)$  can be interpreted as an over-parameterization penalty. The multiplication of the log likelihood by  $-2$  is traditional, as it puts the criterion into the same units as a log-likelihood statistic.

In the context of normal linear regression, we have calculated in (5.9) that

$$\log L(\hat{\boldsymbol{\theta}}) = -\frac{n}{2} (\log(2\pi) + 1) - \frac{n}{2} \log(\hat{\sigma}^2)$$

where  $\hat{\sigma}^2$  is the residual variance estimate. Hence

$$\text{BIC} = n \log(2\pi\hat{\sigma}^2) + n + K \log(n).$$

with  $K = k + 1$ .

Since  $n \log(2\pi) + n$  does not vary across models this term is often omitted. It is better, however, to define the BIC correctly using all terms so that comparisons across different parametric families is done correctly. It is also useful to know that some authors define the BIC by dividing the above expression by  $n$  (e.g.  $\text{BIC} = \log(2\pi\hat{\sigma}^2) + K \log(n)/n$ ) which does not change the rankings between models. However, this is an unwise choice for it alters the scaling which makes it difficult to assess if two models are similar or not with respect to the BIC metric.

Now suppose that we have two models  $\mathcal{M}_1$  and  $\mathcal{M}_2$  which have marginal likelihoods  $p_1(\mathbf{y})$  and  $p_2(\mathbf{y})$ . Assume that both models have equal prior probability. Bayes Theorem states that the probability that a model is the true model given the data is proportional to the marginal likelihood of the model. Specifically

$$\begin{aligned}\Pr(\mathcal{M}_1 | \mathbf{y}) &= \frac{p_1(\mathbf{y})}{p_1(\mathbf{y}) + p_2(\mathbf{y})} \\ \Pr(\mathcal{M}_2 | \mathbf{y}) &= \frac{p_2(\mathbf{y})}{p_1(\mathbf{y}) + p_2(\mathbf{y})}.\end{aligned}$$

Bayes selection picks the model with highest probability. Thus if  $p_1(\mathbf{y}) > p_2(\mathbf{y})$  we select  $\mathcal{M}_1$ . If  $p_1(\mathbf{y}) < p_2(\mathbf{y})$  we select  $\mathcal{M}_2$ .

Finding the model with largest marginal likelihood is the same as finding the model with lowest value of  $-2\log p(\mathbf{y})$ . Theorem 24.2 shows that the latter approximately equals the BIC. BIC selection picks the model with the lowest<sup>1</sup> value of BIC. Thus BIC selection is approximate Bayes selection.

The above discussion concerned two models but applies to any number of models. BIC selection picks the model with the smallest BIC. For implementation you simply estimate each model, calculate its BIC, and compare.

The BIC may be obtained in Stata by using the command `estimates stats` after an estimated model.

## 24.4 Akaike Information Criterion for Regression

The **Akaike Information Criterion (AIC)** was introduced by Akaike (1973). It is used to select the model whose estimated density is closest to the true density. It is designed for parametric models estimated by maximum likelihood.

---

<sup>1</sup>When the BIC is negative this means taking the most negative value.

Let  $\hat{f}(\mathbf{y})$  be an estimate of the unknown density  $g(\mathbf{y})$  of the observation vector  $\mathbf{y} = (y_1, \dots, y_n)$ . For example, the normal linear regression estimate of  $g(\mathbf{y})$  is  $\hat{f}(\mathbf{y}) = \prod_{i=1}^n \phi_{\hat{\sigma}}(y_i - \mathbf{x}'_i \hat{\beta})$ .

To measure the distance between densities  $g$  and  $f$  Akaike used the Kullback-Leibler information criterion (KLIC)

$$\text{KLIC}(g, f) = \int g(\mathbf{y}) \log \left( \frac{g(\mathbf{y})}{f(\mathbf{y})} \right) d\mathbf{y}.$$

Notice that  $\text{KLIC}(g, f) = 0$  when  $f(\mathbf{y}) = g(\mathbf{y})$ . By Jensen's inequality,

$$\text{KLIC}(g, f) = - \int g(\mathbf{y}) \log \left( \frac{f(\mathbf{y})}{g(\mathbf{y})} \right) d\mathbf{y} \geq - \log \int f(\mathbf{y}) d\mathbf{y} = 0.$$

Thus  $\text{KLIC}(g, f)$  is a non-negative measure of the deviation of  $f$  from  $g$ , with small values indicating a smaller deviation.

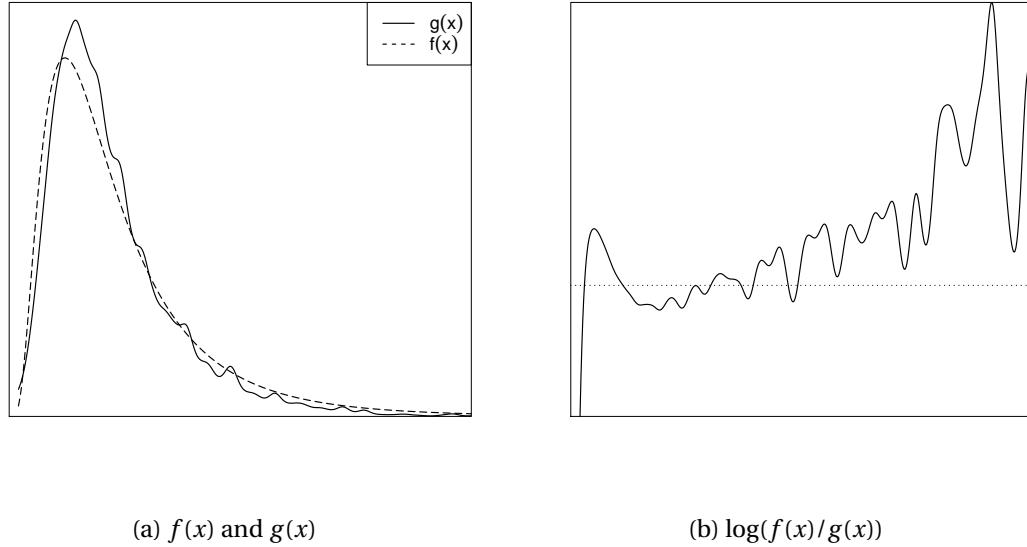


Figure 24.2: Kullback-Leibler Distance Measure

To illustrate, in Figure 24.2 we display two densities and their log ratio. In the left panel we display two densities  $f(x)$  and  $g(x)$ . For concreteness, the density  $g(x)$  is the nonparametric estimate of the log wage density displayed in Figure 2.1. The density  $f(x)$  is the MLE of a log-normal parametric model. You can see that the two densities are quite similar and have the same general shape. The parametric model, however, is somewhat lower at the peak, and may over-state the right tail of the density. In the right panel of the figure you see the log ratio  $\log(f(x)/g(x))$ . The dotted line is 0 for reference. If the two densities were the same then this plot would be the zero line. Negative values indicate regions where  $f(x) < g(x)$ . Positive values indicate regions where  $f(x) > g(x)$ . In this plot we see that the largest deviations are in the right tail, as the deviations are measured as percentage deviations. The KLIC is the weighted integral of this log ratio function. It is a weighted average, with weights given by the density  $g(x)$ . Since  $g(x)$  puts most probability mass in the left-middle of the plot, this is the region emphasized by the KLIC calculation. Thus while the right tail has the largest deviations, it does not receive a large weight in the KLIC because the density  $g(x)$  has little probability mass there.

The KLIC distance between the true and estimated densities is

$$\begin{aligned} \text{KLIC}(g, \hat{f}) &= \int g(\mathbf{y}) \log \left( \frac{g(\mathbf{y})}{\hat{f}(\mathbf{y})} \right) d\mathbf{y} \\ &= \int g(\mathbf{y}) \log g(\mathbf{y}) d\mathbf{y} - \int g(\mathbf{y}) \log \hat{f}(\mathbf{y}) d\mathbf{y}. \end{aligned}$$

This is random as it depends on the estimator  $\hat{f}$ . Akaike proposed examining the expected KLIC distance

$$\mathbb{E}(\text{KLIC}(g, \hat{f})) = \int g(\mathbf{y}) \log g(\mathbf{y}) d\mathbf{y} - \mathbb{E}\left(\int g(\mathbf{y}) \log \hat{f}(\mathbf{y}) d\mathbf{y}\right). \quad (24.10)$$

The first term in (24.10) does not depend on the model. So minimization of expected KLIC distance is minimization of the second term. Multiplied by 2 (similarly to the BIC) this is

$$T = -2\mathbb{E}\left(\int g(\mathbf{y}) \log \hat{f}(\mathbf{y}) d\mathbf{y}\right). \quad (24.11)$$

The expectation is over the random estimator  $\hat{f}$ .

An alternative interpretation is to notice that the integral in (24.11) is an expectation over  $\mathbf{y}$  with respect to the true data density  $g(\mathbf{y})$ . Thus we can write (24.11) as

$$T = -2\mathbb{E}(\log \hat{f}(\tilde{\mathbf{y}})) \quad (24.12)$$

where  $\tilde{\mathbf{y}}$  is an independent copy of  $\mathbf{y}$ . The key to understand this expression is that both the estimator  $\hat{f}$  and the evaluation points  $\tilde{\mathbf{y}}$  are random and independently distributed. This is the expected log-likelihood fit using the estimated model  $\hat{f}$  of an out-of-sample realization  $\tilde{\mathbf{y}}$ . Thus  $T$  can be interpreted as an expected predictive log likelihood. Models with low values of  $T$  have good fit based on the out-of-sample log-likelihood.

To gain further understanding we consider the simple case of the normal linear regression model with  $K$  regressors. The log density of the model for the observations is

$$\log f(\mathbf{y}, \theta) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2. \quad (24.13)$$

The expected value at the true parameter values is  $-\frac{n}{2} \log(2\pi\sigma^2) - \frac{n}{2}$ . This means that the idealized value of  $T$  is  $T_0 = n \log(2\pi\sigma^2) + n$ . This would be the value obtained if there were no estimation error.

We now add the assumption that the variance  $\sigma^2$  is known. This is not realistic but simplifies the calculations.

**Theorem 24.3** Suppose  $\hat{f}(\mathbf{y})$  is an estimated normal linear regression model with  $K$  regressors and a known variance  $\sigma^2$ . Suppose that the true density  $g(\mathbf{y})$  is a conditionally homoskedastic regression with variance  $\sigma^2$ . Then

$$T = n \log(2\pi\sigma^2) + n + K \quad (24.14)$$

$$\mathbb{E}(-2 \log L(\hat{\boldsymbol{\theta}})) = n \log(2\pi\sigma^2) + n - K. \quad (24.15)$$

The proof is given in Section 24.40.

These expressions are interesting. Expression (24.14) shows that  $T$  equals the idealized value  $T_0$  plus  $K$ . The latter is the cost of parameter estimation, measured in terms of expected KLIC distance. By estimating parameters (rather than using the true values) the expected KLIC distance increases linearly with  $K$ .

Expression (24.15) shows the converse story. It shows that the sample log-likelihood function is smaller than the idealized value  $T_0$  by  $K$ . This is the cost of in-sample over-fitting. The sample log-likelihood is an in-sample measure of fit, and therefore understates the population log-likelihood. The two expressions together show that the sample log-likelihood is smaller than the target value  $T$  by  $2K$ . This is the combined cost of over-fitting and parameter estimation.

Combining these expressions we can suggest an unbiased estimator for  $T$ . In the normal regression model we use

$$\text{AIC} = n \log(2\pi\hat{\sigma}^2) + n + 2K. \quad (24.16)$$

Since  $n \log(2\pi) + n$  does not vary across models it are often omitted. Thus for linear regression it is common to define the AIC as

$$\text{AIC} = n \log(\hat{\sigma}^2) + 2K.$$

Interestingly the AIC takes a very similar form to the BIC. Both the AIC and BIC are penalized log likelihoods, and both penalties are proportional to the number of estimated parameters  $K$ . The difference is that the AIC penalty is  $2K$  while the BIC penalty is  $K \log(n)$ . Since  $2 < \log(n)$  if  $n \geq 8$  the BIC uses a stronger parameterization penalty.

Selecting a model by the AIC is equivalent to calculating the AIC for each model and selecting the model with the lowest<sup>2</sup> value.

**Theorem 24.4** Under the assumptions of Theorem 24.3

$$\mathbb{E}(\text{AIC}) = T.$$

AIC is thus an unbiased estimator of  $T$ .

One of the interesting features of these results are that they are exact – there is no approximation error – and they do not require that the true error is normally distributed. The critical assumption is conditional homoskedasticity. If homoskedasticity fails then the AIC loses its validity. In more general contexts these exact results do not hold but instead hold as approximations (as discussed in the next section).

The AIC may be obtained in Stata by using the command `estimates stats` after an estimated model.

## 24.5 Akaike Information Criterion for Likelihood

For the general likelihood context Akaike proposed the criterion

$$\text{AIC} = -2 \log L(\hat{\theta}) + 2K.$$

Here,  $\hat{\theta}$  is the maximum likelihood estimator,  $\log L(\hat{\theta})$  is the maximized log-likelihood function, and  $K$  is the number of estimated parameters. This specializes to (24.16) for the case of a normal linear regression model.

As for regression, AIC selection is performed by estimating a set of models, calculating AIC for each, and selecting the model with the smallest AIC.

The advantages of the AIC are that it is simple to calculate, easy to implement, and straightforward to interpret. It is intuitive as it is a simple penalized likelihood.

The disadvantage is that its simplicity may be deceptive. The proof shows that the criterion is based on a quadratic approximation to the log likelihood function and an asymptotic chi-square approximation to the classical Wald statistic. When these conditions fail then the AIC may not be accurate. For example, if the model is an approximate (quasi) likelihood rather than a true likelihood, then the failure of the information matrix equality implies that the classical Wald statistic is not asymptotically normal. In this case the accuracy of AIC fails. Another problem is that many nonlinear models have parameter regions where parametric identification fails. In these models the quadratic approximation to the log

---

<sup>2</sup>When the AIC is negative this means taking the most negative value.

likelihood function fails to hold uniformly in the parameter space, so the accuracy of the AIC fails. These qualifications point to challenges in interpretation of the AIC in nonlinear models.

The following is an analog of Theorem 24.4.

**Theorem 24.5** Under standard regularity conditions for maximum likelihood estimation, plus the assumption that certain statistics (identified in the proof) are uniformly integrable,

$$\mathbb{E}(\text{AIC}) = T + O(n^{1/2}).$$

AIC is thus an approximately unbiased estimator of  $T$ .

A sketch of the proof is given in Section 24.40.

This result shows that the AIC is, in general, a reasonable estimator of the KLIC fit of an estimated parametric model. The theorem holds broadly for maximum likelihood estimation and thus the AIC can be used in a wide variety of contexts.

## 24.6 Mallows Criterion

The Mallows Criterion was proposed by Mallows (1973) and is often called the  $C_p$  criterion. It is appropriate for linear estimators of homoskedastic regression models.

Take the homoskedastic regression framework

$$\begin{aligned} y_i &= m_i + e_i \\ m_i &= m(\mathbf{x}_i) \\ \mathbb{E}(e_i | \mathbf{x}_i) &= 0 \\ \mathbb{E}(e_i^2 | \mathbf{x}_i) &= \sigma^2. \end{aligned}$$

Write the first equation in vector notation for the  $n$  observations as  $\mathbf{y} = \mathbf{m} + \mathbf{e}$ . Let  $\hat{\mathbf{m}} = \mathbf{A}\mathbf{y}$  be a linear estimator of  $\mathbf{m}$ , meaning that  $\mathbf{A}$  is some  $n \times n$  function of the regressor matrix  $\mathbf{X}$  only. The residuals are  $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{m}}$ . The class of linear estimators includes least squares, weighted least squares, kernel regression, local linear regression, and series regression. For example, the least squares estimator using a regressor matrix  $\mathbf{Z}$  is the case  $\mathbf{A} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ .

Mallows (1973) proposed the criterion

$$C_p = \hat{\mathbf{e}}'\hat{\mathbf{e}} + 2\tilde{\sigma}^2 \text{tr}(\mathbf{A}) \quad (24.17)$$

where  $\tilde{\sigma}^2$  is a preliminary estimator of  $\sigma^2$  (typically based on fitting a large model). In the case of least squares regression this simplifies to

$$C_p = n\hat{\sigma}^2 + 2K\tilde{\sigma}^2. \quad (24.18)$$

The Mallows criterion can be used similarly to the AIC. A set of regression models are estimated and the criterion  $C_p$  calculated for each. The model with the smallest value of  $C_p$  is the Mallows-selected model.

Mallows designed the criterion  $C_p$  as an unbiased estimator of the following measure of regression fit

$$R = \mathbb{E}\left(\sum_{i=1}^n (\hat{m}_i - m_i)^2\right).$$

This is the expected squared difference between the estimated and true regression evaluated at the observations.

An alternative motivation for  $R$  is in terms of prediction accuracy. Consider an independent set of observations  $\tilde{y}_i$ ,  $i = 1, \dots, n$ , which have the same regressors  $\mathbf{x}_i$  as those in sample. Consider prediction of  $\tilde{y}_i$  given  $\mathbf{x}_i$  and the fitted regression. The least squares predictor is  $\hat{m}_i$ . The sum of expected squared prediction errors is

$$\text{MSFE} = \sum_{i=1}^n \mathbb{E}(\tilde{y}_i - \hat{m}_i)^2.$$

The best possible (infeasible) value of this quantity is

$$\text{MSFE}_0 = \sum_{i=1}^n \mathbb{E}(\tilde{y}_i - m_i)^2.$$

The difference is the **prediction accuracy** of the estimator:

$$\begin{aligned} \text{MSFE} - \text{MSFE}_0 &= \sum_{i=1}^n \mathbb{E}(\tilde{y}_i - \hat{m}_i)^2 - \sum_{i=1}^n \mathbb{E}(\tilde{y}_i - m_i)^2 \\ &= \mathbb{E}\left(\sum_{i=1}^n (\hat{m}_i - m_i)^2\right) \\ &= R \end{aligned}$$

which equals Mallows' measure of regression fit. Thus  $R$  can be viewed as a measure of prediction accuracy.

We stated that the Mallows criterion is an unbiased estimator of  $R$ . More accurately, the adjusted criterion  $C_p^* = C_p - \mathbf{e}'\mathbf{e}$  is unbiased for  $R$ . When comparing models  $C_p$  and  $C_p^*$  are equivalent so this substitution has no consequence for model selection.

**Theorem 24.6** If  $\hat{\mathbf{m}} = \mathbf{A}\mathbf{y}$  is a linear estimator, the regression error is conditionally mean zero and homoskedastic, and  $\tilde{\sigma}^2$  is unbiased for  $\sigma^2$ , then

$$\mathbb{E}(C_p^*) = R$$

so the adjusted Mallows criterion  $C_p^*$  is an unbiased estimator of  $R$ .

The proof is given in Section 24.40.

## 24.7 Cross-Validation Criterion

In applied statistics and machine learning the default method for model selection and tuning parameter selection is cross-validation. We have introduced some of the concepts throughout the textbook, and review and unify the concepts at this point.

In Section 3.20 we defined the leave-one-out estimator as that obtained by applying an estimation formula to the sample omitting the  $i^{th}$  observation. Equation (3.43), for example, gives the definition for the least squares estimator. Theorem 3.7 gives the convenient computation formula

$$\hat{\boldsymbol{\beta}}_{(-i)} = \hat{\boldsymbol{\beta}} - \frac{1}{(1 - h_{ii})} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i \hat{e}_i$$

where  $\hat{e}_i$  are the least squares residuals and  $h_{ii}$  are the leverage values. We also defined the leave-one-out residual or prediction error as that obtained using the leave-one-out regression estimator, thus

$$\tilde{e}_i = y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{(-i)} = (1 - h_{ii})^{-1} \hat{e}_i$$

where the second equality is from Theorem 3.7. We defined the out-of-sample mean squared error as

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \tilde{e}_i^2 = \frac{1}{n} \sum_{i=1}^n (1 - h_{ii})^{-2} \hat{e}_i^2.$$

In Section 4.12 we defined the mean squared forecast error as the expectation of the squared out-of-sample prediction error

$$\text{MSFE}_n = \mathbb{E}(\tilde{e}_{n+1}^2).$$

In Theorem 4.6 we showed that  $\tilde{\sigma}^2$  is approximately an unbiased estimator of the MSFE.

In our study of nonparametric regression (Section 20.12) we defined the cross-validation criterion for kernel regression as the weighted average of the squared prediction errors

$$\text{CV} = \frac{1}{n} \sum_{i=1}^n \tilde{e}_i^2 w(x_i).$$

Theorem 20.6 showed that CV is approximately unbiased for the integrated mean squared error (IMSE), which is a standard measure of accuracy for nonparametric regression. Since CV and  $\tilde{\sigma}^2$  are identical (in the absence of weights) these results show that CV is an unbiased estimator for both the MSFE and IMSE, showing a close connection between these measures of accuracy.

In Section 21.17 and equation (21.28) we defined the CV criterion for series regression as in (24.7). Selecting the variables for series regression is identical to the problem of model selection. The results as described above show that the CV criterion is an estimator for the MSFE and IMSE of the regression model and is therefore a good candidate for assessing model accuracy. The validity of the CV criterion is much broader than the AIC, as the theorems for CV do not require conditional homoskedasticity. This is not an artifact of the proof method; cross-validation is inherently more robust than AIC or BIC.

Implementation of CV model selection is the same as for the other criteria. A set of regression models are estimated. For each the CV criterion is calculated. The model with the smallest value of CV is the CV-selected model.

The CV method is also much broader in concept and potential application. It applies to any estimation method, so long as a “leave one out” error can be calculated. It can also be applied to other loss functions beyond squared error loss. For example, a cross-validation estimate of absolute loss is

$$\text{CV} = \frac{1}{n} \sum_{i=1}^n |\tilde{e}_i|.$$

Computationally and conceptually it is straightforward to select models by minimizing such criterion. However, the properties of applying CV to general criterion is not known.

Stata does not have a standard command to calculate the CV criterion for regression models.

## 24.8 K-Fold Cross-Validation

One challenge with implementation of cross validation is that it can be computationally costly when sample sizes are very large or the estimation method is other than least squares. For least squares estimation there is a simple expression for the CV criterion but this is not the case for other estimators. In such cases to evaluate the CV criterion  $n$  separate estimators are calculated and this may be computationally costly.

A relatively low cost simplification is to split the sample into  $K$  groups (or “folds”) and treat each group as a hold-out sample. This effectively reduces the number of estimations from  $n$  to  $K$ . (This  $K$  is not the number of estimated coefficients. I apologize for the possible confusion in notation but this is the standard label.) The most common choices<sup>3</sup> are  $K = 5$ ,  $K = 10$ , and  $K = 20$ , leading to what is known as “5-fold”, “10-fold”, and “20-fold cross validation”.

The method works by the following steps. This description is for estimation of a regression model  $y_i = g(\mathbf{x}_i, \boldsymbol{\theta}) + e_i$  with estimator  $\hat{\boldsymbol{\theta}}$ .

---

<sup>3</sup>To obtain good accuracy and reliability the number of “folds” should be taken to be as large as computationally reasonable.

1. Randomly sort the observations.
2. Split the observations into  $K$  groups of (roughly) equal size  $n/K$ .
3. For  $k = 1, \dots, K$ 
  - (a) Exclude the  $k^{th}$  group from the dataset. This produces a sample with  $n - n/K$  observations.
  - (b) Calculate the estimator  $\hat{\theta}_{(-k)}$  on this sample.
  - (c) Calculate the prediction errors  $\tilde{e}_i = y_i - g(\mathbf{x}_i, \hat{\theta}_{(-k)})$  for observations within the  $k^{th}$  group.
4. This produces prediction errors for all observations.
5. Calculate  $CV = \sum_{i=1}^n \tilde{e}_i^2$ .

If  $K = n$  the method is identical to leave-one-out cross validation.

A disadvantage of  $K$ -fold cross-validation is that the results can be sensitive to the initial random sorting of the observations. Consequently some practitioners calculate the criterion  $M$  times and then average the results. A better solution, however, at the same computation cost, is to use  $MK$  folds. The randomness of the method diminishes as the number of folds increase.

$K$ -fold CV can be interpreted of as an approximation to leave-one-out ( $n$ -fold) CV.

## 24.9 Many Selection Criteria are Similar

For the linear regression model many selection criteria have been introduced. However, many of these alternative criteria are quite similar to one another. In this section we review some of these connections.

Considering the WMSE criterion, let  $\tilde{\sigma}^2$  denote the variance estimator of the unconstrained model. Then

$$\tilde{\sigma}^2(M_K + n) = n\hat{\sigma}^2 + 2K\tilde{\sigma}^2 = C_p$$

the Mallows criterion for regression. Minimization of the left and rights sides are identical. Thus WMSE and Mallows selection are identical.

Shibata (1980) proposed the criteria

$$\text{Shibata} = \hat{\sigma}^2 \left( 1 + \frac{2K}{n} \right)$$

as an estimator of the MSFE. Recalling the Mallows criterion for regression (24.18) we see that  $\text{Shibata} = C_p/n$  if we replace  $\tilde{\sigma}^2$  with  $\hat{\sigma}^2$ . Thus the two are quite similar in practice.

Taking logarithms and using the approximation  $\log(1 + x) \approx x$  for small  $x$

$$n \log \text{Shibata} = n \log(\hat{\sigma}^2) + n \log \left( 1 + \frac{2K}{n} \right) \approx n \log(\hat{\sigma}^2) + 2K = \text{AIC}.$$

Thus minimization of Shibata's criterion and AIC are similar.

Akaike (1969) proposed the Final Prediction Error Criteria

$$\text{FPE} = \hat{\sigma}^2 \left( \frac{1 + K/n}{1 - K/n} \right).$$

Using the expansions  $(1 - x)^{-1} \approx 1 + x$  and  $(1 + x)^2 \approx 1 + 2x$  we see that FPE  $\approx$  Shibata.

Craven and Wahba (1979) proposed Generalized Cross Validation

$$\text{GCV} = \frac{n\hat{\sigma}^2}{(n - K)^2}.$$

By the expansion  $(1 - x)^{-2} \approx 1 + 2x$  we find that

$$n\text{GCV} = \frac{\hat{\sigma}^2}{(1 - K/n)^2} \approx \hat{\sigma}^2 \left(1 + \frac{2K}{n}\right) = \text{Shibata.}$$

The above calculations show that the WMSE, AIC, Shibata, FPE, GCV, and Mallows criterion are all close approximations to one another when  $K/n$  is small. Differences arise in finite samples for large  $K$ . However, the above analysis shows that there is no fundamental difference between these criteria. They are all estimating the same target. This is in contrast to BIC which uses a different parameterization penalty and is asymptotically distinct.

Interestingly there also is a connection between CV and the above criteria. Again using the expansion  $(1 - x)^{-2} \approx 1 + 2x$  we find that

$$\begin{aligned} \text{CV} &= \sum_{i=1}^n (1 - h_{ii})^{-2} \hat{e}_i^2 \\ &\approx \sum_{i=1}^n \hat{e}_i^2 + \sum_{i=1}^n 2h_{ii}\hat{e}_i^2 \\ &= n\hat{\sigma}^2 + 2 \sum_{i=1}^n \mathbf{x}'_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i \hat{e}_i^2 \\ &= n\hat{\sigma}^2 + 2 \text{tr} \left( (\mathbf{X}'\mathbf{X})^{-1} \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \hat{e}_i^2 \right) \right) \\ &\approx n\hat{\sigma}^2 + 2 \text{tr} \left( (\mathbb{E}(\mathbf{x}_i \mathbf{x}'_i))^{-1} (\mathbb{E}(\mathbf{x}_i \mathbf{x}'_i \hat{e}_i^2)) \right) \\ &= n\hat{\sigma}^2 + 2K\sigma^2 \\ &\approx \text{Shibata}. \end{aligned}$$

The third-to-last line holds asymptotically by the WLLN. The following equality holds under conditional homoskedasticity. The final approximation replaces  $\sigma^2$  by the estimator  $\hat{\sigma}^2$ . This calculation shows that under the assumption of conditional homoskedasticity the CV criterion is similar to the other criteria. It differs under heteroskedasticity, however, which is one of its primary advantages.

## 24.10 Relation with Likelihood Ratio Testing

Since the AIC and BIC are penalized log-likelihoods, AIC and BIC selection are related to likelihood ratio testing. Suppose we have two nested models  $\mathcal{M}_1$  and  $\mathcal{M}_2$  with log-likelihoods  $\log L_1(\hat{\boldsymbol{\theta}}_1)$  and  $\log L_2(\hat{\boldsymbol{\theta}}_2)$  and  $K_1 < K_2$  estimated parameters. AIC selects  $\mathcal{M}_1$  if  $\text{AIC}(K_1) < \text{AIC}(K_2)$  which occurs when

$$-2\log L_1(\hat{\boldsymbol{\theta}}_1) + 2K_1 < -2\log L_2(\hat{\boldsymbol{\theta}}_2) + 2K_2$$

or

$$LR = 2(\log L_2(\hat{\boldsymbol{\theta}}_2) - \log L_1(\hat{\boldsymbol{\theta}}_1)) < 2r$$

where  $r = K_2 - K_1$ . Thus AIC selection is similar to selection by likelihood ratio testing with a different critical value. Rather than using a critical value from the chi-square distribution, the “critical value” is  $2r$ . This is not to say that AIC selection is testing (it is not). But rather that there is a similar structure in the decision.

There are two useful practical implications. One is that when test statistics are reported in their  $F$  form (which divide by the difference in coefficients  $r$ ) then the AIC “critical value” is 2. The AIC selects the restricted (smaller) model if  $F < 2$ . It selects the unrestricted (larger) model if  $F > 2$ .

Another useful implication is in the case of considering a single coefficient (when  $r = 1$ ). AIC selects the coefficient (the larger model) if  $LR > 2$ . In contrast a 5% significance test “selects” the larger model (rejects the smaller) if  $LR > 3.84$ . Thus AIC is more generous in terms of selecting larger models. An

equivalent way of seeing this is that AIC selects the coefficient if the t-ratio exceeds 1.41, while the 5% significance test selects if the t-ratio exceeds 1.96.

Similar comments apply to BIC selection, though the effective critical values are different. For comparing models with coefficients  $K_1 < K_2$ , the BIC selects  $\mathcal{M}_1$  if  $LR < \log(n)r$ . The “critical value” for an  $F$  statistic is  $\log(n)$ . Hence BIC selection becomes more strict as sample sizes increase.

## 24.11 Consistent Selection

An important property of a model selection procedure is whether it selects a true model in large samples. We call such a procedure **consistent**.

To discuss this further we need to thoughtfully define what is a “true” model. The answer depends on the type of model.

When a model is a parametric density or distribution  $f(y, \boldsymbol{\theta})$  with  $\boldsymbol{\theta} \in \Theta$  (as in likelihood estimation) then the model is true if there is some  $\boldsymbol{\theta}_0 \in \Theta$  such that  $f(y, \boldsymbol{\theta}_0)$  equals the true density or distribution. Notice that it is important in this context both that the function class  $f(y, \boldsymbol{\theta})$  and parameter space  $\Theta$  are appropriately defined.

In a semiparametric conditional moment condition model which states  $\mathbb{E}(g(y_i, \mathbf{x}_i, \boldsymbol{\theta}) | \mathbf{x}_i) = 0$  with  $\boldsymbol{\theta} \in \Theta$  then the model is true if there is some  $\boldsymbol{\theta}_0 \in \Theta$  such that  $\mathbb{E}(g(y_i, \mathbf{x}_i, \boldsymbol{\theta}_0) | \mathbf{x}_i) = 0$ . This includes the regression model  $y_i = m(\mathbf{x}_i, \boldsymbol{\theta}) + e_i$  with  $\mathbb{E}(e_i | \mathbf{x}_i) = 0$  where the model is true if there is some  $\boldsymbol{\theta}_0 \in \Theta$  such that  $m(\mathbf{x}_i, \boldsymbol{\theta}_0) = \mathbb{E}(y_i | \mathbf{x}_i)$ . It also includes the homoskedastic regression model which adds the requirement that  $\mathbb{E}(e_i^2 | \mathbf{x}_i) = \sigma^2$  is a constant. A semiparametric model does not require, however, that the true data distribution is specified.

In a semiparametric unconditional moment condition model which states  $\mathbb{E}g(y_i, \mathbf{x}_i, \boldsymbol{\theta}) = 0$  with  $\boldsymbol{\theta} \in \Theta$  then the model is true if there is some  $\boldsymbol{\theta}_0 \in \Theta$  such that  $\mathbb{E}g(y_i, \mathbf{x}_i, \boldsymbol{\theta}_0) = 0$ . A subtle issue here is that when the model is just identified and  $\Theta$  is unrestricted then this condition typically holds and so the model is typically true. This includes least squares regression interpreted as a projection and just-identified instrumental variables regression.

In a nonparametric model such as  $y \sim f \in \mathcal{F}$ , where  $\mathcal{F}$  is some function class (such as second-order differentiable densities) then the model is true if the true density is a member of the function class  $\mathcal{F}$ .

A complication arises that there may be multiple true models. This cannot occur when models are strictly non-nested (meaning that there is no common element in both model classes) but strictly non-nested models are rare. Most models have non-trivial intersections. For example, the linear regression models  $y_i = \alpha + \mathbf{x}'_{1i}\boldsymbol{\beta}_1 + e_i$  and  $y_i = \alpha + \mathbf{x}'_{2i}\boldsymbol{\beta}_2 + e_i$  with  $\mathbf{x}_{1i}$  and  $\mathbf{x}_{2i}$  containing no common elements may appear non-nested, but they intersect when  $\boldsymbol{\beta}_1 = 0$  and  $\boldsymbol{\beta}_2 = 0$ . As another example consider the linear model  $y_i = \alpha + \mathbf{x}'_i\boldsymbol{\beta} + e_i$  and log-linear model  $\log(y_i) = \alpha + \mathbf{x}'_i\boldsymbol{\beta} + e_i$ . If we add the assumption that  $e_i \sim N(0, \sigma^2)$  then the models are non-intersecting. But if we relax normality and instead use the conditional mean assumption  $\mathbb{E}(e_i | \mathbf{x}_i) = 0$  then the models are intersecting when  $\boldsymbol{\beta}_1 = 0$  and  $\boldsymbol{\beta}_2 = 0$ .

The most common type of intersecting models are nested. In regression this occurs when the two models are  $y_i = \mathbf{x}'_{1i}\boldsymbol{\beta}_1 + e_i$  and  $y_i = \mathbf{x}'_{1i}\boldsymbol{\beta}_1 + \mathbf{x}'_{2i}\boldsymbol{\beta}_2 + e_i$ . If  $\boldsymbol{\beta}_2 \neq 0$  then only the second model is true. But if  $\boldsymbol{\beta}_2 = 0$  then both are true models.

In general, given a set of models  $\overline{\mathcal{M}} = \{\mathcal{M}_1, \dots, \mathcal{M}_M\}$ , a subset  $\overline{\mathcal{M}}^*$  are true models (as described above) while the remainder are not true models.

A model selection rule  $\widehat{\mathcal{M}}$  selects one model from the set  $\overline{\mathcal{M}}$ . We say a method is consistent if it asymptotically selects a true model.

**Definition 24.1** A model selection rule is **model selection consistent** if

$$\Pr(\widehat{\mathcal{M}} \in \overline{\mathcal{M}}^*) \rightarrow 1$$

as  $n \rightarrow \infty$ .

This states that the model selection rule selects a true model with probability tending to 1 as the sample size diverges.

A broad class of model selection methods satisfy this definition of consistency. To see this consider the class of information criteria

$$\text{IC} = -2 \log L(\hat{\boldsymbol{\theta}}) + c(n, K).$$

This includes AIC ( $c = 2K$ ), BIC ( $c = K \log(n)$ ), and testing-based selection ( $c$  equals a fixed quantile of the  $\chi_K^2$  distribution).

**Theorem 24.7** Under standard regularity conditions for maximum likelihood estimation, selection based on IC is model selection consistent if  $c(n, K) = o(n)$  as  $n \rightarrow \infty$ .

The proof is given in Section 24.40.

This result covers AIC, BIC and testing-based selection. Thus all are model selection consistent.

A major limitation with this result is that the definition of model selection consistency is weak. A model may be true but over parameterized. To understand the distinction consider the models  $y_i = \mathbf{x}'_{1i}\boldsymbol{\beta}_1 + e_i$  and  $y_i = \mathbf{x}'_{1i}\boldsymbol{\beta}_1 + \mathbf{x}'_{2i}\boldsymbol{\beta}_2 + e_i$ . If  $\boldsymbol{\beta}_2 = 0$  then both  $\mathcal{M}_1$  and  $\mathcal{M}_2$  are true, but  $\mathcal{M}_1$  would be the preferred model as it is more parsimonious. When two nested models are both true models, it is conventional to think of the more parsimonious model as the correct model. In this context we do not describe the larger model as an incorrect model, but rather as over-parameterized. If a selection rule asymptotically selects an over-parameterized model we say that it “over-selects”.

**Definition 24.2** A model selection rule **asymptotically over-selects** if there are models  $\mathcal{M}_1 \subset \mathcal{M}_2$  such that

$$\liminf_{n \rightarrow \infty} \Pr\left(\widehat{\mathcal{M}} = \mathcal{M}_2 \mid \mathcal{M}_1\right) > 0.$$

The definition states that over-selection occurs when two models are nested and the smaller (short) model is true (so both models are true models but the smaller model is more parsimonious), if the larger model is asymptotically selected with positive probability.

**Theorem 24.8** Under standard regularity conditions for maximum likelihood estimation, selection based on IC asymptotically over-selects if  $c(n, K) = O(1)$  as  $n \rightarrow \infty$ .

The proof is given in Section 24.40.

This result includes both AIC and testing-based selection. Thus these procedures over-select. For example, if the models are  $y_i = \mathbf{x}'_{1i}\boldsymbol{\beta}_1 + e_i$  and  $y_i = \mathbf{x}'_{1i}\boldsymbol{\beta}_1 + \mathbf{x}'_{2i}\boldsymbol{\beta}_2 + e_i$  and  $\boldsymbol{\beta}_2 = 0$  holds, then these procedures select the over-parameterized regression with positive probability.

Following this line of reasoning, it is useful to draw a distinction between true and parsimonious models. We define the set of **parsimonious models**  $\overline{\mathcal{M}}^0 \subset \overline{\mathcal{M}}^*$  as the set of true models with the fewest number of parameters. When the models in  $\overline{\mathcal{M}}^*$  are nested then  $\overline{\mathcal{M}}^0$  will be a singleton. In the regression example with  $\boldsymbol{\beta}_2 = 0$  then  $\mathcal{M}_1$  is the unique parsimonious model among  $\{\mathcal{M}_1, \mathcal{M}_2\}$ . We introduce a stronger consistency definition for procedures which asymptotically select parsimonious models.

**Definition 24.3** A model selection rule is **consistent for parsimonious models** if

$$\Pr\left(\widehat{\mathcal{M}} \in \overline{\mathcal{M}}^0\right) \rightarrow 1$$

as  $n \rightarrow \infty$ .

Of the methods we have reviewed, only BIC selection is consistent for parsimonious models, as we now show.

**Theorem 24.9** Under standard regularity conditions for maximum likelihood estimation, selection based on IC is consistent for parsimonious models if for all  $K_2 > K_1$

$$c(n, K_2) - c(n, K_1) \rightarrow \infty \quad (24.19)$$

as  $n \rightarrow \infty$ , yet  $c(n, K) = o(n)$  as  $n \rightarrow \infty$ .

The proof is given in Section 24.40.

The condition includes BIC, as  $c(n, K_2) - c(n, K_1) = (K_2 - K_1) \log(n) \rightarrow \infty$  if  $K_2 > K_1$ .

Some economists have interpreted Theorem 24.9 as indicating that BIC selection is preferred over the other methods. This is a narrow reading of the result. In the next section we show that the other selection procedures are asymptotically optimal in terms of model fit and in terms of out-of-sample forecasting. Thus consistent model selection is only one of several desirable statistical properties.

## 24.12 Asymptotic Selection Optimality

Regressor selection by the WMSE/AIC/Shibata/Mallows/CV class turns out to be asymptotically optimal with respect to out-of-sample prediction under quite broad conditions. This may appear to conflict with the results of the previous section but it does not as there is a critical difference between the goals of consistent model selection and accurate prediction.

Our analysis will be in the homoskedastic regression model, conditioning on the regressor matrix  $X$ . All stated expectations are conditional on  $X$ , but to keep the notation uncluttered we will often write the expectations without explicit conditioning.

We write the regression model as

$$\begin{aligned} y_i &= m_i + e_i \\ m_i &= \sum_{j=1}^{\infty} x_{ji} \beta_j \\ \mathbb{E}(e_i | \mathbf{x}_i) &= 0 \\ \mathbb{E}(e_i^2 | \mathbf{x}_i) &= \sigma^2 \end{aligned}$$

where  $\mathbf{x}_i = (x_{1i}, x_{2i}, \dots)$ . We can also write the regression equation in matrix notation as  $\mathbf{y} = \mathbf{m} + \mathbf{e}$ .

The  $K^{th}$  regression model uses the first  $K$  regressors  $\mathbf{x}_{Ki} = (x_{1i}, x_{2i}, \dots, x_{Ki})$ . The least squares estimates in matrix notation are

$$\mathbf{y} = \mathbf{X}_K \widehat{\boldsymbol{\beta}}_K + \widehat{\mathbf{e}}_K.$$

As in Section 24.6 define the fitted values  $\widehat{\mathbf{m}} = \mathbf{X}_K \widehat{\boldsymbol{\beta}}_K$  and regression fit (out-of-sample sum of expected squared prediction errors) as

$$R_n(K) = \mathbb{E}\left((\widehat{\mathbf{m}} - \mathbf{m})' (\widehat{\mathbf{m}} - \mathbf{m})\right)$$

though now we index  $R$  by sample size  $n$  and model  $K$  for precision.

In any sample there is an optimal model  $K$  which minimizes  $R_n(K)$ :

$$K_n^{\text{opt}} = \underset{K}{\operatorname{argmin}} R_n(K).$$

Model  $K_n^{\text{opt}}$  obtains the minimized value of  $R_n(K)$

$$R_n^{\text{opt}} = R_n(K_n^{\text{opt}}) = \min_K R_n(K).$$

Now consider model selection using the Mallow's criterion for regression models

$$C_p(K) = \hat{\mathbf{e}}_K' \hat{\mathbf{e}}_K + 2\sigma^2 K$$

where we explicitly index by  $K$ , and for simplicity we assume the penalty depends on the true error variance  $\sigma^2$ . (The results are unchanged if it is replaced by a consistent estimator.) Let the selected model be

$$\hat{K}_n = \underset{K}{\operatorname{argmin}} C_p(K).$$

The prediction accuracy using the Mallows-selected model is  $R_n(\hat{K}_n)$ . We say that a selection procedure is **asymptotically optimal** if the prediction accuracy is asymptotically equivalent with the infeasible optimum. This can be written as

$$\frac{R_n(\hat{K}_n)}{R_n^{\text{opt}}} \xrightarrow{p} 1. \quad (24.20)$$

We consider convergence in (24.20) in terms of the risk ratio since  $R_n^{\text{opt}}$  diverges as the sample size increases.

Li (1987) established the asymptotic optimality (24.20). His result depends on the following conditions.

#### Assumption 24.1

1. The observations  $(y_i, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ , are independent and identically distributed.
2.  $\mathbb{E}(e_i | \mathbf{x}_i) = 0$ .
3.  $\mathbb{E}(e_i^2 | \mathbf{x}_i) = \sigma^2$ .
4.  $\mathbb{E}(|e_i|^{4r} | \mathbf{x}_i) \leq B < \infty$  for some  $r > 1$
5.  $R_n^{\text{opt}} \rightarrow \infty$  as  $n \rightarrow \infty$ .
6. The estimated models are nested.

Assumptions 24.1.2 and 24.1.3 state that the true model is a conditionally homoskedastic regression. This is important for the result. Assumption 24.1.4 is a technical condition, that a conditional moment of the error is uniformly bounded. Assumption 24.1.5 is subtle. It effectively states that there is no correctly specified finite-dimensional model. To see this, suppose that there is a  $K$  such that the model is correctly specified, meaning that  $m_i = \sum_{j=1}^K x_{ji} \beta_j$ . In this case we can show that  $R_n(K) = 0$ , violating Assumption 24.1.5. This is an important assumption for the optimality result. Assumption 24.1.6 is a technical condition that restricts the number of estimated models. Non-nested models can be allowed but then an alternative restriction on the number of estimated models is needed.

**Theorem 24.10** Under Assumption 24.1, (24.20) holds. Thus Mallows selection is asymptotically equivalent to using the infeasible optimal model.

The proof is given in Section 24.40.

Theorem 24.10 states that Mallows selection in a conditional homoskedastic regression is asymptotically optimal. The key assumptions are homoskedasticity and that all finite-dimensional models are misspecified (incomplete), meaning that there are always omitted variables. The latter means that regardless of the sample size there is always a trade-off between omitted variables bias and estimation variance. The theorem as stated is specific for Mallows selection, but extends to AIC, Shibata, GCV, FPE, and CV with some additional technical considerations. The primary message is that the selection methods discussed in the previous section asymptotically select a sequence of models which are best-fitting in the sense of minimizing the prediction error.

Using a similar argument Andrews (1991c) showed that selection by cross-validation satisfies the same asymptotic optimality condition without requiring conditional homoskedasticity. The treatment is a bit more technical so we do not review it here. This indicates an important advantage for cross-validation selection over the other methods.

## 24.13 Focused Information Criterion

Claeskens and Hjort (2003) introduced the **Focused Information Criterion (FIC)** as an estimator of the MSE of a scalar parameter of interest. The criterion is appropriate in correctly-specified likelihood models when one of the estimated models nests all other models. Let  $f(y, \boldsymbol{\theta})$  be a parametric density with a  $K \times 1$  parameter  $\boldsymbol{\theta}$ . The likelihood  $L(\boldsymbol{\theta}) = f(\mathbf{y}, \boldsymbol{\theta}) = \prod_{i=1}^n f(y_i, \boldsymbol{\theta})$  is the density evaluated at the observations, and the unrestricted maximum likelihood estimator  $\hat{\boldsymbol{\theta}}$  maximizes  $L(\boldsymbol{\theta})$ .

The class of models (sub-models) allowed are those defined by a set of differentiable restrictions  $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$ . Let  $\tilde{\boldsymbol{\theta}}$  be the restricted MLE which maximizes the likelihood subject to  $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$ .

A key feature of the FIC is that it focuses on a real-valued parameter of interest  $\mu = g(\boldsymbol{\theta})$  where  $g$  is some differentiable function. Claeskens and Hjort call  $\mu$  the **target parameter**. The choice of  $\mu$  is made by the researcher, and is a critical choice. In most applications  $\mu$  is the key coefficient of interest in the application (for example, the returns to schooling in a wage regression). The unrestricted MLE for  $\mu$  is  $\hat{\mu} = g(\hat{\boldsymbol{\theta}})$ , the restricted MLE is  $\tilde{\mu} = g(\tilde{\boldsymbol{\theta}})$ .

Estimation accuracy is measured by the MSE of the estimator of the target parameter, which is the squared bias plus the variance:

$$\begin{aligned}\text{mse}(\tilde{\mu}) &= \mathbb{E}(\tilde{\mu} - \mu)^2 \\ &= (\mathbb{E}(\tilde{\mu}) - \mu)^2 + \text{var}(\tilde{\mu}).\end{aligned}$$

It turns out to be convenient to normalize the MSE by that of the unrestricted estimator. We define this as the Focus

$$F = \text{mse}(\tilde{\mu}) - \text{mse}(\hat{\mu}).$$

The Claeskens-Hjort FIC is an estimator of  $F$ . Specifically,

$$\text{FIC} = (\tilde{\mu} - \hat{\mu})^2 - 2\hat{\mathbf{G}}'\hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}\hat{\mathbf{R}}\left(\hat{\mathbf{R}}'\hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}\hat{\mathbf{R}}\right)^{-1}\hat{\mathbf{R}}'\hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}\hat{\mathbf{G}}$$

where  $\hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}$ ,  $\hat{\mathbf{G}}$  and  $\hat{\mathbf{R}}$  are estimators of  $\text{var}(\hat{\boldsymbol{\theta}})$ ,  $\mathbf{G} = \frac{\partial}{\partial \boldsymbol{\theta}}g(\boldsymbol{\theta})$  and  $\mathbf{R} = \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{r}(\boldsymbol{\theta})$ .

In a least squares regression  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$  with a linear restriction  $\mathbf{R}'\boldsymbol{\beta} = 0$  and linear parameter of interest  $\mu = \mathbf{G}'\boldsymbol{\beta}$  the FIC equals

$$\begin{aligned}\text{FIC} &= \left( \mathbf{G}'\mathbf{R}\left(\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\right)^{-1}\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\hat{\boldsymbol{\beta}} \right)^2 \\ &\quad - 2\hat{\sigma}^2\mathbf{G}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\left(\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\right)^{-1}\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}.\end{aligned}$$

The FIC is used similarly to AIC. The FIC is calculated for each sub-model of interest, and the model with the lowest value of FIC is selected. All estimated models need to be sub-models of the unrestricted model.

The advantage of the FIC is that it is specifically targeted to minimize the MSE of the target parameter. The FIC is therefore appropriate when the goal is to estimate a specific target parameter. A disadvantage is that it does not necessarily produce a model with good estimates of the other parameters. For example, in a linear regression  $y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + e_i$ , if  $x_{1i}$  and  $x_{2i}$  are uncorrelated and the focus parameter is  $\beta_1$ , then the FIC will tend to select the sub-model without  $x_{2i}$ , and thus the selected model will produce a highly biased estimate of  $\beta_2$ . Consequently when using the FIC it is dubious if attention should be paid to estimates other than those of  $\mu$ .

Computationally it may be convenient to implement the FIC using an alternative formulation. Define the adjusted focus

$$F^* = n(F + 2\text{mse}(\hat{\mu})) = n(\text{mse}(\tilde{\mu}) + \text{mse}(\hat{\mu})).$$

This adds the same quantity to all models and therefore does not alter the minimizing model. Multiplication by  $n$  puts the FIC in units which are easier for reporting. The estimate of the adjusted focus is an adjusted FIC and can be written as

$$\text{FIC}^* = n(\tilde{\mu} - \hat{\mu})^2 + 2n\hat{V}_{\tilde{\mu}} \quad (24.21)$$

$$= n(\tilde{\mu} - \hat{\mu})^2 + 2ns(\tilde{\mu})^2 \quad (24.22)$$

where

$$\hat{V}_{\tilde{\mu}} = \hat{\mathbf{G}}' \left( \mathbf{I}_k - \hat{V}_{\hat{\theta}} \hat{\mathbf{R}} \left( \hat{\mathbf{R}}' \hat{V}_{\hat{\theta}} \hat{\mathbf{R}} \right)^{-1} \hat{\mathbf{R}}' \hat{V}_{\hat{\theta}} \right) \hat{\mathbf{G}}$$

is an estimator of  $\text{var}(\tilde{\mu})$  and  $s(\tilde{\mu}) = \hat{V}_{\tilde{\mu}}^{1/2}$  is a standard error for  $\tilde{\mu}$ .

This means that  $\text{FIC}^*$  can be easily calculated using conventional software without additional programming. The estimator  $\hat{\mu}$  can be calculated from the full model (the long regression), the estimator  $\tilde{\mu}$  and its standard error  $s(\tilde{\mu})$  from the restricted model (the short regression). The formula (24.22) can then be applied to obtain  $\text{FIC}^*$ .

The formula (24.21) also provides an intuitive understanding of the FIC. When we minimize  $\text{FIC}^*$  we are minimizing the variance of the estimator of the target parameter ( $\hat{V}_{\tilde{\mu}}$ ) while not altering the estimate  $\tilde{\mu}$  too much from the unrestricted estimate  $\hat{\mu}$ .

When selecting from amongst just two models, the FIC selects the restricted model if

$$(\tilde{\mu} - \hat{\mu})^2 + 2\hat{V}_{\tilde{\mu}} < 0$$

which is the same as

$$\frac{(\tilde{\mu} - \hat{\mu})^2}{\hat{V}_{\tilde{\mu}}} < 2.$$

The statistic to the left of the inequality is the squared t-statistic in the restricted model for testing the hypothesis that  $\mu$  equals the unrestricted estimator  $\hat{\mu}$ , but ignoring the estimation error in the latter. Thus a simple implementation (when just comparing two models) is to estimate the long and short regressions, take the difference in the two estimates of the coefficient of interest, and compute a t-ratio using the standard error from the short (restricted) regression. If this t-ratio exceeds 1.4 the FIC selects the long regression estimate. If the t-ratio is smaller than 1.4 the FIC selects the short regression estimate.

Claeskens and Hjort motivate the FIC using a local misspecification asymptotic framework. We use a simpler heuristic motivation. First take the unrestricted MLE. Under standard conditions  $\hat{\mu}$  has asymptotic variance  $\mathbf{G}' \mathbf{V}_{\theta} \mathbf{G}$  where  $\mathbf{V}_{\theta} = \mathcal{J}^{-1}$ . As the estimator is asymptotically unbiased it follows that

$$\text{mse}(\hat{\mu}) \simeq \text{var}(\hat{\mu}) \simeq n^{-1} \mathbf{G}' \mathbf{V}_{\theta} \mathbf{G}.$$

Second take the restricted MLE. Under standard conditions  $\tilde{\mu}$  has asymptotic variance

$$\mathbf{G}' \left( \mathbf{V}_{\theta} - \mathbf{V}_{\theta} \mathbf{R} \left( \mathbf{R}' \mathbf{V}_{\theta} \mathbf{R} \right)^{-1} \mathbf{R} \mathbf{V}_{\theta} \right) \mathbf{G}.$$

$\tilde{\mu}$  also has a probability limit, say  $\mu_R$ , which (generally) differs from  $\mu$ . Together we find that

$$\text{mse}(\tilde{\mu}) \simeq B + n^{-1} \mathbf{G}' \left( \mathbf{V}_\theta - \mathbf{V}_\theta \mathbf{R} (\mathbf{R}' \mathbf{V}_\theta \mathbf{R})^{-1} \mathbf{R} \mathbf{V}_\theta \right) \mathbf{G}$$

where  $B = (\mu - \mu_R)^2$ . Subtracting, we find that the Focus is

$$\mathbf{F} \simeq B - n^{-1} \mathbf{G}' \mathbf{V}_\theta \mathbf{R} (\mathbf{R}' \mathbf{V}_\theta \mathbf{R})^{-1} \mathbf{R} \mathbf{V}_\theta \mathbf{G}.$$

A plug-in estimator of  $B$  is  $\hat{B} = (\hat{\mu} - \tilde{\mu})^2$ . However it is biased since

$$\begin{aligned} \mathbb{E}(\hat{B}) &= (\mathbb{E}(\hat{\mu} - \tilde{\mu}))^2 + \text{var}(\hat{\mu} - \tilde{\mu}) \\ &\simeq B + \text{var}(\hat{\mu}) - \text{var}(\tilde{\mu}) \\ &\simeq B + n^{-1} \mathbf{G}' \mathbf{V}_\theta \mathbf{R} (\mathbf{R}' \mathbf{V}_\theta \mathbf{R})^{-1} \mathbf{R} \mathbf{V}_\theta \mathbf{G}. \end{aligned}$$

It follows that an approximately unbiased estimator for  $\mathbf{F}$  is

$$\hat{B} - 2n^{-1} \mathbf{G}' \mathbf{V}_\theta \mathbf{R} (\mathbf{R}' \mathbf{V}_\theta \mathbf{R})^{-1} \mathbf{R} \mathbf{V}_\theta \mathbf{G}.$$

The FIC is obtained by replacing the unknown  $\mathbf{G}$ ,  $\mathbf{R}$ , and  $n^{-1} \mathbf{V}_\theta$  by estimates.

## 24.14 Best Subset and Stepwise Regression

Suppose that we have a set of potential regressors  $\{x_{1i}, \dots, x_{Ki}\}$  where  $K$  is possibly very large, and we want to select a subset of the regressors to use in a regression. Let  $S_m$  denote a subset of the regressors, and let  $m = 1, \dots, M$  denote the set of potential subsets. Given a model selection criterion (e.g. AIC, Mallows, or CV), the best subset model is the one which minimizes the criterion across the  $M$  models. This is conventionally implemented by estimating the  $M$  models and comparing the model selection criteria.

If  $K$  is small this is computationally feasible, but it is not feasible when  $K$  is large. This is because the number of potential subsets is  $M = 2^K$ , which grows quickly with  $K$ . For example,  $K = 10$  implies  $M = 1024$ ,  $K = 20$  implies  $M \geq 1,000,000$ , and  $K = 40$  implies  $M$  exceeds one trillion. It simply does not make sense to contemplate estimating all subset regressions!

If the goal is to find the set of regressors which produces the smallest selection criterion, it seems likely that we should be able to find an approximating set of regressors at much reduced computation cost. Some specific algorithms to implement this goal are as called stepwise, stagewise, and least angle regression. None of these procedures are believed to actually achieve the goal of minimizing any specific selection criterion; rather they are viewed as useful computational approximations. There is also some potential confusion as different authors seem to use the same terms for somewhat different implementations. We use the terms here as described in Hastie, Tibshirani, and Friedman (2008).

In the following descriptions we use  $\text{SSE}(m)$  to refer to the sum of squared residuals from a fitted model, and  $C(m)$  to refer to the selection criterion used for model comparison (AIC is most typically used).

### Backward Stepwise Regression

1. Start with all regressors  $\{x_{1i}, \dots, x_{Ki}\}$  included in the “active set”.
2. For  $m = 0, \dots, K - 1$ 
  - (a) Estimate the regression of  $y_i$  on the active set.
  - (b) Identify the regressor whose omission will have the smallest impact on  $C(m)$ .
  - (c) Put this regressor in slot  $K - m$  and delete from the active set.

- (d) Calculate  $C(m)$  and store in slot  $K - m$ .
3. The model with the smallest value of  $C(m)$  is the selected model.

Backward stepwise regression requires that  $K < n$  so that regression with all variables is feasible. It produces an ordering of the regressors from “most relevant” to “least relevant”. A simplified version is to exit the loop when  $C(m)$  increases. (This may not yield the same result as completing the loop.) For the case of AIC selection, step (b) can be implemented by calculating the classical (homoskedastic) t-ratio for each active regressor and find the regressor with the smallest absolute t-ratio. (See Exercise 24.4.)

### Forward Stepwise Regression

1. Start with the null set  $\{\emptyset\}$  as the “active set” and all regressors  $\{x_{1i}, \dots, x_{Ki}\}$  as the “inactive set”.
2. For  $m = 1, \dots, \min(n - 1, K)$ 
  - (a) Estimate the regression of  $y_i$  on the active set.
  - (b) Identify the regressor in the inactive set whose inclusion will have the largest impact on  $C(m)$ .
  - (c) Put this regressor in slot  $m$  and move it from the inactive to the active set.
  - (d) Calculate  $C(m)$  and store in slot  $m$ .
3. The model with the smallest value of  $C(m)$  is the selected model.

A simplified version is to exit the loop when  $C(m)$  increases. (This may not yield the same answer as completing the loop.) For the case of AIC selection, step (b) can be implemented by finding the regressor in the inactive set with the largest absolute correlation with the residual from step (a). (See Exercise 24.5.)

There are combined algorithms which check both forward and backward movements at each step. The algorithms can also be implemented with the regressors organized into groups (so that all elements are either included or excluded at each step). There are also old-fashioned versions which use significance testing rather than selection criterion (however this is unadvised unless implemented to mimic AIC).

Stepwise regression based on old-fashioned significance testing can be implemented in Stata using the `stepwise` command. If attention is confined to models which include regressors one-at-a-time, AIC selection can be implemented by setting the significance level equal to  $p = 0.32$ . Thus the command `stepwise, pr(.32)` implements backward stepwise regression with the AIC criterion, and `stepwise, pe(.32)` implements forward stepwise regression with the AIC criterion.

Stepwise regression can be implemented in R using the `lars` command.

## 24.15 The MSE of Model Selection Estimators

While model selection intuitively makes sense, it can lead to estimators with poor sampling performance. In this section we show that the mean squared error of estimation is not necessarily improved, and can be considerably worsened, by model selection.

To keep things simple, consider an estimator with an exact normal distribution and known covariance matrix. Normalizing the latter to the identity, we consider the setting

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \mathbf{I}_K)$$

and the class of model selection estimators

$$\hat{\boldsymbol{\theta}}^* = \begin{cases} \hat{\boldsymbol{\theta}} & \text{if } \hat{\boldsymbol{\theta}}' \hat{\boldsymbol{\theta}} > c \\ \mathbf{0} & \text{if } \hat{\boldsymbol{\theta}}' \hat{\boldsymbol{\theta}} \leq c \end{cases}$$

for some  $c$ . AIC sets  $c = 2K$ , BIC sets  $c = K \log(n)$ , and 5% significance testing sets  $c$  to equal the 95% quantile of the  $\chi_K^2$  distribution. It is common to call  $\hat{\boldsymbol{\theta}}^*$  a **post-model-selection (PMS)** estimator.

We can explicitly calculate the MSE of  $\hat{\boldsymbol{\theta}}^*$ .

**Theorem 24.11** If  $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \mathbf{I}_K)$  then

$$\text{mse}(\hat{\boldsymbol{\theta}}^*) = K + (2\lambda - K) F_{K+2}(c, \lambda) - \lambda F_{K+4}(c, \lambda)$$

where  $F_r(x, \lambda)$  is the non-central chi-square distribution function with  $r$  degrees of freedom and non-centrality parameter  $\lambda = \boldsymbol{\theta}'\boldsymbol{\theta}$ .

The proof is given in Section 24.40.

The MSE is determined only by  $K$ ,  $\lambda$ , and  $c$ .  $\lambda = \boldsymbol{\theta}'\boldsymbol{\theta}$  turns out to be an important parameter for the MSE. As the squared Euclidean length, it indexes the magnitude of the coefficient  $\boldsymbol{\theta}$ .

We can see the following limiting cases. If  $\lambda = 0$  then  $\text{mse}(\hat{\boldsymbol{\theta}}^*) = K(1 - F_{K+2}(c, 0))$ . As  $\lambda \rightarrow \infty$  then  $\text{mse}(\hat{\boldsymbol{\theta}}^*) \rightarrow K$ . The unrestricted estimator obtains if  $c = 0$ , in which case  $\text{mse}(\hat{\boldsymbol{\theta}}^*) = K$ . As  $c \rightarrow \infty$ ,  $\text{mse}(\hat{\boldsymbol{\theta}}^*) \rightarrow \lambda$ . The latter fact implies that the PMS estimator based on the BIC has unbounded MSE as  $n \rightarrow \infty$ .

Using Theorem 24.11 we can numerically calculate the MSE. In Figure 24.3 we plot the MSE of a set of estimators for a range of values of  $\lambda$ . The left plot is for  $K = 1$ , the right plot is for  $K = 5$ . The dotted line marks the MSE of the unselected estimator  $\hat{\boldsymbol{\theta}}$  which is invariant to  $\lambda$ . The other estimators plotted are AIC selection ( $c = 2K$ ), 5% significance testing selection (chi-square critical value), and BIC selection ( $c = K \log(n)$ ) for  $n = 200$  and  $n = 1000$ .

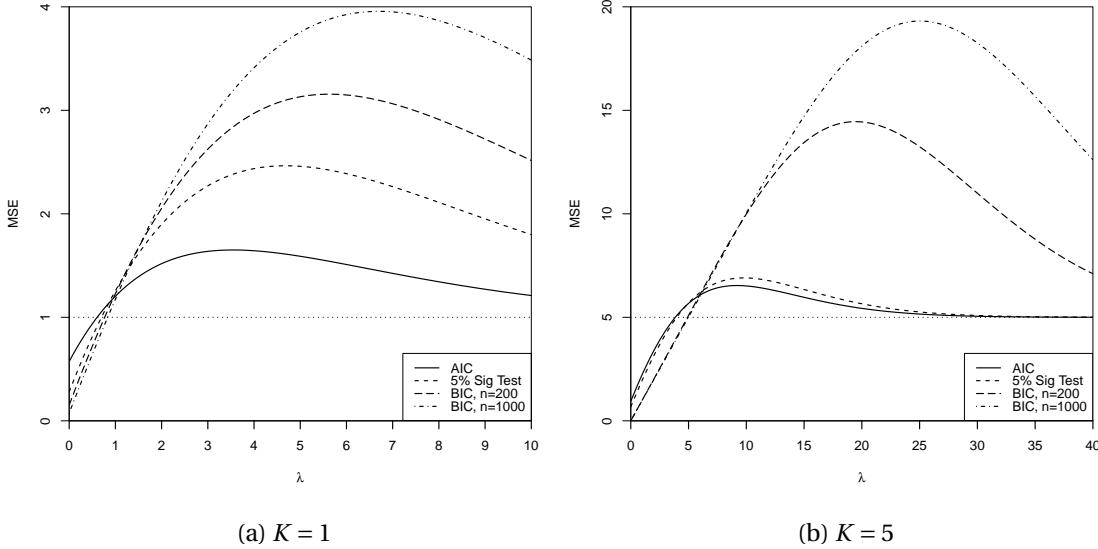


Figure 24.3: Mean Squared Error of Post-Model-Selection Estimators

In the plots you can see that the PMS estimators have lower MSE than the unselected estimator roughly for  $\lambda < K$  but higher MSE for  $\lambda > K$ . The AIC estimator has MSE which is least distorted from the unselected estimator, reaching a peak of about 1.5 for  $K = 1$ . The BIC estimators, however, have very large MSE for larger values of  $\lambda$ , and the distortion is growing as  $n$  increases. The MSE of the selection estimators increases with  $\lambda$  until it reaches a peak, and then slowly decreases and asymptotes back to  $K$ . Furthermore, the MSE of BIC is unbounded as  $n$  diverges. Thus for very large sample sizes the MSE

of a BIC-selected estimator can be a very large multiple of the MSE of the unselected estimator. The plots show that if  $\lambda$  is small then there are advantages to model selection, as MSE can be greatly reduced. However if  $\lambda$  is large then MSE can be greatly increased if BIC is used, and moderately increased if AIC is used. A sensible reading of the plots leads to the practical recommendation to not use the BIC for model selection, and use the AIC with care.

The numerical calculations show that MSE is reduced by selection when  $\lambda$  is small but increased when  $\lambda$  is moderately large. What does this mean in practice?  $\lambda$  is small when  $\theta$  is small, which means the compared models are similar in terms of estimation accuracy. In these contexts model selection can be valuable as it helps select smaller models to improve precision. However when  $\lambda$  is moderately large (which means that  $\theta$  is moderately large) the smaller model has meaningful omitted variable bias, yet the selection criteria have difficulty detecting which model to use. The conservative BIC selection procedure tends to select the smaller model, and thus incurs greater bias resulting in high MSE. These considerations suggest that it is better to use the AIC when selecting among models with similar estimation precision. Unfortunately it is impossible to known *a priori* the appropriate models.

The results of this section may appear to contradict Theorem 24.9 which showed that the BIC is consistent for parsimonious models, as for all  $\lambda > 0$  in the plots the correct parsimonious model is the larger model. Yet BIC is not selecting this model with sufficient frequency to produce a low MSE. There is no contradiction. The consistency of the BIC appears in the lower portion of the plots, where the MSE of the BIC estimator appears to be approximately the straight line  $MSE = \lambda$ . This is the MSE of the restricted estimator. Thus for small  $\lambda$  the BIC properly selects the true model. The fact that the MSE of the AIC estimator somewhat exceeds that of the BIC in this region is illustrating the over-selection property of the AIC.

## 24.16 Inference After Model Selection

Economists are typically interested in inferential questions, such as hypothesis tests and confidence intervals. If an econometric model has been selected by a procedure such as AIC or CV, what are the properties of statistical tests applied to the selected model?

To be concrete consider the regression model  $y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + e_i$  and AIC selection of the variable  $x_{2i}$ . That is, we compare  $y_i = x_{1i}\beta_1 + e_i$  with  $y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + e_i$ . It is not too deep a realization that in this context it is inappropriate to conduct conventional inference for  $\beta_2$  in the selected model. If we select the smaller model there is no estimate of  $\beta_2$ . If we select the larger it is because the t-ratio for  $\beta_2$  exceeds 1.4. The distribution of the t-ratio, conditional on exceeding 1.4, is not conventionally distributed and there seems little point to push this issue further.

The more interesting and subtle question is the impact on inference concerning  $\beta_1$ . This indeed is a context of typical interest. An economist is interested in the impact of  $x_{1i}$  on  $y_i$  given a set of controls  $x_{2i}$ . It is common to select across these controls to find a suitable empirical model. Once this has been obtained we want to make inferential statements about  $\beta_1$ . Has selection over the controls impacted correct inference?

To illustrate the importance of the issue we focus on a stylized setting. The parameter of interest is  $\theta = \mathbb{E}(y) - \beta\mathbb{E}(x)$  where we know  $\text{var}(\bar{y}) = \text{var}(\bar{x}) = 1$ , their correlation is zero, and we know the value of  $\beta$ . We select over the variable  $x$  using an information criterion. This means the PMS estimator is

$$\hat{\theta}^* = \bar{y} - \beta\bar{x}\mathbf{1}(\bar{x}^2 > c).$$

After selection, the standard error is either 1 in the short regression or  $\sqrt{1 + \beta^2}$  in the long regression. Thus the PMS t-ratio can be written as

$$T = \frac{\hat{\theta}^* - \theta}{s(\hat{\theta}^*)} \tag{24.23}$$

where

$$s(\hat{\theta}^*) = \mathbf{1}(\bar{x}^2 \leq c) + \sqrt{1 + \beta^2}\mathbf{1}(\bar{x}^2 > c).$$

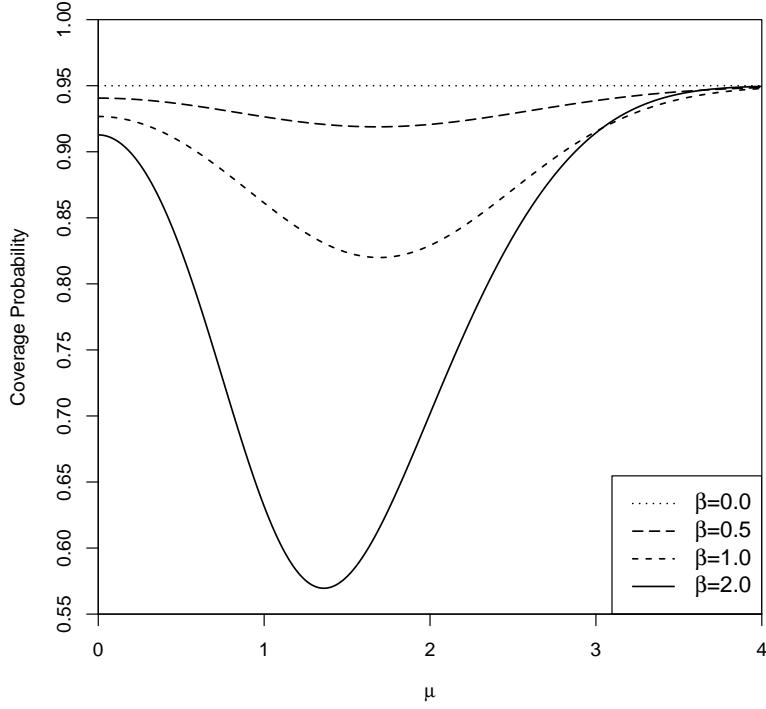


Figure 24.4: Coverage Probabilities of PMS Intervals

A naive hypothesis test compares  $T$  with the normal critical values. A naive confidence interval equals  $\hat{\theta}^* \pm 1.96s(\hat{\theta}^*)$ .

This model is sufficiently simple that we can calculate the distribution of  $T$  explicitly. (An expression<sup>4</sup> for the distribution function is given below.) The distribution is a function only of  $\beta$ ,  $\mu = \mathbb{E}(x)$ , and  $c$ . We focus on AIC for which  $c = 2$ , vary  $\beta$  among 0, 0.5, 1.0, and 2.0, and vary  $\mu$  on a grid between 0 and 4. We plot the coverage probabilities of nominal 95% intervals in Figure 24.4.

The first (dotted) line is the plot of the coverage probability for  $\beta = 0$  as a function of  $\mu$ . We can see that the probability is exactly 95% for all values of  $\mu$ . In this special case the coverage probability is exact. The second through fourth lines are the plots for  $\beta > 0$  as a function of  $\mu$ . These plots all lie below 95% indicating undercovered. As  $\beta$  increases the coverage probability worsens. The distortion is hump-shaped in  $\mu$ , with the largest distortion for  $\mu \in [1, 2]$ . This is the region where it is most difficult to detect if  $\mu = 0$  or not. The distortion is increasing in  $\beta$ , with the worst-case coverage shown equalling 57% (far from the nominal 95%). The coverage can be made worse, however, either by increasing  $\beta$  or  $c$ . It is also useful to observe that the coverage is even distorted at  $\mu = 0$  (which is the ideal case where the restricted estimator is optimal). This distortion is the effect of the AIC property of over-selection.

The message from this display is that inference after model selection is problematic. Conventional inference procedures do not have the same distributions as expected from a non-selection theory, and the distortions are potentially unbounded.

<sup>4</sup>The integral in the expression is evaluated numerically.

**Theorem 24.12** The distribution function of  $T$  defined in (24.23) is

$$\Pr(T \leq t) = \Phi(t) + \Phi(t - \beta\mu) (\Phi(\sqrt{c} - \mu) - \Phi(-\sqrt{c} - \mu)) \\ - \int_{-\sqrt{c}-\mu}^{\sqrt{c}-\mu} \Phi\left(\beta s + t\sqrt{1+\beta^2}\right) \phi(s) ds$$

where  $\phi(t)$  and  $\Phi(t)$  are the standard normal pdf and cdf functions.

A proof of the theorem is provided in Section 24.40.

## 24.17 Empirical Illustration

We illustrate the model selection methods using an empirical application. Take the CPS dataset and consider the sub-sample of Asian women, which has  $n = 1149$  observations. Consider a log wage regression with primary interest focused on the return to experience, measured as the percentage difference between expected wages between 0 and 30 years of experience. We consider and compare nine least squares regressions. All include an indicator for *married* and three indicators for the *region*. The estimated models range in complexity concerning how the impact of education and experience are modeled.

Table 24.1: Estimates of Return to Experience among Asian Women

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Return	13%	22%	20%	29%	40%	37%	33%	47%	45%
s.e.	7	8	7	11	11	11	17	18	17
BIC	956	<b>907</b>	924	964	913	931	977	925	943
AIC	915	861	858	914	858	<b>855</b>	916	860	857
CV	405	387	386	405	385	<b>385</b>	406	387	386
FIC*	86	48	53	58	<b>32</b>	34	86	71	68
Education	College	Spline	Dummy	College	Spline	Dummy	College	Spline	Dummy
Experience	2	2	2	4	4	4	6	6	6

Terms for experience:

- Models 1-3 include experience and its square.
- Models 4-6 include powers of experience up to the power 4.
- Models 7-9 include powers of experience up to the power 6.

Terms for education:

- Models 1, 4, and 7 include a single dummy variable *college* indicating that years of education is 16 or higher.
- Models 2, 5, and 8 is a linear spline with a single knot at 9 years of education.
- Models 3, 6, and 9 include six dummy variables, for education equalling 12, 13, 14, 16, 18, and 20.

Table 24.1 reports some key estimates from the nine models. Reported are the estimate of the return to experience as a percentage wage difference, its standard error (HC1), the BIC, AIC, CV, and FIC\*, the latter treating the return to experience as the focus. What we can see is that the estimates vary meaningfully, ranging from 13% to 47%. Some of the estimates also have moderately large standard errors. (In most models the return to experience is “statistically significant”, but by large standard errors we mean that it is difficult to pin down the precise value of the return to experience.) We can also see that the most important factors impacting the magnitude of the point estimate is going beyond the quadratic specification for experience, and going beyond the simplest specification for education. Another thing to notice is that the standard errors are most affected by the number of experience terms.

The BIC picks a parsimonious model with the linear spline in education and a quadratic in experience. The AIC and CV select a less parsimonious model with the full dummy specification for education and a 4<sup>th</sup> order polynomial in experience. The CV criterion, however, has similar values across six of the nine models. The FIC\* selects an intermediate model, with a linear spline in education and a 4<sup>th</sup> order polynomial in experience.

When selecting a model using information criteria it is useful to examine several criteria. In applications decisions should be made by a combination of judgment as well as the formal criteria. In this case the cross-validation criterion selects model 6 which has the estimate of 37%, but near-similar values of the CV criterion are obtained by models 3 and 9, which have the estimates 20% and 45%. The FIC, which focuses on this specific coefficient, selects model 5, which has the point estimate 40% which is similar to the CV-selected model. Overall based on this evidence the CV-selected model and its point estimate of 37% seems an appropriate choice. However, the uncertainty reflected by the flatness of the CV criterion suggests that uncertainty remains in the choice of specification.

## 24.18 Shrinkage Methods

Shrinkage methods are a broad class of estimators which reduce variance by moving an estimator  $\hat{\boldsymbol{\theta}}$  towards a pre-selected point such as the zero vector. In high dimensions the reduction in variance more than compensates for the increase in bias, resulting in improved efficiency when measured by mean squared error.

The simplest shrinkage estimator takes the form  $\tilde{\boldsymbol{\theta}} = (1 - w)\hat{\boldsymbol{\theta}}$  for some shrinkage weight  $w \in [0, 1]$ . Setting  $w = 0$  we obtain  $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$  (no shrinkage) and setting  $w = 1$  we obtain  $\tilde{\boldsymbol{\theta}} = 0$  (full shrinkage). It is straightforward to calculate the MSE of this estimator. Assume  $\hat{\boldsymbol{\theta}} \sim (\boldsymbol{\theta}, \mathbf{V})$ . Then  $\tilde{\boldsymbol{\theta}}$  has bias

$$\text{bias}(\tilde{\boldsymbol{\theta}}) = \mathbb{E}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{\theta} = -w\boldsymbol{\theta}, \quad (24.24)$$

variance

$$\text{var}(\tilde{\boldsymbol{\theta}}) = (1 - w)^2 \mathbf{V}, \quad (24.25)$$

and weighted mean squared error (using the weight matrix  $\mathbf{W} = \mathbf{V}^{-1}$ )

$$\text{wmse}(\tilde{\boldsymbol{\theta}}) = K(1 - w)^2 + w^2 \lambda \quad (24.26)$$

where  $\lambda = \boldsymbol{\theta}' \mathbf{V}^{-1} \boldsymbol{\theta}$ .

**Theorem 24.13** If  $\hat{\boldsymbol{\theta}} \sim (\boldsymbol{\theta}, \mathbf{V})$  and  $\tilde{\boldsymbol{\theta}} = (1 - w)\hat{\boldsymbol{\theta}}$  then

1.  $\text{wmse}(\tilde{\boldsymbol{\theta}}) < \text{wmse}(\hat{\boldsymbol{\theta}})$  if  $0 < w < 2K/(K + \lambda)$ .
2.  $\text{wmse}(\tilde{\boldsymbol{\theta}})$  is minimized by the shrinkage weight  $w_0 = K/(K + \lambda)$ .
3. The minimized WMSE is  $\text{wmse}(\tilde{\boldsymbol{\theta}}) = K\lambda/(K + \lambda)$ .

For the proof see Exercise 24.7.

Part 1 of the theorem shows that the shrinkage estimator has reduced WMSE for a range of values of the shrinkage weight  $w$ . Part 2 of the theorem shows that the WMSE-minimizing shrinkage weight is a simple function of  $K$  and  $\lambda$ . The latter is a measure of the magnitude of  $\boldsymbol{\theta}$  relative to the estimation variance. When  $\lambda$  is large (the coefficients are large) then the optimal shrinkage weight  $w_0$  is small; when  $\lambda$  is small (the coefficients are small) then the optimal shrinkage weight  $w_0$  is large. Part 3 calculates the associated optimal WMSE. This can be substantially less than the WMSE of the original estimator  $\hat{\boldsymbol{\theta}}$ . For example, if  $\lambda = K$  then  $\text{wmse}(\tilde{\boldsymbol{\theta}}) = K/2$ , one-half the WMSE of the original estimator.

To construct the optimal shrinkage weight we need the unknown  $\lambda$ . An unbiased estimator is  $\hat{\lambda} = \hat{\boldsymbol{\theta}}' \mathbf{V}^{-1} \hat{\boldsymbol{\theta}} - K$  (see Exercise 24.8) implying the shrinkage weight

$$\hat{w} = \frac{K}{\hat{\boldsymbol{\theta}}' \mathbf{V}^{-1} \hat{\boldsymbol{\theta}}}. \quad (24.27)$$

Replacing  $K$  with a free parameter  $c$  (which we call the shrinkage coefficient) we obtain

$$\tilde{\boldsymbol{\theta}} = \left(1 - \frac{c}{\hat{\boldsymbol{\theta}}' \mathbf{V}^{-1} \hat{\boldsymbol{\theta}}}\right) \hat{\boldsymbol{\theta}}. \quad (24.28)$$

This class of estimators is often called a Stein-rule estimator.

This estimator has many appealing properties. It can be viewed as a smoothed selection estimator. The quantity  $\hat{\boldsymbol{\theta}}' \mathbf{V}^{-1} \hat{\boldsymbol{\theta}}$  is a Wald statistic for the hypothesis  $H_0 : \boldsymbol{\theta} = \mathbf{0}$ . Thus when this Wald statistic is large (when the evidence suggests the hypothesis of a zero coefficient is false) the shrinkage estimator is close to the original estimator  $\hat{\boldsymbol{\theta}}$ . However when this Wald statistic is small (when the evidence is consistent with the hypothesis of a zero coefficient) then the shrinkage estimator moves the original estimator towards zero.

## 24.19 James-Stein Shrinkage Estimator

James and Stein (1961) made the following discovery.

**Theorem 24.14** Assume that  $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \mathbf{V})$ ,  $\tilde{\boldsymbol{\theta}}$  is defined in (24.28), and  $K > 2$ .

1. If  $0 < c < 2(K - 2)$  then  $\text{wmse}(\tilde{\boldsymbol{\theta}}) < \text{wmse}(\hat{\boldsymbol{\theta}})$ .
2. The WMSE is minimized by setting  $c = K - 2$ .

Theorem 24.14 follows fairly directly from Theorem 24.15 below. See Exercise 24.9.

This result stunned the world of statistics. Part 1 shows that the shrinkage estimator has strictly smaller WMSE for all values of the parameters, and thus dominates the original estimator. The latter is the MLE, so this result shows that the MLE is dominated and thus inadmissible. This is a stunning result because it had previously been assumed that it would be impossible to find an estimator which dominates the MLE.

Theorem 24.14 critically depends on the condition  $K > 2$ . This means that shrinkage achieves uniform improvements only in dimensions three or larger. In smaller dimensions shrinkage can reduce MSE only over subsets of the parameter space.

The minimizing choice for the shrinkage coefficient  $c = K - 2$  leads to what is commonly known as the James-Stein estimator

$$\tilde{\boldsymbol{\theta}} = \left(1 - \frac{K-2}{\hat{\boldsymbol{\theta}}' \mathbf{V}^{-1} \hat{\boldsymbol{\theta}}}\right) \hat{\boldsymbol{\theta}}.$$

In practice  $V$  is unknown so we substitute an estimator  $\hat{V}$ . This leads to

$$\tilde{\boldsymbol{\theta}}_{JS} = \left(1 - \frac{K-2}{\hat{\boldsymbol{\theta}}' \hat{V}^{-1} \hat{\boldsymbol{\theta}}}\right) \hat{\boldsymbol{\theta}}$$

which is fully feasible as it does not depend on unknowns or tuning parameters. The substitution of  $\hat{V}$  for  $V$  can be justified by finite sample or asymptotic arguments.

It is possible to explicitly calculate the MSE of the James-Stein estimator.

**Theorem 24.15** Under the assumptions of Theorem 24.14

$$wmse(\tilde{\boldsymbol{\theta}}) = K - c(2(K-2) - c) J_K(\lambda)$$

where

$$J_K(\lambda) = \mathbb{E}(Q_K^{-1}). \quad (24.29)$$

and  $Q_K \sim \chi_K^2(\lambda)$ , a non-central chi-square random variable.

Using Theorem 24.15 (and the formula for  $J_K(\lambda)$  given in Theorem 24.17 below) we can calculate the MSE of  $\tilde{\boldsymbol{\theta}}$ . In Figure 24.5 we plot  $wmse(\tilde{\boldsymbol{\theta}})/K$  as a function of  $\lambda/K$  for  $K = 4, 6, 12$ , and  $48$ . The plots are uniformly below 1 (the normalized WMSE of the MLE) and substantially so for small and moderate values of  $\lambda$ . The WMSE functions fall as  $K$  increases, demonstrating that the MSE reductions are more substantial when  $K$  is large.

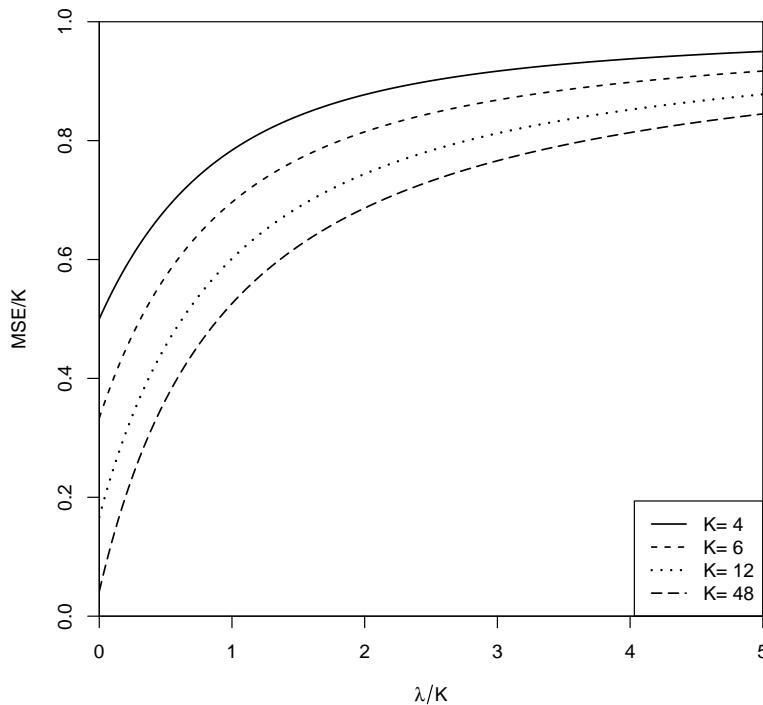


Figure 24.5: WMSE of James-Stein Estimator

## 24.20 Derivation of James-Stein Theorem\*

This section presents the technical derivation of the James-Stein theory and can be skipped by most readers.

While technical, the derivation of Theorem 24.15 is reasonably straightforward. The key is a simple yet famous application of integration by parts.

**Theorem 24.16** (Stein's Lemma) If  $Z \sim N(\boldsymbol{\theta}, V)$  and  $\mathbf{g}(z) : \mathbb{R}^K \rightarrow \mathbb{R}^K$  is absolutely continuous then

$$\mathbb{E}(\mathbf{g}(Z)' V^{-1} (Z - \boldsymbol{\theta})) = \mathbb{E} \operatorname{tr}\left(\frac{\partial}{\partial z} \mathbf{g}(Z)'\right)$$

We prove Stein's Lemma under the simplification that  $V = I_K$ . Since the multivariate normal density is  $\phi(\mathbf{x}) = (2\pi)^{-K/2} \exp(-\mathbf{x}' \mathbf{x}/2)$  then

$$\frac{\partial}{\partial \mathbf{x}} \phi(\mathbf{x} - \boldsymbol{\theta}) = -(\mathbf{x} - \boldsymbol{\theta}) \phi(\mathbf{x} - \boldsymbol{\theta}).$$

By integration by parts

$$\begin{aligned} \mathbb{E}(\mathbf{g}(Z)' (Z - \boldsymbol{\theta})) &= \int \mathbf{g}(\mathbf{x})' (\mathbf{x} - \boldsymbol{\theta}) \phi(\mathbf{x} - \boldsymbol{\theta}) d\mathbf{x} \\ &= - \int \mathbf{g}(\mathbf{x})' \frac{\partial}{\partial \mathbf{x}} \phi(\mathbf{x} - \boldsymbol{\theta}) d\mathbf{x} \\ &= \int \operatorname{tr}\left(\frac{\partial}{\partial \mathbf{x}} \mathbf{g}(\mathbf{x})'\right) \phi(\mathbf{x} - \boldsymbol{\theta}) d\mathbf{x} \\ &= \mathbb{E} \operatorname{tr}\left(\frac{\partial}{\partial z} \mathbf{g}(Z)'\right). \end{aligned}$$

This is the stated result.

We now derive Theorem 24.15 allowing for general  $V$ . Using the definition of  $\tilde{\boldsymbol{\theta}}$  and expanding the quadratic

$$\begin{aligned} \text{wmse}(\tilde{\boldsymbol{\theta}}) &= \mathbb{E}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})' V^{-1} (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\ &= \mathbb{E}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \frac{c}{\hat{\boldsymbol{\theta}}' V^{-1} \hat{\boldsymbol{\theta}}}\right)' V^{-1} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \frac{c}{\hat{\boldsymbol{\theta}}' V^{-1} \hat{\boldsymbol{\theta}}}\right) \\ &= \mathbb{E}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})' V^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + c^2 \mathbb{E}\left(\frac{1}{\hat{\boldsymbol{\theta}}' V^{-1} \hat{\boldsymbol{\theta}}}\right) - 2c \mathbb{E}\left(\frac{\hat{\boldsymbol{\theta}}' V^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})}{\hat{\boldsymbol{\theta}}' V^{-1} \hat{\boldsymbol{\theta}}}\right) \\ &= K + c^2 J_K(\lambda) - 2c \mathbb{E}(\mathbf{g}(\hat{\boldsymbol{\theta}})' V^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})) \end{aligned}$$

using the facts  $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})' V^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim \chi_K^2$  and  $\hat{\boldsymbol{\theta}}' V^{-1} \hat{\boldsymbol{\theta}} \sim \chi_K^2(\lambda)$ , and defining  $\mathbf{g}(\mathbf{x}) = \mathbf{x}(\mathbf{x}' V^{-1} \mathbf{x})^{-1}$ . Using the rules of matrix differentiation

$$\operatorname{tr}\left(\frac{\partial}{\partial \mathbf{x}} \mathbf{g}(\mathbf{x})'\right) = \operatorname{tr}\left(I_K (\mathbf{x}' V^{-1} \mathbf{x})^{-1} - 2V^{-1} \mathbf{x} \mathbf{x}' (\mathbf{x}' V^{-1} \mathbf{x})^{-2}\right) = \frac{K-2}{\mathbf{x}' V^{-1} \mathbf{x}}.$$

Using Stein's Lemma, the assumption  $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, V)$ , and  $\hat{\boldsymbol{\theta}}' V^{-1} \hat{\boldsymbol{\theta}} \sim \chi_K^2(\lambda)$

$$\begin{aligned} \mathbb{E}(\mathbf{g}(\hat{\boldsymbol{\theta}})' V^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})) &= \mathbb{E} \operatorname{tr}\left(\frac{\partial}{\partial \mathbf{x}} \mathbf{g}(\hat{\boldsymbol{\theta}})'\right) \\ &= \mathbb{E}\left(\frac{K-2}{\hat{\boldsymbol{\theta}}' V^{-1} \hat{\boldsymbol{\theta}}}\right) \\ &= (K-2) J_K(\lambda). \end{aligned}$$

Together we find

$$\begin{aligned}\text{wmse}(\tilde{\boldsymbol{\theta}}) &= K + c^2 J_K(\lambda) - 2c(K-2)J_K(\lambda) \\ &= K - c(2(K-2) - c) J_K(\lambda)\end{aligned}$$

as stated.

We now discuss calculation of  $J_K(\lambda)$  and its trimmed version

$$J_K(c, \lambda) = \mathbb{E}(\chi_K^2(\lambda)^{-1} \mathbf{1}(\chi_K^2(\lambda) < c)). \quad (24.30)$$

**Theorem 24.17** For  $J_K(\lambda)$  and  $J_K(c, \lambda)$  defined in (24.29) and (24.30), for  $K > 2$ ,

$$\begin{aligned}J_K(\lambda) &= \sum_{i=0}^{\infty} \frac{e^{-\lambda/2}}{i!} \frac{(\lambda/2)^i}{K+2i-2} \\ J_K(c, \lambda) &= \sum_{i=0}^{\infty} \frac{e^{-\lambda/2}}{i!} \frac{(\lambda/2)^i}{K+2i-2} F_{K+2i-2}(K-2, \lambda)\end{aligned}$$

where  $F_r(x, \lambda)$  is the distribution function of  $\chi_r^2(\lambda)$ .

Notice that when  $\lambda = 0$  the expressions simplify to  $J_K(0) = 1/(K-2)$  and  $J_K(c, 0) = F_{K-2}(K-2, 0)/(K-2)$ .

The expressions in Theorem 24.17 are convergent series. For our reported calculations we truncate after the first 200 terms.

To demonstrate Theorem 24.17, recall that the density of the non-central chi-square is

$$f_K(x, \lambda) = \sum_{i=0}^{\infty} \frac{e^{-\lambda/2}}{i!} \left(\frac{\lambda}{2}\right)^i f_{K+2i}(x)$$

where

$$f_r(x) = \frac{x^{r/2-1} e^{-x/2}}{2^{r/2} \Gamma(r/2)}$$

is the  $\chi_r^2$  density. Simple manipulations reveal that for  $r > 2$ ,  $x^{-1} f_r(x) = (r-2)^{-1} f_{r-2}(x)$ . Then

$$\int_0^c x^{-1} f_r(x) dx = \frac{1}{r-2} \int_0^c f_{r-2}(x) dx = \frac{F_{r-2}(c)}{r-2}.$$

Thus for  $K > 2$

$$\begin{aligned}J_K(c, \lambda) &= \int_0^c x^{-1} f_K(x, \lambda) dx \\ &= \sum_{i=0}^{\infty} \frac{e^{-\lambda/2}}{i!} \left(\frac{\lambda}{2}\right)^i \int_0^c x^{-1} f_{K+2i}(x) dx \\ &= \sum_{i=0}^{\infty} \frac{e^{-\lambda/2}}{i!} \left(\frac{\lambda}{2}\right)^i \frac{F_{K+2i-2}(c)}{K+2i-2}\end{aligned}$$

as claimed. This specializes to the statement for  $J_K(\lambda)$  when  $c \rightarrow \infty$ .

## 24.21 Interpretation of the Stein Effect

The James-Stein Theorem appears to conflict with classical statistical theory. The original estimator  $\hat{\boldsymbol{\theta}}$  is the maximum likelihood estimator. It is unbiased. It is minimum variance unbiased. It is Cramer-Rao efficient. It achieves the minimax efficiency bound. How can it be that the James-Stein shrinkage estimator achieves uniformly smaller mean squared error?

Part of the answer is that classical theory has caveats. The Cramer-Rao Theorem, for example, restricts attention to unbiased estimators, and thus precludes consideration of shrinkage estimators. The James-Stein estimator has reduced MSE, but is not Cramer-Rao efficient since it is biased. Therefore the James-Stein Theorem does not conflict with the Cramer-Rao Theorem. Rather, they are complementary results. On the one hand, the Cramer-Rao Theorem describes the best possible variance when unbiasedness is an important property for estimation. On the other hand, the James-Stein Theorem shows that if unbiasedness is not a critical property, but instead MSE is important, then there are better estimators than the MLE.

The James-Stein Theorem may also appear to conflict with our results from Section 24.15 which showed that selection estimators do not achieve uniform MSE improvements over the MLE. This may appear to be a conflict since the James-Stein estimator has a similar form to a selection estimator. The difference is that selection estimators are **hard threshold** rules – they are discontinuous functions of the data – while the James-Stein estimator is a **soft threshold** rule – it is a continuous function of the data. Hard thresholding tends to result in high variance; soft thresholding tends to result in low variance. The James-Stein estimator is able to achieve reduced variance because it is a soft threshold function.

The MSE improvements achieved by the James-Stein estimator are greatest when  $\lambda$  is small. This occurs when the parameters  $\boldsymbol{\theta}$  are small in magnitude relative to the estimation variance  $\mathbf{V}$ . This means that the user needs to choose the centering point wisely.

## 24.22 Positive Part Estimator

The simple James-Stein estimator has the odd property that it can “over-shrink”. When  $\hat{\boldsymbol{\theta}}' \mathbf{V}^{-1} \hat{\boldsymbol{\theta}} < K - 2$  then  $\tilde{\boldsymbol{\theta}}$  has opposite sign with  $\hat{\boldsymbol{\theta}}$ . This does not make sense and suggests that further improvements can be made. The standard solution is to use “positive-part” trimming by bounding the shrinkage weight (24.27) below one. This estimator can be written as

$$\begin{aligned}\tilde{\boldsymbol{\theta}}^+ &= \begin{cases} \tilde{\boldsymbol{\theta}}, & \hat{\boldsymbol{\theta}}' \mathbf{V}^{-1} \hat{\boldsymbol{\theta}} \geq K - 2 \\ 0, & \hat{\boldsymbol{\theta}}' \mathbf{V}^{-1} \hat{\boldsymbol{\theta}} < K - 2 \end{cases} \\ &= \left(1 - \frac{K - 2}{\hat{\boldsymbol{\theta}}' \mathbf{V}^{-1} \hat{\boldsymbol{\theta}}}\right)_+ \hat{\boldsymbol{\theta}}\end{aligned}$$

where  $(a)_+ = \max[a, 0]$  is the “positive-part” function. Alternatively, it can be written as

$$\tilde{\boldsymbol{\theta}}^+ = \hat{\boldsymbol{\theta}} - \left(\frac{K - 2}{\hat{\boldsymbol{\theta}}' \mathbf{V}^{-1} \hat{\boldsymbol{\theta}}}\right)_1 \hat{\boldsymbol{\theta}}$$

where  $(a)_1 = \min[a, 1]$

The positive part estimator simultaneously performs “selection” as well as “shrinkage”. If  $\hat{\boldsymbol{\theta}}' \mathbf{V}^{-1} \hat{\boldsymbol{\theta}}$  is sufficiently small,  $\tilde{\boldsymbol{\theta}}^+$  “selects” 0. When  $\hat{\boldsymbol{\theta}}' \mathbf{V}^{-1} \hat{\boldsymbol{\theta}}$  is of moderate size,  $\tilde{\boldsymbol{\theta}}^+$  shrinks  $\hat{\boldsymbol{\theta}}$  towards zero. When  $\hat{\boldsymbol{\theta}}' \mathbf{V}^{-1} \hat{\boldsymbol{\theta}}$  is very large,  $\tilde{\boldsymbol{\theta}}^+$  is close to the original estimator  $\hat{\boldsymbol{\theta}}$ .

Consistent with our intuition, the positive part estimator has uniformly lower WMSE than the unadjusted James-Stein estimator. We can also present an explicit expression for the WMSE.

**Theorem 24.18** Under the assumptions of Theorem 24.14

$$\text{wmse}(\tilde{\boldsymbol{\theta}}^+) < \text{wmse}(\tilde{\boldsymbol{\theta}}). \quad (24.31)$$

The WMSE has the explicit expression

$$\begin{aligned} \text{wmse}(\tilde{\boldsymbol{\theta}}^+) = & \text{wmse}(\tilde{\boldsymbol{\theta}}) - 2KF_K(K-2, \lambda) + KF_{K+2}(K-2, \lambda) \\ & + \lambda F_{K+4}(K-2, \lambda) + (K-2)^2 J_K(K-2, \lambda) \end{aligned} \quad (24.32)$$

where  $F_r(x, \lambda)$  is the non-central chi-square distribution function and  $J_K(c, \lambda)$  is defined in (24.30).

The proof is in Section 24.40.

To illustrate the improvement in MSE, Figure 24.6 plots the WMSE of the unadjusted and positive-part James-Stein estimators for  $K = 4$  and  $K = 12$ . For  $K = 4$  the positive-part estimator has meaningfully reduced MSE relative to the unadjusted estimator, especially for small values of  $\lambda$ . For  $K = 12$  the difference between the estimators is smaller.

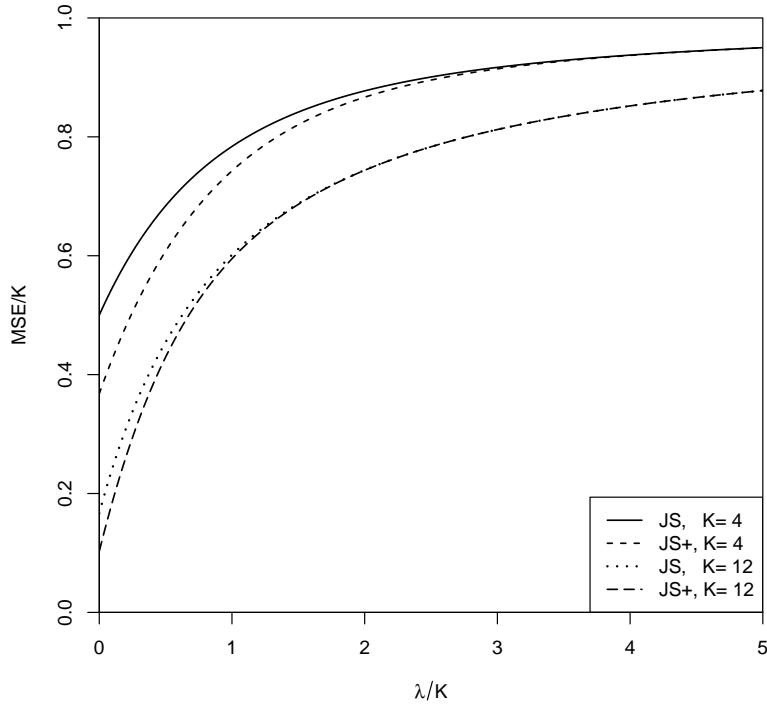


Figure 24.6: WMSE of Positive-Part James-Stein Estimator

In summary, the positive-part transformation is an important improvement over the unadjusted James-Stein estimator. It is more reasonable, and reduces the mean squared error. The broader message is that imposing boundary conditions can often help to regularize estimators and improve their performance.

## 24.23 Shrinkage Towards Restrictions

The classical James-Stein estimator does not have much use in applications because it is rare that a user wishes to shrink an entire parameter vector towards a specific point. Rather, it is more common to wish to shrink a parameter vector towards a set of restrictions. Here are a few examples:

1. Shrink a long regression towards a short regression.
2. Shrink a regression towards an intercept-only model.
3. Shrink the regression coefficients towards a set of restrictions.
4. Shrink a set of estimates (or coefficients) towards their common mean.
5. Shrink a set of estimates (or coefficients) towards a parametric model.
6. Shrink a nonparametric series model towards a parametric model.

The way to think generally about these applications is that the researcher wants to allow for generality with the large model, but believes that the smaller model may be a useful approximation. A shrinkage estimator allows the data to smoothly select between these two options depending on the strength of information for the two specifications.

Let  $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, V)$  be the original estimator, for example a set of regression coefficient estimates. The normality assumption is used for the exact theory, but can be justified based on an asymptotic approximation as well. The researcher considers a set of  $q > 2$  linear restrictions which can be written as  $\mathbf{R}'\boldsymbol{\theta} = \mathbf{r}$  where  $\mathbf{R}$  is  $K \times q$  and  $\mathbf{r}$  is  $q \times 1$ . A minimum distance estimator for  $\boldsymbol{\theta}$  is

$$\hat{\boldsymbol{\theta}}_R = \hat{\boldsymbol{\theta}} - \mathbf{V}\mathbf{R}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}(\mathbf{R}'\hat{\boldsymbol{\theta}} - \mathbf{r}).$$

The Stein-rule estimator (with positive-part trimming) is

$$\tilde{\boldsymbol{\theta}}^+ = \hat{\boldsymbol{\theta}} - \left( \frac{q-2}{(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_R)' \mathbf{V}^{-1} (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_R)} \right)_1 (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_R).$$

As before, the function  $(a)_1 = \min[a, 1]$  bounds the shrinkage weight below one.

**Theorem 24.19** Under the assumptions of Theorem 24.14, if  $q > 2$  then

$$\text{wmse}(\tilde{\boldsymbol{\theta}}^+) < \text{wmse}(\hat{\boldsymbol{\theta}}).$$

Thus the shrinkage estimator achieves uniformly smaller MSE if the number of restrictions is three or greater. The number of restrictions  $q$  plays the same role as the number of parameters  $K$  in the classical James-Stein estimator. The same theoretical properties apply. Shrinkage achieves greater gains when there are more restrictions  $q$ , and achieves greater gains when the restrictions are close to being satisfied in the population. If the imposed restrictions are far from satisfied then the shrinkage estimator will have similar performance as the original estimator. It is therefore important to select the restrictions carefully.

In practice the variance matrix  $V$  is unknown so it is replaced by an estimator  $\hat{V}$ . Thus the feasible version of the estimators equal

$$\hat{\boldsymbol{\theta}}_R = \hat{\boldsymbol{\theta}} - \hat{\mathbf{V}}\mathbf{R}(\mathbf{R}'\hat{\mathbf{V}}\mathbf{R})^{-1}(\mathbf{R}'\hat{\boldsymbol{\theta}} - \mathbf{r})$$

and

$$\tilde{\boldsymbol{\theta}}^+ = \hat{\boldsymbol{\theta}} - \left( \frac{q-2}{J} \right)_1 (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_R) \quad (24.33)$$

where

$$J = (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_R)' \hat{V}^{-1} (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_R).$$

It is insightful to notice that  $J$  is the minimum distance statistic for the test of the hypothesis  $\mathbb{H}_0 : \mathbf{R}'\boldsymbol{\theta} = \mathbf{r}$  against  $\mathbb{H}_1 : \mathbf{R}'\boldsymbol{\theta} \neq \mathbf{r}$ . Thus the degree of shrinkage is a smoothed version of the standard test of the restrictions. When  $J$  is large (so the evidence indicates that the restrictions are false) the shrinkage estimator is close to the unrestricted estimator  $\hat{\boldsymbol{\theta}}$ . When  $J$  is small (so the evidence indicates that the restrictions could be correct) the shrinkage estimator equals the restricted estimator  $\hat{\boldsymbol{\theta}}_R$ . For intermediate values of  $J$  the shrinkage estimator shrinks  $\hat{\boldsymbol{\theta}}$  towards  $\hat{\boldsymbol{\theta}}_R$ .

We can substitute for  $J$  any similar asymptotically chi-square statistic. This includes the Wald, Likelihood Ratio, and Score tests. This also includes the F statistic (which is commonly produced by statistical software) if we multiply by  $q$ . These substitutions do not produce the same exact finite sample distribution, but are asymptotically equivalent.

In linear regression we have some very convenient simplifications available. In general,  $\hat{V}$  can be a heteroskedastic-robust or cluster-robust covariance matrix estimator. However, if the dimension  $K$  of the unrestricted estimator is quite large, or has sparse dummy variables, then these covariance matrix estimators are ill-behaved and it may be better to use a classical covariance matrix estimator to perform the shrinkage. If this is done then  $\hat{V} = (\mathbf{X}'\mathbf{X})^{-1}s^2$ ,  $\hat{\boldsymbol{\theta}}_R$  is the constrained least-squares estimator (in most applications the least squares estimator of the short regression), and  $J$  is a conventional (homoskedastic) Wald statistic for a test of the restrictions. We can write the latter in F statistic form

$$J = \frac{n(\hat{\sigma}_R^2 - \hat{\sigma}^2)}{s^2} \tag{24.34}$$

where  $\hat{\sigma}_R^2$  and  $\hat{\sigma}^2$  are the least squares error variance estimators from the restricted and unrestricted models. Thus the shrinkage weight  $((q-2)/J)_1$  can be easily calculated from standard regression output.

## 24.24 Group James-Stein

The James-Stein estimator can be applied to groups (blocks) of parameters. Suppose we have the parameter vector  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_G)$  partitioned into  $G$  groups, each of dimension  $K_g \geq 3$ . We have a standard estimator  $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \dots, \hat{\boldsymbol{\theta}}_G)$  (for example, least squares regression or MLE) with variance matrix  $V$ . Let The group James-Stein estimator is

$$\begin{aligned} \tilde{\boldsymbol{\theta}} &= (\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2, \dots, \tilde{\boldsymbol{\theta}}_G) \\ \tilde{\boldsymbol{\theta}}_g &= \hat{\boldsymbol{\theta}}_g \left( 1 - \frac{K_g - 2}{\hat{\boldsymbol{\theta}}_g' V_g^{-1} \hat{\boldsymbol{\theta}}_g} \right)_+ \end{aligned}$$

where  $V_g$  is the  $g^{th}$  diagonal block of  $V$ . A feasible version of the estimator replaces  $V$  with  $\hat{V}$  and  $V_g$  with  $\hat{V}_g$ .

The group James-Stein estimator shrinks each block of coefficients separately. The advantage (relative to the classical James-Stein estimator) is that this allows the shrinkage weight to vary across blocks. Some parameter blocks can use a large amount of shrinkage while others a minimal amount. Since the positive-part trimming is used, the estimator simultaneously performs shrinkage and selection. Blocks with small effects will be shrunk to zero and eliminated. The disadvantage of the estimator is that the benefits of shrinkage can be reduced since the shrinkage dimension has been reduced. The trade-off between these factors will depend on how heterogeneous the optimal shrinkage weight varies across the parameters.

The groups should be selected based on two criteria. First, they should be selected so that the groups separate variables by expected amount of shrinkage. Thus coefficients which are expected to be “large” relative to their estimation variance should be grouped together, and coefficients which are expected to be “small” should be grouped together. This will allow the estimated shrinkage weights to vary according

to the group. For example, a researcher may expect high-order coefficients in a polynomial regression to be small relative to their estimation variance. Hence it is appropriate to group the polynomial variables into “low order” and “high order”. Second, the groups should be selected so that the researcher’s loss (utility) is separable across groups of coefficients. This is because the optimality theory (given below) relies on the assumption that the loss is separable. To understand the implications of these recommendations consider a wage regression. Our interpretation of the education and experience coefficients are separable if we use them for separate purposes, such as for estimation of the return to education and the return to experience. In this case it is appropriate to separate the education and experience coefficients into different groups.

For an optimality theory we define weighted MSE with respect to the weight matrix  $\mathbf{W} = \text{diag}(\mathbf{V}_1^{-1}, \dots, \mathbf{V}_G^{-1})$ .

**Theorem 24.20** Under the assumptions of Theorem 24.14, if WMSE is defined with respect to  $\mathbf{W} = \text{diag}(\mathbf{V}_1^{-1}, \dots, \mathbf{V}_G^{-1})$  and  $K_g > 2$  for all  $g = 1, \dots, G$  then

$$\text{wmse}(\tilde{\boldsymbol{\theta}}) < \text{wmse}(\hat{\boldsymbol{\theta}}).$$

The proof is a simple extension of the classical James-Stein theory. The block diagonal structure of  $\mathbf{W}$  means that the WMSE is the sum of the WMSE of each group. The classical James-Stein theory can be applied to each group, finding that the WMSE is reduced by shrinkage group-by-group. Thus the total WMSE is reduced by shrinkage.

## 24.25 Empirical Illustrations

We illustrate James-Stein shrinkage with three empirical applications.

The first application is to the same sample as used in Section 24.17, the CPS dataset with the subsample of Asian women ( $n = 1149$ ) focusing on the return to experience profile. We consider shrinkage of Model 9 ( $6^{th}$  order polynomial in experience) towards Model 3 ( $2^{nd}$  order polynomial in experience). The difference in the number of estimated coefficients is 4. We set  $\hat{\mathbf{V}}$  to equal the heteroskedasticity-consistent-consistent covariance matrix estimator. The shrinkage weight is 0.46, meaning that the Stein Rule estimator is approximately an equal weighted average of the estimates from the two models. The estimated experience profiles are displayed in Figure 24.7.

The two least squares estimates are visually distinct. The  $6^{th}$  order polynomial (Model 9) shows a steep return to experience for the first 10 years, then a wobbly experience profile up to 40 years, and declining above that. It also shows a dip around 25 years. The quadratic specification misses some of these features. The Stein Rule estimator is essentially an average of the two profiles. It retains most features of the quartic specification, except that it smooths out the unappealing 25-year dip.

The second application uses the CPS dataset with the subsample of Black men ( $n = 2413$ ) focusing on the return to education across U.S. regions (Northeast, Midwest, South, West). Suppose you are asked to flexibly estimate the return to education for Black men, allowing for the return to education to vary across the regions. Given the model selection information from Section 24.17, a natural baseline is model 6, augmented to allow for greater variation across regions. A flexible specification interacts the six education dummy variables with the four regional dummies (omitting the intercept), which adds 18 coefficients and allows the return to education to vary without restriction in each region.

The least squares estimate of the return to education by region is displayed in panel (a) of Figure 24.8. For simplicity we label the omitted education group (less than 12 years education) as “11 years”. The estimates appear noisy due to the small samples. One feature which we can see is that the four lines track one another for years of education between 12 and 18. That is, they are roughly linear in years of education with the same slope but different intercepts.

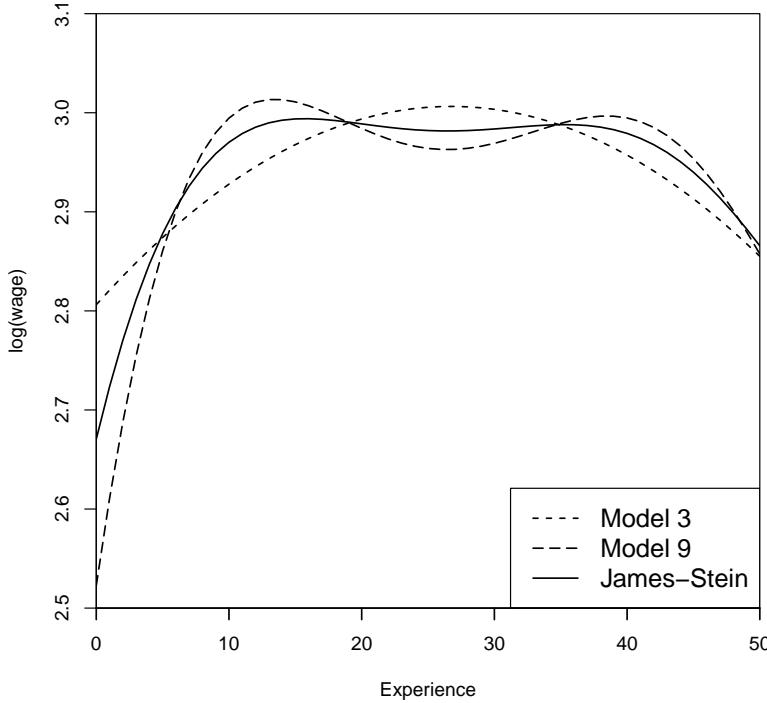


Figure 24.7: Stein Rule Estimation of Experience Profile

To improve the precision of the estimates we shrink the four profiles towards Model 6. This means that we are shrinking the profiles not towards each other, but towards the model with the same effect of education but regional-specific intercepts. Again we use the heteroskedastic covariance matrix estimate. The number of restrictions is 18. The shrinkage weight is 0.49, which means that the Stein Rule estimator puts equal weight on the two models.

The Stein Rule estimates are displayed in panel (b) of Figure 24.8. The estimates are less noisy than panel (a) and it is easier to see the patterns. The four lines track each other, and are approximately linear over 12-18. For 20 years of education the four lines disperse, which seems likely due to small samples. In panel (b) it is easier to see the patterns across regions. It appears that the northeast region has the highest wages (conditional on education) while the west region has the lowest wages. This ranking is constant for nearly all levels of education.

While the Stein Rule estimates shrink the nonparametric estimates towards the common-education-factor specification, it does not impose the latter specification. The Stein Rule estimator has the ability to put near zero weight on the common-factor model. The fact that the estimates put 1/2 weight on both models is the choice selected by the Stein Rule and is thus data-driven.

The third application is to the Invest1993 data set used throughout Chapter 17. This is a panel data set of annual observations on investment decisions by corporations. We focus on the firm-specific effects. These are of interest when studying firm heterogeneity, and is of particular importance for firm-specific forecasting. Accurate estimation of firm effects is challenging when the number of time series observations per firm is small.

To keep the analysis focused we restrict attention to firms which are traded on either the NYSE or AMEX, and to the last ten years of the sample (1982-1991). Since the regressors are lagged this means that there are at most nine time-series observations per firm. The sample has a total of  $N = 786$  firms and  $n = 5692$  observations for estimation. Our baseline model is the two-way fixed effects linear regression as reported in the fourth column of Table 17.2. Our restricted model replaces the firm fixed effects with 19 industry-specific dummy variables. This is similar to the first column of Table 17.2, except that the

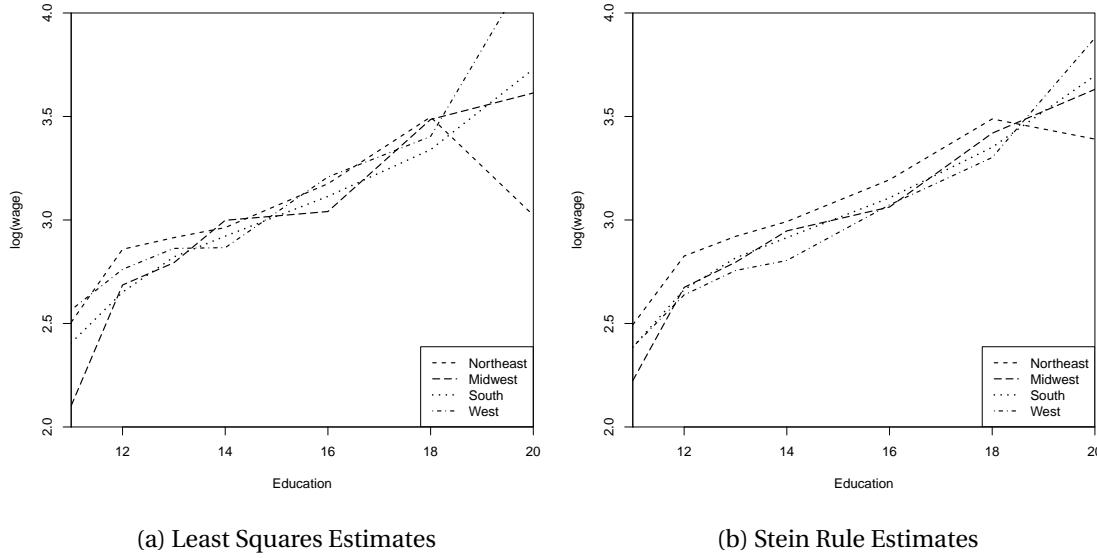


Figure 24.8: Stein Rule Estimation of Education Profiles Across Regions

trading dummy is omitted and time dummies are added. The Stein Rule estimator thus shrinks the fixed effects model towards the industry effects model. The latter will do well if most of the fixed effects are explained by industry rather than firm-specific variation.

Due to the large number of estimated coefficients in the unrestricted model we use the homoskedastic weight matrix as a simplification. This allows the calculation of the shrinkage weight using the simple formula (24.34) for the statistic  $J$ . The heteroskedastic covariance matrix is not appropriate, and the cluster-robust covariance matrix will not be reliable due to the sparse dummy specification.

The estimated shrinkage weight is 0.35, which means that the Stein Rule estimator puts about 1/3 weight on the industry-effect specification and 2/3 weight on the firm-specific specification.

To report our results we focus on the distribution of the firm-specific effects. For the fixed effects model these are the estimated fixed effects. For the industry-effect model these are the estimated industry dummy coefficients (for each firm). For the Stein Rule estimates they are a weighted average of the two. We estimate<sup>5</sup> the densities of the estimated firm-specific effects from the fixed-effects and Stein Rule estimators, and plot them in Figure 24.9.

You can see that the fixed-effects estimate of the firm-specific density is more dispersed, while the Stein estimator is sharper and more peaked, indicating that the fixed effects estimator attributes more variation in firm-specific factors than the Stein estimator. The Stein estimator pulls the fixed effects towards their common mean, adjusting for the randomness due to their estimation. Our expectation is that the Stein estimates, if used for an application such as firm-specific forecasting, will be more accurate because they will have reduced variance relative to the fixed effects estimates.

The message from these three applications is that the James-Stein shrinkage approach can be constructively used to reduce estimation variance in economic applications. These applications illustrate common forms of potential applications: Shrinkage of a flexible specification towards a simpler specification; Shrinkage of heterogeneous estimates towards homogeneous estimates; Shrinkage of fixed effects towards group dummy estimates. These three applications also employed moderately large sample sizes ( $n = 1149, 2413$ , and  $5692$ ) yet found shrinkage weights near 50%. This shows that the benefits of Stein shrinkage are not confined to “small” samples, but rather can be constructively used in moderately large samples with complicated structures.

<sup>5</sup>The two densities are estimated with a common bandwidth to aid comparison. The bandwidth was selected to compromise between those selected for the two samples. The Gaussian kernel was used.

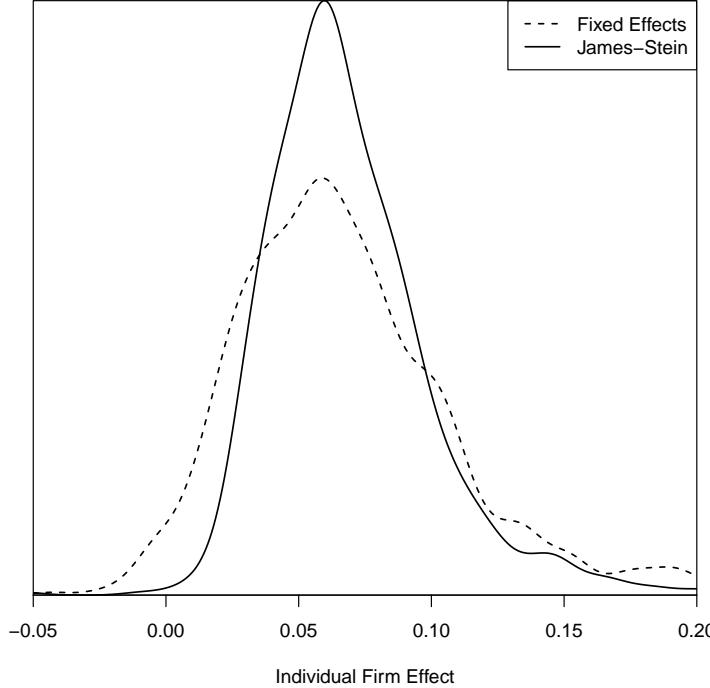


Figure 24.9: Stein Rule Estimation of Firm Specific Effects

## 24.26 Model Averaging

Recall that the problem of model selection is how to select a single model from a general set of models. The James-Stein shrinkage estimator smooths between two nested models by taking a weighted average of two estimators. More generally we can take an average of an arbitrary number of estimators. These estimators are known as model averaging estimators. The key issue for estimation is how to select the averaging weights.

Suppose we have a set of  $M$  models  $\bar{\mathcal{M}} = \{\mathcal{M}_1, \dots, \mathcal{M}_M\}$ . For each model there is an estimator  $\hat{\boldsymbol{\theta}}_m$  of the parameter  $\boldsymbol{\theta}$ . The natural way to think about multiple models, parameters, and estimators is the same as for model selection. All models are subsets of a general superset (overlapping) model which contains all submodels as special cases.

Corresponding to the set of models we introduce a set of weights  $\mathbf{w} = \{w_1, \dots, w_M\}$ . It is common to restrict the weights to be non-negative and sum to one. The set of such weights is called the  $\mathbb{R}^M$  probability simplex.

**Definition 24.4 Probability Simplex.** The set  $\mathcal{S} \subset \mathbb{R}^M$  of vectors such that  $\sum_{m=1}^M w_m = 1$  and  $w_i \geq 1$  for  $i = 1, \dots, M$ .

The probability simplex in  $\mathbb{R}^2$  and  $\mathbb{R}^3$  is shown in the two panels of Figure 24.10. The simplex in  $\mathbb{R}^2$  (the left panel) is the line between the vertices  $(1, 0)$  and  $(0, 1)$ . An example element is the point  $(.7, .3)$  indicated by the dot. This is the weight vector which puts weight 0.7 on model 1 and weight 0.3 on model 2. The vertex  $(1, 0)$  is the weight vector which puts all weight on model 1, corresponding to model selection, and similarly the vertex  $(0, 1)$  is the weight vector which puts all weight on model 2.

The simplex in  $\mathbb{R}^3$  (the right panel) is the equilateral triangle formed between  $(1,0,0)$ ,  $(0,1,0)$ , and  $(0,0,1)$ . An example element is the point  $(.1,.5,.4)$  indicated by the dot. The edges are weight vectors which are averages between two of the three models. For example the bottom edge are weight vectors which divide the weight between models 1 and 2, placing no weight on model 3. The vertices are weight vectors which put all weight on one of the three models and correspond to model selection.

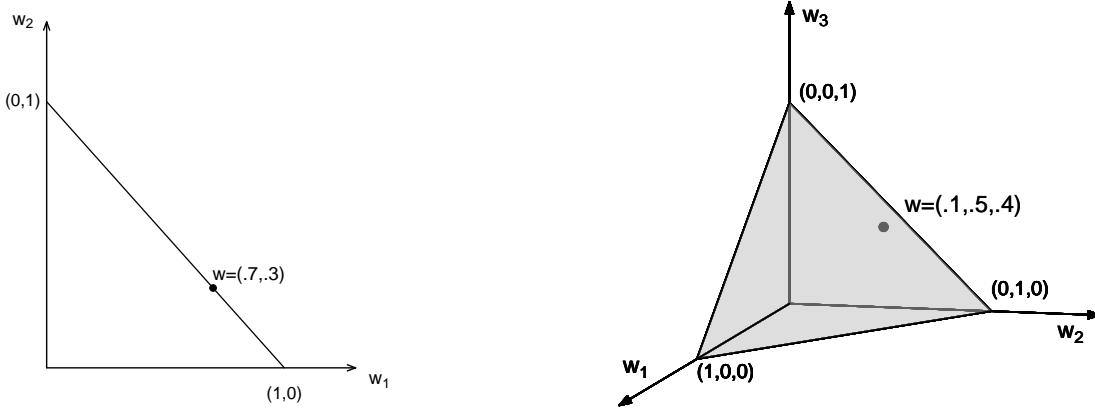


Figure 24.10: Probability Simplex in  $\mathbb{R}^2$  and  $\mathbb{R}^3$

Since the weights on the probability simplex sum to one, an alternative representation is to eliminate one weight by substitution. Thus we can set  $w_M = 1 - \sum_{m=1}^{M-1} w_m$  and define the set of vectors  $\mathbf{w} = \{w_1, \dots, w_{M-1}\}$  which lie in the  $\mathbb{R}^{M-1}$  unit simplex, which is the region bracketed by the probability simplex and the origin.

Given a weight vector we define the averaging estimator

$$\hat{\boldsymbol{\theta}}(\mathbf{w}) = \sum_{m=1}^M w_m \hat{\boldsymbol{\theta}}_m. \quad (24.35)$$

Selection estimators emerge as the special case where the weight vector  $\mathbf{w}$  is a unit vector, e.g. the vertices in Figure 24.10.

It is not absolutely necessary to restrict the weight vector of an averaging estimator to lie in the probability simplex  $\mathcal{S}$ , but in most cases it is a sensible restriction which improves performance. The unadjusted James-Stein estimator, for example, is an averaging estimator which does not enforce non-negativity of the weights. The positive-part version, however, imposes non-negativity and achieves reduced MSE as a result.

In Section 24.18 and Theorem 24.13 we explored the MSE of a simple shrinkage estimator which shrinks an unrestricted estimator towards the zero vector. This is the same as a model averaging estimator where one of the two estimators is the zero vector. In Theorem 24.13 we showed that the MSE of the optimal shrinkage (model averaging) estimator is less than the unrestricted estimator. This result extends to the case of averaging between an arbitrary number of estimators. The MSE of the optimal averaging estimator is less than the MSE of the estimator of the full model, in any given sample.

The optimal averaging weights, however, are unknown. A number of methods have been proposed for selection of the averaging weights.

One simple method is **equal weighting**. This is achieved by setting  $w_m = 1/M$  and results in the

estimator

$$\hat{\boldsymbol{\theta}}^* = \frac{1}{M} \sum_{m=1}^M \hat{\boldsymbol{\theta}}_m.$$

The advantages of equal weighting are that it is simple, easy to motivate, and no randomness is introduced by estimation of the weights. The variance of the equal weighting estimator can be calculated since the weights are fixed. Another important advantage is that the estimator can be constructed in contexts where it is unknown how to construct empirical-based weights, for example when averaging models from completely different probability families. The disadvantages of equal weighting are that the method can be sensitive to the set of models considered, there is no guarantee that the estimator will perform better than the unrestricted estimator, and sample information is inefficiently used. In practice, equal weighting is best used in contexts where the set of models have been pre-screened so that all are considered “reasonable” models. From the standpoint of econometric methodology, equal weighting is not a proper statistical method, as it is an incomplete methodology.

Despite these concerns, equal weighting can be constructively employed when summarizing information for a non-technical audience. The relevant context is when you have a small number of reasonable but distinct estimates, typically made using different assumptions. The distinct estimates are presented to illustrate the range of possible results, and the average taken to represent the “consensus” or “recommended” estimate.

As mentioned above, a number of methods have been proposed for selection of the averaging weights. In the following sections we outline four popular methods: Smoothed BIC, Smoothed AIC, Mallows averaging, and Jackknife averaging.

## 24.27 Smoothed BIC and AIC

Recall that Schwarz's Theorem 24.2 states that for a probability model  $f(\mathbf{y}, \boldsymbol{\theta})$  and a diffuse prior, then the marginal likelihood  $p(\mathbf{y})$  satisfies

$$-2 \log p(\mathbf{y}) \simeq -2 \log L(\hat{\boldsymbol{\theta}}) + K \log(n) = \text{BIC}.$$

This has been interpreted to mean that the model with the highest value of the right-hand-side approximately has the highest marginal likelihood, and is thus the model with the highest probability of being the true model.

There is another interpretation of Schwarz's result. We can write the approximation as

$$p(\mathbf{y}) \simeq \exp(-\text{BIC}/2).$$

This shows that the marginal likelihood is approximately proportional to the probability that the model is true, conditional on the data. Thus we can set the model weight to be proportional to the right-hand-side. These are known as BIC weights and produce the smoothed BIC estimator.

To describe the method completely, we have a set of models  $\bar{\mathcal{M}} = \{\mathcal{M}_1, \dots, \mathcal{M}_M\}$ . Each model  $f_m(\mathbf{y}, \boldsymbol{\theta}_m)$  depends on a  $K_m \times 1$  parameter vector  $\boldsymbol{\theta}_m$  which is estimated by the maximum likelihood estimator  $\hat{\boldsymbol{\theta}}_m$ . The maximized likelihood is  $L_m(\hat{\boldsymbol{\theta}}_m) = f_m(\mathbf{y}, \hat{\boldsymbol{\theta}}_m)$ . The BIC for model  $m$  is  $\text{BIC}_m = -2 \log L_m(\hat{\boldsymbol{\theta}}_m) + K_m \log(n)$ .

The BIC weights are

$$w_m = \frac{\exp(-\text{BIC}_m/2)}{\sum_{j=1}^M \exp(-\text{BIC}_j/2)}.$$

Some properties of the BIC weights are as follows. They are non-negative, so all models receive positive weight. However models can receive weight arbitrarily close to zero, and in practice it is common that most estimated models receive BIC weight that is essentially zero. The model which is selected by BIC receives the greatest weight, and models which have BIC values close to the minimum receive weights closest to the largest weight. Models whose BIC is not close to the minimum receive weight near zero.

The Smoothed BIC (SBIC) estimator is

$$\hat{\boldsymbol{\theta}}_{\text{sbic}} = \sum_{m=1}^M w_m \hat{\boldsymbol{\theta}}_m.$$

The SBIC estimator is a smoother function of the data than BIC selection as there are no discontinuous jumps across models.

An advantage of the smoothed BIC weights and estimator is that it can be used to combine models from different probability families. As for the BIC, it is important that all models are estimated on the same sample. It is also important that the full formula is used for the BIC (no omission of constants) when combining models from different probability families.

Computationally it is better to implement smoothed BIC with what are called “BIC differences” rather than the actual values of the BIC, as the formula as written can produce numerical overflow problems. The difficulty is due to the exponentiation in the formula. This problem can be eliminated as follows. Let

$$\text{BIC}^* = \min_{1 \leq m \leq M} \text{BIC}_m$$

denote the lowest BIC among the models and define the BIC differences

$$\Delta\text{BIC}_m = \text{BIC}_m - \text{BIC}^*.$$

Then

$$\begin{aligned} w_m &= \frac{\exp(-\text{BIC}_m/2)}{\sum_{j=1}^M \exp(-\text{BIC}_j/2)} \\ &= \frac{\exp(-\text{BIC}_m/2) \exp(\text{BIC}^*/2)}{\sum_{j=1}^M \exp(-\text{BIC}_j/2) \exp(\text{BIC}^*/2)} \\ &= \frac{\exp(-\Delta\text{BIC}_m/2)}{\sum_{j=1}^M \exp(-\Delta\text{BIC}_j/2)}. \end{aligned}$$

Thus the weights are algebraically identically whether computed on  $\text{BIC}_m$  or  $\Delta\text{BIC}_m$ . Since  $\Delta\text{BIC}_m$  are of smaller magnitude than  $\text{BIC}_m$  overflow problems are less likely to occur.

Because of the properties of the exponential, if  $\Delta\text{BIC}_m \geq 10$  then  $w_m \leq 0.01$ . Thus smoothed BIC typically concentrates weight on models whose BIC values are close to the minimum. This means that in practice smoothed BIC puts effective non-zero weight on a small number of models.

Burnham and Anderson (1998) follow a suggestion they credit to Akaike that if we make the same transformation to the AIC as to the BIC to obtain the smoothed BIC weights, we obtain frequentist approximate probabilities for the models. Specifically they propose the weights

$$w_m = \frac{\exp(-\text{AIC}_m/2)}{\sum_{j=1}^M \exp(-\text{AIC}_j/2)}.$$

They do not provide a strong theoretical justification for this specific choice of transformation, but it seems natural given the smoothed BIC formula and does work well in simulations.

The algebraic properties of the AIC weights are similar to those of the BIC weights. All models receive positive weight though some receive weight which is arbitrarily close to zero. The model with the smallest AIC receives the greatest AIC weight, and models with similar AIC values receive similar AIC weights.

Computationally the AIC weights should be computed using AIC differences. Define

$$\begin{aligned} \text{AIC}^* &= \min_{1 \leq m \leq M} \text{AIC}_m \\ \Delta\text{AIC}_m &= \text{AIC}_m - \text{AIC}^*. \end{aligned}$$

We can calculate that the AIC weights algebraically equal

$$w_m = \frac{\exp(-\Delta \text{AIC}_m \text{AIC}_m / 2)}{\sum_{j=1}^M \exp(-\Delta \text{AIC}_j / 2)}.$$

As for the BIC weights,  $w_m \leq 0.01$  if  $\text{AIC}_m \geq 10$  so the AIC weights will concentrated on models whose AIC values are close to the minimum. However, in practice it is common that the AIC criterion is less concentrated than the BIC criterion, as the AIC puts a smaller penalty on large penalizations, so the AIC weights tend to be more spread out across models than the corresponding BIC weights.

The Smoothed AIC (SAIC) estimator is

$$\hat{\boldsymbol{\theta}}_{\text{saic}} = \sum_{m=1}^M w_m \hat{\boldsymbol{\theta}}_m.$$

The SAIC estimator is a smoother function of the data than AIC selection.

Recall that both AIC selection and BIC selection are model selection consistent, in the sense that as the sample size gets large the probability that the selected model is a true model is arbitrarily close to one. Furthermore BIC is consistent for parsimonious models, and AIC asymptotically over-selects.

These properties extend to SBIC and SAIC. In large samples, SAIC and SBIC weights will concentrate exclusively on true models; the weight on incorrect models will asymptotically approach zero. However, SAIC will asymptotically spread weight across both parsimonious true models and overparameterized true models, which SBIC asymptotically will concentrate weight only on parsimonious true models.

An interesting property of the smoothed estimators is the possibility of asymptotically spreading weight across equal-fitting parsimonious models. Suppose we have two non-nested models with the same number of parameters with the same KLIC value so they are equally good approximations. In large samples both SBIC and SAIC will be weighted averages of the two estimators, rather than simply selecting one of the two.

## 24.28 Mallows Model Averaging

In linear regression the Mallows criterion (24.17) applies directly to the model averaging estimator (24.35). The homoskedastic regression model is

$$\begin{aligned} y_i &= m_i + e_i \\ m_i &= m(\mathbf{x}_i) \\ \mathbb{E}(e_i | \mathbf{x}_i) &= 0 \\ \mathbb{E}(e_i^2 | \mathbf{x}_i) &= \sigma^2. \end{aligned}$$

Suppose that there are  $M$  models for  $m(\mathbf{x}_i)$ , each which takes the form  $\boldsymbol{\beta}'_m \mathbf{x}_{mi}$  for some  $K_m \times 1$  regression vector  $\mathbf{x}_{mi}$ . The  $m^{th}$  model estimator of the coefficient is  $\hat{\boldsymbol{\beta}}_m = (\mathbf{X}'_m \mathbf{X}_m)^{-1} \mathbf{X}'_m \mathbf{y}$ , and the estimator of the vector  $\mathbf{m}$  is  $\hat{\mathbf{m}}_m = \mathbf{P}_m \mathbf{y}$  where  $\mathbf{P}_m = \mathbf{X}_m (\mathbf{X}'_m \mathbf{X}_m)^{-1} \mathbf{X}'_m$ . The corresponding residual vector is  $\hat{\mathbf{e}}_m = (\mathbf{I}_n - \mathbf{P}_m) \mathbf{y}$ .

The model averaging estimator for fixed weights is

$$\hat{\mathbf{m}}_m(\mathbf{w}) = \sum_{m=1}^M w_m \mathbf{P}_m \mathbf{y} = \mathbf{P}(\mathbf{w}) \mathbf{y}$$

where

$$\mathbf{P}(\mathbf{w}) = \sum_{m=1}^M w_m \mathbf{P}_m.$$

The model averaging residual is

$$\hat{\mathbf{e}}(\mathbf{w}) = (\mathbf{I}_n - \mathbf{P}(\mathbf{w})) \mathbf{y} = \sum_{m=1}^M w_m (\mathbf{I}_n - \mathbf{P}_m) \mathbf{y}.$$

The estimator  $\hat{\mathbf{m}}_m(\mathbf{w})$  is linear in  $\mathbf{y}$  so the Mallows criterion can be applied. It equals

$$\begin{aligned} C(\mathbf{w}) &= \hat{\mathbf{e}}(\mathbf{w})' \hat{\mathbf{e}}(\mathbf{w}) + 2\tilde{\sigma}^2 \text{tr}(\mathbf{P}(\mathbf{w})) \\ &= \hat{\mathbf{e}}(\mathbf{w})' \hat{\mathbf{e}}(\mathbf{w}) + 2\tilde{\sigma}^2 \sum_{m=1}^M w_m K_m \end{aligned}$$

where  $\tilde{\sigma}^2$  is a preliminary<sup>6</sup> estimator of  $\sigma^2$ .

In the case of model selection the Mallows penalty is proportional to the number of estimated coefficients. In the model averaging case the Mallows penalty is the average number of estimated coefficients.

The Mallows-selected weight vector is that which minimizes the Mallows criterion. It can be written as

$$\hat{\mathbf{w}}_{\text{mma}} = \underset{\mathbf{w} \in \mathcal{S}}{\operatorname{argmin}} C(\mathbf{w}). \quad (24.36)$$

Computationally it is useful to observe that  $C(\mathbf{w})$  is a quadratic function in  $\mathbf{w}$ . Indeed, by defining the  $n \times M$  matrix  $\hat{\mathbf{E}} = [\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_M]$  of residual vectors, and the  $M \times 1$  vector  $\mathbf{K} = [K_1, \dots, K_M]$ , the criterion is

$$C(\mathbf{w}) = \mathbf{w}' \hat{\mathbf{E}}' \hat{\mathbf{E}} \mathbf{w} + 2\tilde{\sigma}^2 \mathbf{K}' \mathbf{w}.$$

The probability simplex  $\mathcal{S}$  is defined by one equality and  $2M$  inequality constraints. The minimization problem (24.36) falls in the category of **quadratic programming**, which means optimization of a quadratic subject to linear equality and inequality constraints. This is a well-studied area of numerical optimization and numerical solutions are widely available. In R use the command `solve.QP` in the package `quadprog`. In MATLAB use the command `quadprog`.

Figure 24.11 illustrates the Mallows weight computation problem. Displayed is the probability simplex  $\mathcal{S}$  in  $\mathbb{R}^3$ . The axes are the weight vectors. The ellipses are the contours of the unconstrained sum of squared errors as a function of the weight vectors projected onto the constrained set  $\sum_{m=1}^M w_m = 1$ . This is the extension of the probability simplex as a two-dimensional plane in  $\mathbb{R}^3$ . The midpoint of the contours is the minimizing weight vector allowing for weights outside  $[0, 1]$ . The point where the lowest contour ellipse hits the probability simplex is the solution (24.36), the Mallows selected weight vector. In the left panel is displayed an example where the solution is the vertex  $(0, 1, 0)$  so the selected weight vector puts all weight on model 2. In the right panel is displayed an example where the solution lies on the edge between  $(1, 0, 0)$  and  $(0, 0, 1)$ , meaning that the selected weight vector averages models 1 and 3, but puts no weight on model 2. Since the contour sets are ellipses and the constraint set is a simplex, solution points tend to be on edges and vertices, meaning that some models receive zero weight. In fact, where there are a large number of models a generic feature of the solution is that most models receive zero weight; the selected weight vector puts positive weight on a small subset of the eligible models.

Once the weights  $\hat{\mathbf{w}}$  are obtained the model averaging estimator of the coefficients are found by averaging the model estimates  $\hat{\beta}_m$  using the weights.

In the special case of two nested models the Mallows criterion can be written as

$$\begin{aligned} C(w) &= (w, 1-w) \begin{pmatrix} \hat{\mathbf{e}}_1' \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_1' \hat{\mathbf{e}}_2 \\ \hat{\mathbf{e}}_2' \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2' \hat{\mathbf{e}}_2 \end{pmatrix} \begin{pmatrix} w \\ 1-w \end{pmatrix} + 2\tilde{\sigma}^2 (wk_1 + (1-w)k_2) \\ &= (w, 1-w) \begin{pmatrix} \hat{\mathbf{e}}_1' \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2' \hat{\mathbf{e}}_2 \\ \hat{\mathbf{e}}_2' \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2' \hat{\mathbf{e}}_2 \end{pmatrix} \begin{pmatrix} 1-w \\ w \end{pmatrix} + 2\tilde{\sigma}^2 (wk_1 + (1-w)k_2) \\ &= w^2 (\hat{\mathbf{e}}_1' \hat{\mathbf{e}}_1 - \hat{\mathbf{e}}_2' \hat{\mathbf{e}}_2) + \hat{\mathbf{e}}_2' \hat{\mathbf{e}}_2 - 2\tilde{\sigma}^2 (k_2 - k_1) w + 2\tilde{\sigma}^2 \end{aligned}$$

where we assume  $k_1 < k_2$  so that  $\hat{\mathbf{e}}_1' \hat{\mathbf{e}}_2 = \mathbf{y}' (\mathbf{I}_n - \mathbf{P}_1) (\mathbf{I}_n - \mathbf{P}_2) \mathbf{y} = \mathbf{y}' (\mathbf{I}_n - \mathbf{P}_2) \mathbf{y} = \hat{\mathbf{e}}_2' \hat{\mathbf{e}}_2$ . The minimizer of this criterion is

$$\hat{w} = \left( \frac{\tilde{\sigma}^2 (k_2 - k_1)}{\hat{\mathbf{e}}_1' \hat{\mathbf{e}}_1 - \hat{\mathbf{e}}_2' \hat{\mathbf{e}}_2} \right)_1.$$

---

<sup>6</sup>It is typical to use the bias-corrected least squares variance estimator from the largest model.

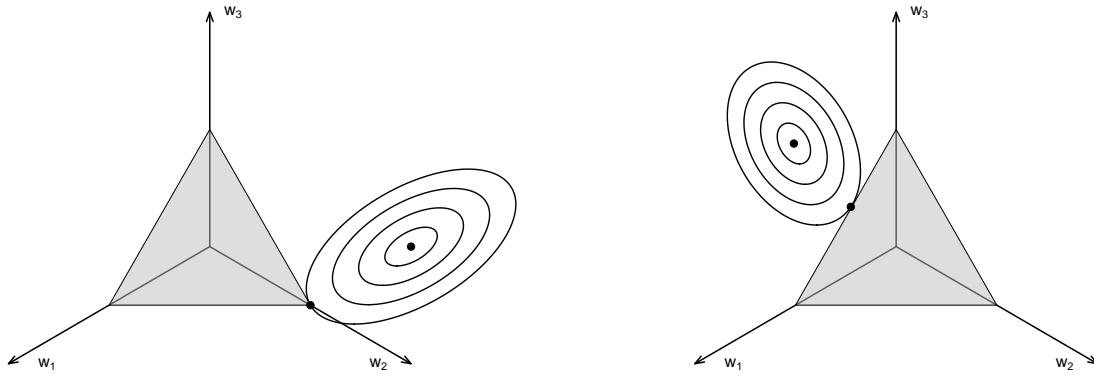


Figure 24.11: Mallows Weight Selection

This is the same as the Stein Rule weight (24.33) with a slightly different shrinkage constant. Thus the Mallows averaging estimator for  $M = 2$  is a member of the Stein Rule family. Hence for  $M > 2$  the Mallows averaging estimator is a generalization of the James-Stein estimator to multiple models.

Based on the latter observation, Hansen (2014) shows that the MMA estimator has lower WMSE than the unrestricted least squares estimator when the models are nested linear regressions, the errors are homoskedastic, and the models are separated by 4 coefficients or greater. The latter condition is analogous to the conditions for improvements in the Stein Rule theory.

Hansen (2007) showed that the MMA estimator asymptotically achieves the same MSE as the infeasible optimal best weighted average using the theory of Li (1987) under similar conditions. This shows that using model selection tools to select the averaging weights is asymptotically optimal for regression fitting and point forecasting.

## 24.29 Jackknife (CV) Model Averaging

A disadvantage of Mallows selection is that the criterion is valid only when the errors are conditional homoskedastic. Selection by cross-validation does not require homoskedasticity. Therefore it seems sensible to use cross-validation rather than Mallows to select the weight vectors. It turns out that this is a simple extension with excellent finite sample performance.

A fitted averaging regression (with fixed weights) can be written as

$$y_i = \sum_{m=1}^M w_m \mathbf{x}'_{mi} \hat{\boldsymbol{\beta}}_m + \hat{e}_i(\mathbf{w})$$

where  $\hat{\boldsymbol{\beta}}_m$  are the least squares coefficient estimates from Model  $m$ . The corresponding leave-one-out equation is

$$y_i = \sum_{m=1}^M w_m \mathbf{x}'_{mi} \hat{\boldsymbol{\beta}}_{m,(-i)} + \tilde{e}_i(\mathbf{w})$$

where  $\hat{\boldsymbol{\beta}}_{m,(-i)}$  are the least squares coefficient estimates from Model  $m$  when observation  $i$  is deleted.

The leave-one-out prediction errors satisfy the simple relationship

$$\tilde{e}_i(\mathbf{w}) = \sum_{m=1}^M w_m \tilde{e}_{mi}$$

where  $\tilde{e}_{mi}$  are the leave-one-out prediction errors for model  $m$ . In matrix notation  $\tilde{\mathbf{e}}(\mathbf{w}) = \tilde{\mathbf{E}}\mathbf{w}$  where  $\tilde{\mathbf{E}}$  is the  $n \times M$  matrix of leave-one-out prediction errors.

This means that the jackknife estimate of variance (or equivalently the cross-validation criterion) equals

$$\text{CV}(\mathbf{w}) = \mathbf{w}' \tilde{\mathbf{E}}' \tilde{\mathbf{E}} \mathbf{w}$$

which is a quadratic function of the weight vector. The cross-validation choice for weight vector is the minimizer

$$\hat{\mathbf{w}}_{\text{jma}} = \underset{\mathbf{w} \in \mathcal{S}}{\operatorname{argmin}} \text{CV}(\mathbf{w}). \quad (24.37)$$

Given the weights the coefficient estimates (and any other parameter of interest) are found by taking weighted averages of the model estimates using the weight vector  $\hat{\mathbf{w}}_{\text{jma}}$ . Hansen and Racine (2012) call this the **Jackknife Model Averaging (JMA)** estimator.

The algebraic properties of the solution are similar to Mallows. Since (24.37) minimizes a quadratic function subject to a simplex constraint, solutions tend to be on edges and vertices, which means that many (or most) models receive zero weight. Hence JMA weight selection simultaneously performs selection and shrinkage. The solution is found numerically by quadratic programming, which is computationally simple and fast even when the number of models  $M$  is large.

Hansen and Racine (2012) showed that the JMA estimator is asymptotically equivalent to the infeasible optimal weighted average across least squares estimates, based on a regression fit criteria. Their results hold under quite mild conditions, including allowing for conditional heteroskedasticity. This result is similar to Andrews (1991c) generalization of Li (1987)'s result for model selection.

The implication of this theory is that JMA weight selection is computationally simple and has excellent sampling performance.

## 24.30 Empirical Illustration

We illustrate the model averaging methods with the empirical application from Section 24.17, which reported wage regression estimates for the CPS sub-sample of Asian women, focusing on the return to experience between 0 and 30 years.

Table 24.2 reports the model averaging weights obtained using the methods of SBIC, SAIC, Mallows model averaging (MMA) and jackknife model averaging (JMA). Also reported in the final column is the weighted average estimate of the return to experience as a percentage.

The results show that the methods put weight on somewhat different models, and different degrees of dispersion. The SBIC puts nearly all weight on model 2. The SAIC puts nearly 1/2 of the weight on model 6, with most of the remainder split between models 5 and 9. MMA puts nearly 1/2 of the weight on model 9, 30% on 5 and 9% on model 1. JMA is similar to MMA but more emphasis on parsimony, with 1/2 of the weight on model 5, 17% on model 9, 17% on model 1, and 8% on model 3. One of the interesting things about the MMA/JMA methods is that they can split weight between quite different models, e.g. models 1 and 9.

The averaging estimators from the non-BIC methods are similar to one another, but SBIC produces a much smaller estimate than the other methods.

## 24.31 Ridge Regression

Ridge regression is a shrinkage-type estimator with similar but distinct properties from the James-Stein estimator. There are two competing motivations for ridge regression. The traditional motivation is

Table 24.2: Model Averaging Weights and Estimates of Return to Experience among Asian Women

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Return
SBIC	.02	.96	.00	.00	.04	.00	.00	.00	.00	22%
SAIC	.00	.02	.10	.00	.15	.44	.00	.06	.22	38%
MMA	.09	.02	.02	.00	.30	.00	.00	.00	.57	39%
JMA	.17	.00	.08	.00	.57	.01	.00	.00	.17	34%

to reduce the degree of collinearity among the regressors. The modern motivation (though in mathematics it predates the “traditional” motivation) is regularization of high-dimensional and ill-posed inverse problems. We discuss both in turn.

Take a linear regression model  $y_i = \mathbf{x}'_i \boldsymbol{\beta} + e_i$ . In nonparametric series and “machine learning” applications the dimension of  $\boldsymbol{\beta}$  can be very large, and often the regressors are highly correlated. In these cases the least squares estimator may be undefined and/or the  $\mathbf{X}'\mathbf{X}$  matrix ill-conditioned, which can mean that the least squares coefficient estimates are numerically unreliable. As a numerical solution to this dilemma, Hoerl and Kennard (1970) proposed the ridge regression estimator

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1} \mathbf{X}'\mathbf{y}$$

where  $\lambda > 0$  is a shrinkage parameter. The ridge regression estimator has the property that it is well-defined and does not suffer from multicollinearity or ill-conditioning so long as  $\lambda > 0$ . This even holds if  $k > n$ ! That is, the ridge regression estimator can be calculated even when the number of regressors exceeds the sample size.

The constant  $\lambda$  is a tuning parameter. We discuss how to select  $\lambda$  below.

To see how  $\lambda > 0$  ensures that the inverse problem is solved, use the spectral decomposition to write  $\mathbf{X}'\mathbf{X} = \mathbf{H}'\mathbf{D}\mathbf{H}$  where  $\mathbf{H}$  is orthonormal and  $\mathbf{D} = \text{diag}\{r_1, \dots, r_k\}$  is a diagonal matrix with the eigenvalues  $r_j$  of  $\mathbf{X}'\mathbf{X}$  on the diagonal. Let  $\boldsymbol{\Lambda} = \lambda \mathbf{I}_k$ . We can write

$$\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k = \mathbf{H}'\boldsymbol{\Lambda}\mathbf{H} + \lambda \mathbf{H}'\mathbf{H} = \mathbf{H}'\boldsymbol{\Lambda}\mathbf{H} + \mathbf{H}'\boldsymbol{\Lambda}\mathbf{H} = \mathbf{H}'(\mathbf{D} + \boldsymbol{\Lambda})\mathbf{H}$$

which has eigenvalues  $r_j + \lambda > 0$ . Thus all eigenvalues are bounded away from zero so  $\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k$  is full rank and well conditioned.

The second motivation is based on penalization. When  $\mathbf{X}'\mathbf{X}$  is ill-conditioned computing its inverse is “ill-posed”. Techniques to deal with ill-posed estimators are called “regularization” and date back to Tikhonov (1943). A leading method is penalization. Consider the penalized regression criterion

$$\begin{aligned} \text{SSE}_2(\boldsymbol{\beta}, \lambda) &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2 \\ &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \end{aligned}$$

where  $\|\mathbf{a}\|_2 = (\mathbf{a}'\mathbf{a})^{1/2}$  is the 2-norm. The minimizer of  $\text{SSE}_2(\boldsymbol{\beta}, \lambda)$  is a regularized least squares estimator.

The first order condition for minimization of  $\text{SSE}_2(\boldsymbol{\beta}, \lambda)$  over  $\boldsymbol{\beta}$  is

$$-2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta} = 0. \quad (24.38)$$

The solution is  $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ . Thus the regularized (penalized) least squares estimator equals ridge regression. This shows that the ridge regression estimator minimizes the sum of squared errors subject to a penalty on the  $L_2$  (2-norm) magnitude of the regression coefficient. Penalizing large coefficient vectors keeps the latter from being too large and erratic. Hence one interpretation of  $\lambda$  is the degree of penalty on the magnitude of the coefficient vector.

Minimization subject to a penalty has a dual representation as constrained minimization. The latter is

$$\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

subject to  $\beta' \beta \leq \tau$  for some  $\tau > 0$ . To see the connection, the Lagrangian for the constrained problem is

$$\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda (\beta' \beta - \tau)$$

where  $\lambda$  is a Lagrange multiplier. The first order condition is (24.38), which is the first order condition for the penalization problem. This shows that they have the same solution.

The practical difference between the penalization and constraint problems is that in the first you specify the ridge parameter  $\lambda$  while in the second you specify the constraint  $\tau$ . They are connected, since the values of  $\lambda$  and  $\tau$  satisfy the relationship

$$\mathbf{y}' \mathbf{X} (\mathbf{X}' \mathbf{X} + \lambda \mathbf{I}_k)^{-1} (\mathbf{X}' \mathbf{X} + \lambda \mathbf{I}_k)^{-1} \mathbf{X}' \mathbf{y} = \tau.$$

Thus to find  $\lambda$  given  $\tau$  it is sufficient to (numerically) solve this equation.

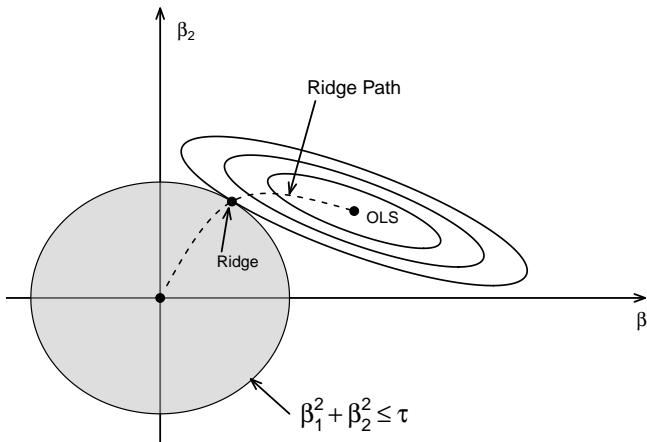


Figure 24.12: Ridge Regression Dual Minimization Solution

To visualize the constraint problem see Figure 24.12 which plots an example in  $\mathbb{R}^2$ . The constraint set  $\beta' \beta \leq \tau$  is displayed as the ball about the origin and the contour sets of the sum of squared errors are displayed as ellipses. The least squares estimator is the center of the ellipses, while the ridge regression estimator is the point on the circle where the contour is tangent. This shrinks the least squares coefficient towards the zero vector. Unlike the Stein estimator, however, it does not shrink along the line segment connecting least squares with the origin, rather it shrinks along a trajectory determined by the degree of correlation between the variables. This trajectory is displayed with the dashed lines, marked as "Ridge path". This is the sequence of ridge regression coefficients obtained as  $\lambda$  (or  $\tau$ ) is varied from small to large. When  $\lambda = 0$  (or  $\tau$  is large) the ridge estimator equals least squares. For small  $\lambda$  the ridge estimator

moves slightly towards the origin by sliding along the ridge of the contour set. As  $\lambda$  increases the ridge estimator takes a more direct path towards the origin. This is unlike the Stein estimator, which shrinks the least squares estimator towards the origin along the connecting line segment.

The ridge parameter  $\lambda$  affects the sampling performance of estimation by decreasing variance and increasing bias. Define the MSE matrix

$$\text{mse}(\hat{\boldsymbol{\beta}}) = \mathbb{E}\left((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\right).$$

**Theorem 24.21** In the homoskedastic linear regression model, if  $0 < \lambda < 2\sigma^2/\boldsymbol{\beta}'\boldsymbol{\beta}$ ,

$$\text{mse}(\hat{\boldsymbol{\beta}}_{\text{ridge}}) < \text{mse}(\hat{\boldsymbol{\beta}}_{\text{ols}}).$$

Theorem 24.21 shows that the ridge estimator dominates the least squares estimator for  $\lambda$  satisfying a range of values. This holds regardless of the dimension of  $\boldsymbol{\beta}$ . Since the upper bound  $2\sigma^2/\boldsymbol{\beta}'\boldsymbol{\beta}$  is unknown, however, it is unclear if feasible ridge regression dominates least squares. The upper bound does not give practical guidance for selection of  $\lambda$ .

It is straightforward to generalize ridge regression to allow different penalties on different groups of regressors. Take the model

$$y_i = \mathbf{x}'_{1i}\boldsymbol{\beta}_1 + \cdots + \mathbf{x}'_{Gi}\boldsymbol{\beta}_G + e_i$$

and minimize the SSE subject to the penalty

$$\lambda_1\boldsymbol{\beta}'_1\boldsymbol{\beta}_1 + \cdots + \lambda_G\boldsymbol{\beta}'_G\boldsymbol{\beta}_G.$$

The solution is

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}'\mathbf{X} + \boldsymbol{\Lambda})^{-1}\mathbf{X}'\mathbf{y}$$

where

$$\boldsymbol{\Lambda} = \text{diag}\{\lambda_1\mathbf{I}_{k_1}, \dots, \lambda_G\mathbf{I}_{k_G}\}.$$

This allows for some coefficients to be penalized more (or less) than other coefficients. This added flexibility comes at the cost of needing to select the shrinkage parameters  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_G)$ . One important special case is where  $\lambda_1 = 0$ , thus one group of coefficients are not penalized. This enables the simple partition of the coefficients into two groups: penalized and non-penalized.

The most popular method to select the shrinkage parameter  $\lambda$  is cross validation. The leave-one-out ridge regression estimator, prediction errors, and CV criterion are

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{-i}(\boldsymbol{\lambda}) &= \left(\sum_{j \neq i} \mathbf{x}_j \mathbf{x}'_j + \boldsymbol{\Lambda}\right)^{-1} \left(\sum_{j \neq i} \mathbf{x}_j y_j\right) \\ \tilde{e}_i(\boldsymbol{\lambda}) &= y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{-i} \\ \text{CV}(\boldsymbol{\lambda}) &= \sum_{i=1}^n \tilde{e}_i(\boldsymbol{\lambda})^2.\end{aligned}$$

The CV-selected shrinkage parameter  $\hat{\boldsymbol{\lambda}}$  minimizes  $\text{CV}(\boldsymbol{\lambda})$ . The cross-validation ridge estimator is calculated using  $\hat{\boldsymbol{\lambda}}$ .

In practice it can be tricky to minimize  $\text{CV}(\boldsymbol{\lambda})$ . The minimum may occur at  $\lambda = 0$  (ridge equals least squares), or as  $\lambda$  tends to infinity (full shrinkage), or have multiple local minima. The scale of the minimizing  $\lambda$  depends on the scaling of the regressors and in particular the singular values of  $\mathbf{X}'\mathbf{X}$ . It can be important to explore  $\text{CV}(\boldsymbol{\lambda})$  for very small values of  $\lambda$ .

As for least squares there is a simple formula to calculate the CV criterion for ridge regression which greatly speeds computation.

**Theorem 24.22** The leave-one-out ridge regression prediction errors are

$$\tilde{e}_i(\boldsymbol{\lambda}) = \left(1 - \mathbf{x}'_i (\mathbf{X}'\mathbf{X} + \boldsymbol{\Lambda})^{-1} \mathbf{x}_i\right)^{-1} \hat{e}_i(\boldsymbol{\lambda})$$

where  $\hat{e}_i(\boldsymbol{\lambda}) = y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{\text{ridge}}(\boldsymbol{\lambda})$  are the ridge regression residuals.

The proof is very similar to that of Theorem 3.7 so is omitted.

An alternative method for selection of  $\lambda$  is to minimize the Mallows criterion, which equals

$$C(\boldsymbol{\lambda}) = \sum_{i=1}^n \hat{e}_i(\boldsymbol{\lambda})^2 + 2\hat{\sigma}^2 \text{tr}\left((\mathbf{X}'\mathbf{X} + \boldsymbol{\Lambda})^{-1} (\mathbf{X}'\mathbf{X})\right).$$

where  $\hat{\sigma}^2$  is the variance estimator from least squares estimation. The Mallows-selected shrinkage parameter  $\hat{\boldsymbol{\lambda}}$  minimizes  $C(\boldsymbol{\lambda})$ . The Mallows-selected ridge estimate is calculated using  $\hat{\boldsymbol{\lambda}}$ . Li (1986) showed that in the normal regression model the Mallows-selected shrinkage estimator is asymptotically equivalent to the infeasible best shrinkage parameter in terms of regression fit. I am unaware of a similar optimality result for cross-validated-selected ridge estimation.

An important caveat is that the ridge regression estimator is not invariant to rescaling the regressors, nor other linear transformations. Therefore it is common to consider applying ridge regression after applying standardizing transformations to the regressors.

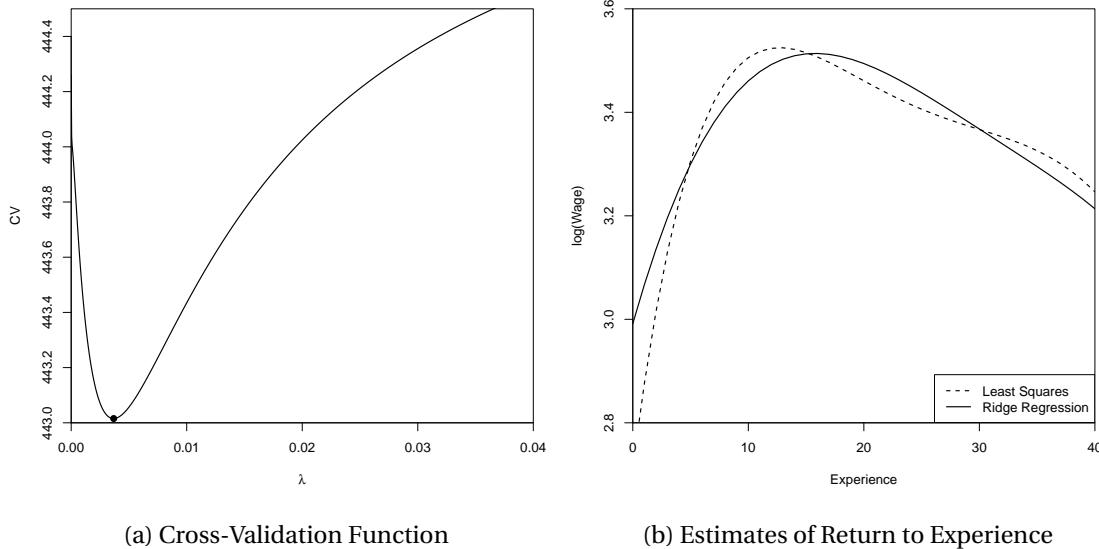


Figure 24.13: Least Squares and Ridge Regression Estimates of the Return to Experience

To illustrate ridge regression we use the CPS dataset with the sample of Asian men with a college education (16 years of education or more) to estimate the experience profile. We standardize experience by dividing by the largest experience level (69) in the sample so that experience lies in the interval  $[0,1]$ . We regress log wages on a fifth-order polynomial in experience. We estimate the same regression by ridge regression, shrinking all coefficients (the five experience coefficients and the intercept) with a common penalty. Panel (a) of Figure 24.13 displays the cross-validation function calculated over the interval  $[0, 0.04]$ . It has an internal minimum at  $\hat{\lambda} = 0.0037$ . Panel (b) displays the estimated experience profiles, least squares displayed by dashes and ridge regression (using the CV-selected shrinkage parameter) by the solid line. The ridge regression estimate is smoother and more compelling. In this example, if we alternatively estimate the model by ridge regression by grouping the regressors, either separating out the

intercept, or separating the intercept and first two polynomial terms, and setting the shrinkage parameter on this first group equal to zero, the estimated experience profile is nearly identical. That is, the estimates obtained by shrinking the full regression towards zero, just the coefficients on experience, or just the coefficients on experience, are all roughly equivalent.

In summary, ridge regression is a very useful shrinkage tool, though tricky to use because of selection of the shrinkage parameter.

## 24.32 LASSO

In the previous section we learned that ridge regression minimizes the sum of squared errors plus an  $L_2$  penalty on the coefficient vector. Model selection (e.g. Mallows) minimizes the sum of squared errors plus an  $L_0$  norm penalty on the coefficient vector (the number of non-zero coefficients). An intermediate case uses an  $L_1$  penalty. This is known as the LASSO (for **Least Absolute Shrinkage and Selection Operator**). The  $L_1$  penalized least squares criterion is

$$\begin{aligned} \text{SSE}_1(\boldsymbol{\beta}, \lambda) &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^k |\beta_j| \\ &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \end{aligned}$$

where  $\|\boldsymbol{\alpha}\|_1 = \sum_{j=1}^k |a_j|$  is the 1-norm ( $L_1$ ). The LASSO estimator is the minimizer of this penalized criterion

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \text{SSE}_1(\boldsymbol{\beta}, \lambda).$$

Except for special cases the solution must be found numerically. Fortunately, computational algorithms are surprisingly simple and fast. An important property is that when  $\lambda > 0$  the LASSO estimator is well-defined even if  $k > n$ .

The LASSO minimization problem has the dual constrained optimization problem

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \underset{\|\boldsymbol{\beta}\|_1 \leq \tau}{\operatorname{argmin}} \text{SSE}(\boldsymbol{\beta}).$$

To see that the two problems are the same observe that the constrained optimization problem has the Lagrangian

$$\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \left( \sum_{j=1}^k |\beta_j| - \tau \right)$$

which has first order conditions

$$-2\mathbf{X}'_j(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \operatorname{sgn}(\beta_j) = 0$$

which are the same as those for minimization of the penalized criterion. Thus the solutions are identical.

The constraint set  $\{\|\boldsymbol{\beta}\|_1 \leq \tau\}$  for the dual problem is a cross-polytope, resembling a multi-faceted diamond. The constrained minimization problem in  $\mathbb{R}^2$  is illustrated in Figure 24.14. The sum of squared error contour sets are the ellipses with the least squares solution at the center. The constraint set is the shaded polytope. The LASSO estimator is the intersection point between the constraint set and the largest ellipse drawn, and in this example hits a vertex of the constraint set, and so the constrained estimator sets  $\hat{\beta}_1 = 0$ . This is a typical outcome in LASSO estimation. Since we are minimizing a quadratic subject to a polytope constraint, the solution tends to be at vertices which eliminate a subset of the coefficients.

The LASSO path is drawn with the dashed line. This is the sequence of solution paths obtained as the constraint set is varied. The solution path has the property that it is a straight line from the least squares estimator to the  $y$ -axis (in this example), at which point  $\beta_2$  is set to zero, and then the solution path follows the  $y$ -axis to the origin. With a general number of coefficients the solution path has a similar property, where the solution path is linear on segments until each coefficient hits zero, at which point it

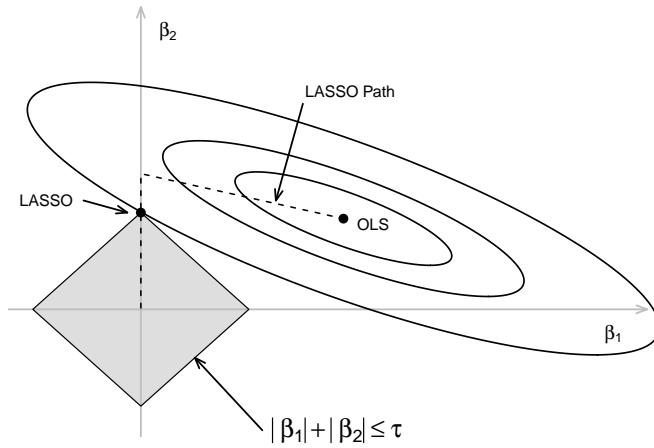


Figure 24.14: LASSO Dual Minimization Solution

is eliminated. In this particular example the solution path shows  $\beta_2$  increasing while  $\beta_1$  decreases. Thus while LASSO is a shrinkage estimator it does not necessarily shrink the individual coefficients monotonically.

It is instructive to compare Figures 24.12 and 24.14 which have the same sum of squares contours. The ridge estimator is generically an interior solution, with no individual coefficient set to zero. The LASSO estimator typically sets some coefficients equal to zero. However both estimators follow similar solution paths, following the ridge of the sum of squared criterion rather than taking a direct path towards the origin.

One case where we can explicitly calculate the LASSO estimates is when the regressors are orthogonal. Suppose that  $\mathbf{X}'\mathbf{X} = \mathbf{I}_k$  and  $k < n$ . Then the first order condition for minimization simplifies to

$$-2(\hat{\beta}_{\text{ols},j} - \hat{\beta}_{\text{lasso},j}) + \lambda \text{sgn}(\hat{\beta}_{\text{lasso},j}) = 0$$

which has the explicit solution

$$\hat{\beta}_{\text{lasso},j} = \begin{cases} \hat{\beta}_{\text{ols},j} - \lambda/2 & \hat{\beta}_{\text{ols},j} > \lambda/2 \\ 0 & |\hat{\beta}_{\text{ols},j}| \leq \lambda/2 \\ \hat{\beta}_{\text{ols},j} + \lambda/2 & \hat{\beta}_{\text{ols},j} < -\lambda/2 \end{cases} .$$

Thus the LASSO coefficient is a continuous transformation of the least squares coefficient estimate. For small values of the least squares estimate the LASSO estimate is set to zero. For all other values the LASSO estimate moves the least squares estimate towards zero by  $\lambda/2$ .

It is constructive to contrast this behavior with ridge regression and selection estimation. When  $\mathbf{X}'\mathbf{X} = \mathbf{I}_k$  the ridge estimator equals  $\hat{\beta}_{\text{ridge}} = (1 + \lambda)^{-1} \hat{\beta}_{\text{ols}}$  so shrinks the coefficients towards zero by a common multiple. A selection estimator (for simplicity consider selection based on a homoskedastic t-test with  $\hat{\sigma}^2 = 1$  and critical value  $c$ ) equals  $\hat{\beta}_{\text{ridge}} = \mathbf{1}(|\hat{\beta}_{\text{ols},j}| > c) \hat{\beta}_{\text{ols},j}$ . Thus the LASSO, ridge, and selection estimators are all transformations of the least squares coefficient estimator. We illustrate these

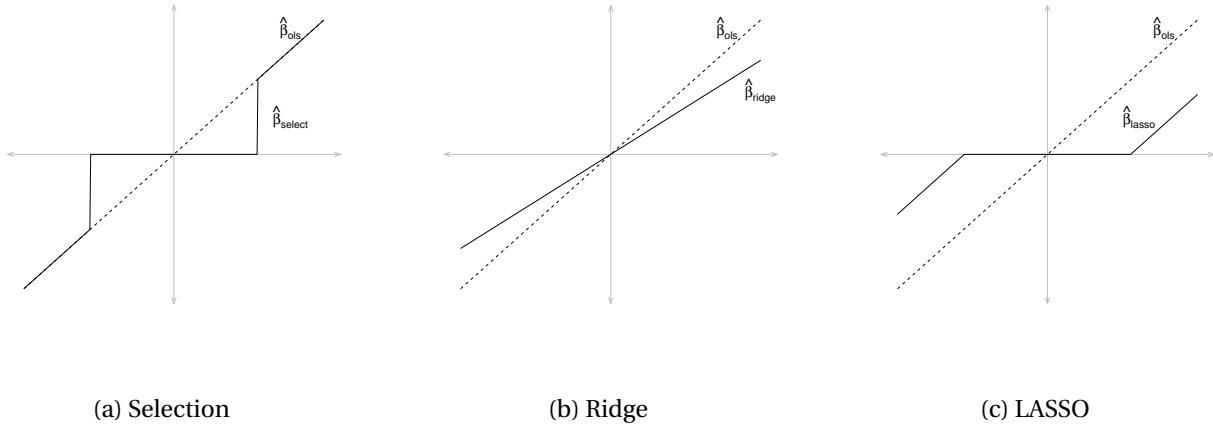


Figure 24.15: Transformations of least squares coefficients by Selection, Ridge, and LASSO estimators

transformations in Figure 24.15. Panel (a) displays the selection transformation, panel (b) displays the ridge transformation, and panel (c) displays the LASSO transformation.

The LASSO and ridge estimators are continuous functions while the selection estimator is a discontinuous function. The LASSO and selection estimators are thresholding functions, meaning that the function equals zero for a region about the origin. Thresholding estimators are selection estimators, since they equal zero when the least squares estimator is sufficiently small. The LASSO function is a “soft thresholding” rule as it is a continuous function with bounded first derivative. The selection estimator is a “hard thresholding” rule as it is discontinuous. Hard thresholding rules tend to have high variance due to the discontinuous transformation. Consequently we expect the LASSO to have reduced variance relative to selection estimators, permitting overall lower MSE.

As for ridge regression LASSO is not invariant to the scaling of the regressors. If you rescale a regressor then the penalty has a completely different meaning. Consequently it is important to scale the regressors appropriately before applying LASSO. It is conventional to scale all the variables to have mean zero and unit variance.

LASSO is also not invariant to rotations of the regressors. For example, LASSO on  $(X_1, X_2)$  is not the same as LASSO on  $(X_1 - X_2, X_2)$  despite having identical least-squares solutions. This is troubling as typically there is no default specification.

Critically important for LASSO estimation is the choice of penalty  $\lambda$ . The most common choice is minimization of K-fold cross validation. Leave-one-out CV is not used as it is computationally expensive. K-fold is a computationally feasible substitute. Many programs set the default number of folds as  $K = 10$ , though some authors use  $K = 5$ , while others recommend  $K = 20$ . It is common to find that the results of K-fold CV can be sensitive across runs (the methods depends on the random sorting of the observations). In this context it is prudent to use a large number of folds  $K$  to reduce the randomness.

## 24.33 Computation of the LASSO Estimator

The constraint representation of LASSO is minimization of a quadratic subject to linear inequality constraints, so can be implemented by standard quadratic programming. This is a computationally simple approach to estimation. For evaluation of the cross-validation function, however, it is useful to compute the entire LASSO path. For this a computationally appropriate method is the modified LARS algorithm. (LARS stands for least angle regression.)

The LARS algorithm produces a path of coefficients starting at the origin and ending at least squares (when  $k < n$ ). The sequence corresponds to the sequence of constraints  $\tau$  which can be calculated by the absolute sum of the coefficients, but these values (nor  $\lambda$ ) is used by the algorithm. The steps are as

follows.

1. Start with all coefficients equal to zero.
2. Find  $x_j$  most correlated with  $y$ .
3. Increase  $\beta_j$  in the direction of correlation.
  - (a) Compute residuals along the way.
  - (b) Stop when some other  $x_\ell$  has the same correlation with the residual as  $x_j$ .
  - (c) If a non-zero coefficient hits zero, drop from the active set of variables and recompute the joint least squares direction.
4. Increase  $(\beta_j, \beta_\ell)$  in their joint least squares direction until some other  $x_m$  has the same correlation with the residual.
5. Repeat until all predictors are in model.

This algorithm produces the LASSO path, but the equality between the two is not immediately apparent so we do not show this here.

### 24.34 Elastic Net

The difference between LASSO and ridge regression is that the LASSO uses an  $L_1$  penalty while ridge uses an  $L_2$  penalty. Since the two procedures both have advantages it seems reasonable that further improvements can be obtained by taking a compromise between the two. While one might try an  $L_q$  penalty for some  $1 < q < 2$  it turns out that this is not computationally attractive. Instead, a similar penalty can be obtained by taking a linear combination of the  $L_1$  and  $L_2$  penalties. This is typically written as

$$\text{SSE}(\boldsymbol{\beta}, \lambda, \alpha) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \left( \alpha \|\boldsymbol{\beta}\|_2^2 + (1 - \alpha) \|\boldsymbol{\beta}\|_1 \right)$$

for  $0 \leq \alpha \leq 1$  and is called the **Elastic Net**. For  $\alpha = 0$  we obtain LASSO and for  $\alpha = 1$  we obtain ridge regression. For small but positive  $\alpha$  the constraint sets are similar to “rounded” versions of the LASSO constraint sets.

Typically the parameters  $(\alpha, \lambda)$  are selected by joint minimization of the K-fold cross-validation criterion. Since the elastic net penalty is linear-quadratic the solution is computationally similar to LASSO.

### 24.35 Regression Sample Splitting

Suppose we have observations  $\{y_i, x_i : i = 1, \dots, n\}$ . Consider the model

$$\begin{aligned} y_i &= \mu_1 \mathbf{1}(x_i \leq \gamma) + \mu_2 \mathbf{1}(x_i > \gamma) + e_i \\ \mathbb{E}(e_i | x_i) &= 0 \end{aligned}$$

where  $(\mu_1, \mu_2, \gamma)$  are unknown parameters. This model specifies that the conditional mean of  $y_i$  is a step function in  $x_i$ , taking the value  $\mu_1$  for small values of  $x_i$ , the value  $\mu_2$  for large values of  $x_i$ , with a step at  $x_i = \gamma$ . For a regression tree this will be viewed as an approximation and many such splits will be applied, but for now let's take the simple case of a single split. To give a concrete example, suppose  $y_i$  is college GPA and  $x_i$  is entrance test score. The model approximates college performance by dividing students into two groups based on their entrance test score, but the optimal split point is treated as unknown.

The question we explore in this section is how to estimate the parameters  $(\mu_1, \mu_2, \gamma)$ .

The standard solution is (nonlinear) least squares. The coefficients  $(\hat{\mu}_1, \hat{\mu}_2, \hat{\gamma})$  minimize the sum of squared errors

$$\text{SSE}(\mu_1, \mu_2, \gamma) = \sum_{i=1}^n (y_i - \mu_1 \mathbf{1}(x_i \leq \gamma) - \mu_2 \mathbf{1}(x_i > \gamma))^2.$$

By nested minimization

$$\hat{\gamma} = \operatorname{argmin}_{\gamma} \min_{\mu_1, \mu_2} \text{SSE}(\mu_1, \mu_2, \gamma) = \operatorname{argmin}_{\gamma} \text{SSE}^*(\gamma),$$

say. The inner minimization holds  $\gamma$  fixed. This is a regression on two dummy variables, which has the simple solution of taking the sample means of the two subsamples. These are

$$\begin{aligned}\hat{\mu}_1(\gamma) &= \frac{\sum_{i=1}^n y_i \mathbf{1}(x_i \leq \gamma)}{\sum_{i=1}^n \mathbf{1}(x_i \leq \gamma)} \\ \hat{\mu}_2(\gamma) &= \frac{\sum_{i=1}^n y_i \mathbf{1}(x_i > \gamma)}{\sum_{i=1}^n \mathbf{1}(x_i > \gamma)}\end{aligned}$$

the sample means for the observations where  $\mathbf{1}(x_i \leq \gamma)$  and  $\mathbf{1}(x_i > \gamma)$ , respectively. We can write

$$\text{SSE}^*(\gamma) = \sum_{i=1}^n (y_i - \hat{\mu}_1(\gamma) \mathbf{1}(x_i \leq \gamma) - \hat{\mu}_2(\gamma) \mathbf{1}(x_i > \gamma))^2,$$

the sum of squared errors after subtracting the split-sample means.

The function  $\text{SSE}^*(\gamma)$  is a step function taking jumps at the sample values of  $x_i$ . Hence it can be minimized by searching over the unique values of the latter. Given the minimizer  $\hat{\gamma}$ , the coefficients  $\hat{\mu}_1$  and  $\hat{\mu}_2$  are found by taking the sample means in the two split samples.

In summary, the algorithm for obtaining the least squares coefficient estimates  $(\hat{\mu}_1, \hat{\mu}_2, \hat{\gamma})$  is as follows.

1. Find the unique  $N \leq n$  sample values  $x_j$  of  $x_i$ .
2. For each of these values
  - (a) Set  $\gamma = x_j$ .
  - (b) Split the sample into two groups: those with  $x_i \leq \gamma$  and those with  $x_i > \gamma$ .
  - (c) Set  $\hat{\mu}_1(\gamma)$  and  $\hat{\mu}_2(\gamma)$  as the sample mean of  $y_i$  on each subsample.
  - (d) Calculate the sum of squared errors  $\text{SSE}^*(\gamma)$  on the full sample.
3. Find  $\hat{\gamma}$  which minimizes  $\text{SSE}^*(\gamma)$ .
4. Split the sample into two groups: those with  $x_i \leq \hat{\gamma}$  and those with  $x_i > \hat{\gamma}$ .
5. Set  $\hat{\mu}_1$  and  $\hat{\mu}_2$  as the sample mean of  $y_i$  on each subsample.

This algorithm requires  $N = O(n)$  regressions.

Now suppose we have observations  $\{y_i, x_{1i}, \dots, x_{ki} : i = 1, \dots, n\}$ . Consider the model

$$\begin{aligned}y_i &= \mu_1 \mathbf{1}(x_{di} \leq \gamma) + \mu_2 \mathbf{1}(x_{di} > \gamma) + e_i \\ \mathbb{E}(e_i | x_i) &= 0\end{aligned}$$

where the index  $d$  is unknown. This means that there is a single way to split the sample, but it is unknown which regressor to use. To give a concrete example once again let  $y_i$  be college GPA and  $x_{di}$  a set of predictors, such as high school GPA, letters of recommendation, participation in sports, extra curricular activities, and region of residence. Again the question is how to divide the students into “high expected

performance" and "low expected performance" but it is unknown which predictor is most useful and which threshold to use. Again the goal is to estimate the coefficients, augmented to include  $d$ .

The least squares estimator  $(\hat{\mu}_1, \hat{\mu}_2, \hat{\gamma}, \hat{d})$  of the coefficients is obtained as a simple extension of the previous algorithm. We simply add an extra loop by searching across the regressors. The least squares estimator can be obtained by the following algorithm.

1. For  $d = 1, \dots, k$ 
  - (a) Perform steps 1-3 of the previous algorithm using variable  $x_{di}$ .
  - (b) Store the minimized  $\text{SSE}^*(d) = \text{SSE}^*(\hat{\gamma})$ .
2. Find  $\hat{d}$  which minimizes  $\text{SSE}^*(d)$ .
3. Given  $\hat{d}$ , estimate model as in the previous algorithm.

This algorithm requires  $\sum_{d=1}^k N_d = O(kn)$  regressions, where  $N_d$  is the number of unique values of  $x_{ji}$ . In most applications the number of regressions is much less than  $kn$ , because many of the regressors will be discrete.

## 24.36 Regression Trees

A regression tree is a nonparametric regression using a large number of step functions. The idea is that if a sufficiently large number of step functions (sample splits) are used then a step function can be a good approximation to any functional form. Regression trees may be especially useful in regression with discrete variables, where traditional kernel and series methods are not appropriate.

The literature on regression trees has developed some colorful language to describe the tools, based on the metaphor of a living tree.

1. A split point is **node**.
2. A subsample is a **branch**.
3. Increasing the set of nodes is **growing** a tree.
4. Decreasing the set of nodes is **pruning** a tree.

The basic structure of the regression tree algorithm is to start with zero nodes. Grow a large (non-parsimonious) tree. Then prune back using an information criterion. The goal of the growth stage is to develop a rich data-determined tree which has small bias but high variance (due to overparameterization). Pruning back is an application of backward stepwise regression, with the goal of reducing over-parameterization and estimation variance.

The basic regression tree growth algorithm is as follows. Assume the observations are  $\{y_i, x_{1i}, \dots, x_{ki} : i = 1, \dots, n\}$ .

1. Select a maximum number  $N$  of nodes.
2. Sequentially apply regression sample splits.
  - (a) Apply the regression sample split algorithm to split the sample into two groups.
  - (b) Apply the regression sample split algorithm on each sub-sample.
  - (c) On each branch  $b$ 
    - i. Take the sample mean  $\hat{\mu}_b$  of  $y_i$  for observations on the branch.
    - ii. This is the estimator of the regression function on this branch.
    - iii. The residuals on the branch are  $\hat{e}_i = y_i - \hat{\mu}_b$ .

- (d) Select the split which produces the lowest sum of squared errors.
- (e) Repeat (b)-(d) until there are  $N$  nodes (splits).

After the regression tree growth algorithm has been run, the estimated regression is a multi-dimensional step function with  $N$  jump points.

The basic pruning algorithm is as follows.

1. Define the Mallows-type information criterion

$$C = \sum_{i=1}^n \hat{e}_i^2 + \alpha N$$

where  $N$  is the number of nodes and  $\alpha$  is a penalty parameter.

2. Compute the criterion  $C$  for the current tree.
3. Use backward stepwise regression to reduce the number of nodes:
  - (a) Identify the set of terminal nodes (those with no further splits).
  - (b) Identify the terminal node whose removal most decreases  $C$ .
  - (c) Prune (remove) this node.
  - (d) If there is no terminal node whose removal decreases  $C$  then stop pruning.
  - (e) Otherwise, repeat (a)-(d).

The use of the Mallows-type criterion for node selection is presumably used because of its simplicity even though there is not a strong theoretical case for the criterion. Regression sample splits is a non-linear regression model for which the traditional Mallows theory does not apply. This means that it is unclear what is a reasonable choice for  $\alpha$ . Consequently a typical implementation is to use K-fold cross-validation to select  $\alpha$ .

The results of a regression tree are difficult to interpret if you are looking for regression coefficients (there are none). Rather the output are direct estimates of the conditional mean for sub-populations. Regression trees are typically used for prediction.

### 24.37 Bagging

Bagging refers to **bootstrap aggregation**. We focus here on its use for estimation of a regression (conditional mean) model. The basic idea is quite simple. You generate a large number  $B$  of bootstrap samples, estimate your regression model on each bootstrap sample, and take the average of the bootstrap regression estimates. The mean of the bootstrap estimates is the bagging estimator.

Bagging is believed to be useful when the conditional mean estimator has low bias but high variance. High variance occurs for hard thresholding estimators such as regression trees and model selection. Bagging is a smoothing operation, which introduces bias but reduces variance. The resulting bagging estimator can have lower MSE as a result. Bagging is believed to be less useful for estimators with high bias, as bagging tends to exaggerate the bias.

We first describe the estimation algorithm. Let  $m(\mathbf{x}) = \mathbb{E}(y_i | \mathbf{x}_i = \mathbf{x})$  be the conditional mean and  $\hat{m}(\mathbf{x})$  an estimator (such as a regression tree). Let  $\hat{m}_b(\mathbf{x})$  be the same estimator constructed on an independent bootstrap sample generated by i.i.d. sampling from the observations. The bagging estimator of  $m(\mathbf{x})$  is

$$\hat{m}_{\text{bag}}(\mathbf{x}) = \frac{1}{B} \sum_{B=1}^b \hat{m}_b(\mathbf{x}).$$

## 24.38 Random Forests

Random forests are a modification of bagged regression trees. The modification is designed to further reduce estimation variance. Random forests are currently quite popular in machine learning applications.

Consider the procedure of applying bagging to regression trees. Since bootstrap samples are similar to one another the estimated bootstrap regression trees are similar to one another, particularly in the sense that they will tend to make sample splits on the same variables. This means that conditional on the sample the bootstrap regression trees are positively correlated. This correlation means that the variance of the bootstrap average remains high even when the number of bootstrap replications  $B$  is large. The modification proposed by random forests is to *decorrelate* the bootstrap regression trees by introducing extra randomness.

The basic random forest algorithm is as follows.

1. Pick a minimum node size  $N_{\min}$  (recommended to set  $N_{\min} = 5$ ).
2. Pick the number of variables  $m < k$  to select at random (recommended to set  $m = k/3$ ).
3. For  $b = 1, \dots, B$ 
  - (a) Draw a random bootstrap sample.
  - (b) Grow a regression tree on the bootstrap sample using the following steps until you have  $N_{\min}$  nodes:
    - i. Select  $m$  variables at random from the  $k$  regressors.
    - ii. Pick the best variable and node for a regression tree from among these  $m$  variables.
    - iii. Split the sample at the node.
  - (c) Set  $\hat{m}_b(\mathbf{x})$  as for a regression tree as the sample mean of  $y_i$  on each branch of the bootstrap tree.
4.  $\hat{m}_{\text{rf}}(\mathbf{x}) = \frac{1}{B} \sum_{B=1}^B \hat{m}_b(\mathbf{x}).$

Using randomization to reduce the number of variables from  $k$  to  $m$  at each step reduces the correlation across the bootstrapped regression trees and hence reduces the variance of the bootstrap average.

## 24.39 Ensembling

Ensembling is the term used by the machine learning literature to signify model averaging across machine learning algorithms. Ensembling is very popular in applied machine learning.

Suppose you have a set of estimators (e.g., CV selection, James-Stein shrinkage, JMA, SBIC, PCA, kernel regression, series regression, ridge regression, LASSO, regression tree, bagged regression tree, and random forest). Which should you use? The principle of model averaging suggests that you can do better by taking a weighted average rather than just selecting one over the other. It is reasonable to expect that one method may work well with some types of data, and other methods may work well with other types of data. Can we use the data to inform us about which is the best weighted average?

We briefly describe here one approach for selection of the averaging weights.

Assume you are trying to predict  $y_i$  and you have  $M$  predictors  $\hat{y}_{Mi}$ . Consider the regression model

$$y_i = w_1 \hat{y}_{1i} + w_2 \hat{y}_{2i} + \cdots + w_M \hat{y}_{Mi} + e_i.$$

The coefficients are the weights. We can estimate the weights by regression of  $y_i$  on the in-sample forecasts. The estimation should not be done by unpenalized least squares as this would simply lead to putting all weight on the most complicated model. Instead, the recommendation is to estimate the weights using a LASSO regression to enforce regularity.

## 24.40 Technical Proofs\*

**Proof of Theorem 24.2:** We establish the theorem under the simplifying assumptions of the normal linear regression model with a  $K \times 1$  coefficient vector  $\beta$  and known variance  $\sigma^2$ . The likelihood function is

$$L(\beta) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}'_i \beta)^2\right).$$

Evaluated at the MLE  $\hat{\beta}$  this equals

$$L(\hat{\beta}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^n \hat{e}_i^2}{2\sigma^2}\right). \quad (24.39)$$

Using (8.21) we can write

$$\begin{aligned} L(\beta) &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \left( \sum_{i=1}^n \hat{e}_i^2 + (\hat{\beta} - \beta)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta) \right) \right) \\ &= L(\hat{\beta}) \exp\left(-\frac{1}{2\sigma^2} (\hat{\beta} - \beta)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta)\right). \end{aligned}$$

For a diffuse prior  $\pi(\beta) = C$  the marginal likelihood is

$$\begin{aligned} p(\mathbf{y}) &= L(\hat{\beta}) \int \exp\left(-\frac{1}{2\sigma^2} (\hat{\beta} - \beta)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta)\right) C d\beta \\ &= L(\hat{\beta}) n^{-K/2} (2\pi\sigma^2)^{K/2} \det\left(\frac{1}{n} \mathbf{X}' \mathbf{X}\right)^{-1/2} C \end{aligned}$$

where the final equality is the multivariate normal integral. Rewriting and taking logs

$$\begin{aligned} -2 \log p(\mathbf{y}) &= -2 \log L(\hat{\beta}) + K \log n - K \log(2\pi\sigma^2) + \log \det\left(\frac{1}{n} \mathbf{X}' \mathbf{X}\right) + \log C \\ &= -2 \log L(\hat{\beta}) + K \log n + O(1). \end{aligned}$$

This is the theorem. ■

**Proof of Theorem 24.3:** From (24.13)

$$\begin{aligned} \int g(\mathbf{y}) \log f(\mathbf{y}, \hat{\theta}) d\mathbf{y} &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n \int (y - \mathbf{x}'_i \hat{\beta})^2 g(y | \mathbf{x}_i) dy \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n \left( \sigma^2 + (\hat{\beta} - \beta)' \mathbf{x}_i \mathbf{x}'_i (\hat{\beta} - \beta) \right) \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{n}{2} - \frac{1}{2\sigma^2} \mathbf{e}' \mathbf{P} \mathbf{e}. \end{aligned}$$

Thus

$$\begin{aligned} T &= -2 \mathbb{E} \left( \int g(\mathbf{y}) \log \hat{f}(\mathbf{y}) d\mathbf{y} \right) \\ &= n \log(2\pi\sigma^2) + n + \frac{1}{\sigma^2} \mathbb{E}(\mathbf{e}' \mathbf{P} \mathbf{e}) \\ &= n \log(2\pi\sigma^2) + n + K. \end{aligned}$$

This is (24.14). The final equality holds under the assumption of conditional homoskedasticity.

Evaluating (24.13) at  $\hat{\boldsymbol{\beta}}$  we obtain the log likelihood

$$\begin{aligned}-2\log L(\hat{\boldsymbol{\beta}}) &= n \log(2\pi\sigma^2) + \frac{1}{\sigma^2} \sum_{i=1}^n \hat{e}_i^2 \\ &= n \log(2\pi\sigma^2) + \frac{1}{\sigma^2} \mathbf{e}' \mathbf{M} \mathbf{e}.\end{aligned}$$

This has expectation

$$\begin{aligned}-\mathbb{E}(2\log L(\hat{\boldsymbol{\theta}})) &= n \log(2\pi\sigma^2) + \frac{1}{\sigma^2} \mathbb{E}(\mathbf{e}' \mathbf{P} \mathbf{e}) \\ &= n \log(2\pi\sigma^2) + n - K.\end{aligned}$$

This is (24.15). The final equality holds under conditional homoskedasticity. ■

**Proof of Theorem 24.5:** The proof uses Taylor expansions similar to those used for the asymptotic distribution theory of the MLE in nonlinear models. We avoid technical details so this is not a full proof.

Write the model density as  $f(\mathbf{y}, \boldsymbol{\theta})$  and the estimated model as  $\hat{f}(\mathbf{y}) = f(\mathbf{y}, \hat{\boldsymbol{\theta}})$ . Recall from (24.12) that we can write the target  $T$  as

$$T = -2\mathbb{E}(\log f(\tilde{\mathbf{y}}, \hat{\boldsymbol{\theta}}))$$

where  $\tilde{\mathbf{y}}$  is an independent copy of  $\mathbf{y}$ . Let  $\tilde{\boldsymbol{\theta}}$  be the MLE calculated on the sample  $\tilde{\mathbf{y}}$ . This is an independent copy of  $\hat{\boldsymbol{\theta}}$ . By symmetry we can write  $T$  as

$$T = -2\mathbb{E}(\log f(\mathbf{y}, \tilde{\boldsymbol{\theta}})). \quad (24.40)$$

Define the Hessian

$$H = -\frac{\partial}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \mathbb{E}(\log f(\mathbf{y}, \boldsymbol{\theta})) > 0.$$

Now take a second-order Taylor series expansion of the log likelihood  $\log f(\mathbf{y}, \hat{\boldsymbol{\theta}})$  about  $\hat{\boldsymbol{\theta}}$ . This is

$$\begin{aligned}\log f(\mathbf{y}, \tilde{\boldsymbol{\theta}}) &= \log f(\mathbf{y}, \hat{\boldsymbol{\theta}}) + \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{y}, \hat{\boldsymbol{\theta}})' (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) - \frac{1}{2} (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})' H (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) + O_p(n^{-1/2}) \\ &= \log f(\mathbf{y}, \hat{\boldsymbol{\theta}}) - \frac{n}{2} (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})' H (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) + O_p(n^{-1/2}).\end{aligned} \quad (24.41)$$

The second equality holds because of the first-order condition for the MLE  $\hat{\boldsymbol{\theta}}$ .

If the  $O_p(n^{-1/2})$  term in (24.41) is uniformly integrable, (24.40) and (24.41) imply that

$$\begin{aligned}T &= -\mathbb{E}(2\log f(\mathbf{y}, \hat{\boldsymbol{\theta}})) + \mathbb{E}\left(n(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})' H (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})\right) + O(n^{-1/2}) \\ &= -\mathbb{E}(2\log L(\hat{\boldsymbol{\theta}})) + \mathbb{E}\left(n(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})' H (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})\right) + \mathbb{E}\left(n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})' H (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\right) \\ &\quad + 2\mathbb{E}\left(n(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})' H (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\right) + O(n^{-1/2}) \\ &= -\mathbb{E}(2\log L(\hat{\boldsymbol{\theta}})) + \mathbb{E}(\chi_K^2) + \mathbb{E}(\tilde{\chi}_K^2) + O(n^{-1/2}) \\ &= -\mathbb{E}(2\log L(\hat{\boldsymbol{\theta}})) + 2K + O(n^{-1/2})\end{aligned}$$

where  $\chi_K^2$  and  $\tilde{\chi}_K^2$  are chi-square random variables with  $K$  degrees of freedom. The second-to-last equality holds if

$$n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})' H (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \chi_K^2 \quad (24.42)$$

and the Wald statistic on the left-side of (24.42) is uniformly integrable. The asymptotic convergence (24.42) holds for the MLE under standard regularity conditions (including correct specification). ■

**Proof of Theorem 24.6:** Our analysis is conditional on the regressors. For simplicity we write conditional expectations as unconditional expectations to reduce notational clutter.

Using matrix notation we can write  $\hat{\mathbf{m}} - \mathbf{m} = -(\mathbf{I}_n - \mathbf{A})\mathbf{m} + \mathbf{A}\mathbf{e}$ . We can then write the fit as

$$\begin{aligned} R &= \mathbb{E}((\hat{\mathbf{m}} - \mathbf{m})'(\hat{\mathbf{m}} - \mathbf{m})) \\ &= \mathbb{E}(\mathbf{m}'(\mathbf{I}_n - \mathbf{A}')(\mathbf{I}_n - \mathbf{A})\mathbf{m} - 2\mathbf{m}'(\mathbf{I}_n - \mathbf{A}')\mathbf{A}\mathbf{e} + \mathbf{e}'\mathbf{A}'\mathbf{A}\mathbf{e}) \\ &= \mathbf{m}'(\mathbf{I}_n - \mathbf{A}')(\mathbf{I}_n - \mathbf{A})\mathbf{m} + \sigma^2 \text{tr}(\mathbf{A}'\mathbf{A}). \end{aligned}$$

Notice that this calculation relies on the assumption of conditional homoskedasticity.

Now consider the Mallows criterion. We find that

$$\begin{aligned} C_p^* &= \hat{\mathbf{e}}'\hat{\mathbf{e}} + 2\tilde{\sigma}^2 \text{tr}(\mathbf{A}) - \mathbf{e}'\mathbf{e} \\ &= (\mathbf{m} + \mathbf{e})'(\mathbf{I}_n - \mathbf{A}')(\mathbf{I}_n - \mathbf{A})(\mathbf{m} + \mathbf{e}) + 2\tilde{\sigma}^2 \text{tr}(\mathbf{A}) - \mathbf{e}'\mathbf{e} \\ &= \mathbf{m}'(\mathbf{I}_n - \mathbf{A}')(\mathbf{I}_n - \mathbf{A})\mathbf{m} + 2\mathbf{m}'(\mathbf{I}_n - \mathbf{A}')(\mathbf{I}_n - \mathbf{A})\mathbf{e} + \mathbf{e}'\mathbf{A}'\mathbf{A}\mathbf{e} - 2\mathbf{e}'\mathbf{A}\mathbf{e} + 2\tilde{\sigma}^2 \text{tr}(\mathbf{A}). \end{aligned}$$

Taking expectations and using the assumptions of conditional homoskedasticity and  $\mathbb{E}(\tilde{\sigma}^2) = \sigma^2$

$$\begin{aligned} \mathbb{E}(C_p^*) &= \mathbf{m}'(\mathbf{I}_n - \mathbf{A}')(\mathbf{I}_n - \mathbf{A})\mathbf{m} + \sigma^2 \text{tr}(\mathbf{A}'\mathbf{A}) \\ &= R. \end{aligned}$$

This is the result as stated. ■

**Proof of Theorem 24.7:** Take any two models  $\mathcal{M}_1$  and  $\mathcal{M}_2$  where  $\mathcal{M}_1 \notin \overline{\mathcal{M}}^*$  and  $\mathcal{M}_2 \in \overline{\mathcal{M}}^*$ . Let their information criterion be written as

$$\begin{aligned} \text{IC}_1 &= -2 \log L_1(\hat{\boldsymbol{\theta}}_1) + c(n, K_1) \\ \text{IC}_2 &= -2 \log L_2(\hat{\boldsymbol{\theta}}_2) + c(n, K_2). \end{aligned}$$

Model  $\mathcal{M}_1$  is selected over  $\mathcal{M}_2$  if

$$LR < c(n, K_2) - c(n, K_1)$$

where

$$LR = 2(\log L_2(\hat{\boldsymbol{\theta}}_2) - \log L_1(\hat{\boldsymbol{\theta}}_1))$$

where  $LR$  is the likelihood ratio statistic for testing  $\mathcal{M}_1$  against  $\mathcal{M}_2$ . Since we have assumed that  $\mathcal{M}_1$  is not a true model while  $\mathcal{M}_2$  is true, then  $LR$  diverges to  $+\infty$  at rate  $n$ . This means that for any  $\alpha > 0$ ,  $n^{-1+\alpha}LR \xrightarrow{P} +\infty$ . Furthermore, the assumptions imply  $n^{-1+\alpha}(c(n, K_1) - c(n, K_2)) \rightarrow 0$ . Fix  $\varepsilon > 0$ . There is an  $n$  sufficiently large such that  $n^{-1+\alpha}(c(n, K_1) - c(n, K_2)) < \varepsilon$ . Thus

$$\begin{aligned} \Pr(\widehat{\mathcal{M}} = \mathcal{M}_1) &\leq \Pr(n^{-1+\alpha}LR < n^{-1+\alpha}(c(n, K_2) - c(n, K_1))) \\ &\leq \Pr(LR < \varepsilon) \\ &\rightarrow 0. \end{aligned}$$

Since this holds for any  $\mathcal{M}_1 \notin \overline{\mathcal{M}}^*$  we deduce that the selected model is in  $\overline{\mathcal{M}}^*$  with probability approaching one. This means that the selection criterion is model selection consistent as claimed. ■

**Proof of Theorem 24.8:** Take the setting as described in the proof of Theorem 24.7 but now assume  $\mathcal{M}_1 \subset \mathcal{M}_2$  and  $\mathcal{M}_1, \mathcal{M}_2 \in \overline{\mathcal{M}}^*$ . The likelihood ratio statistic satisfies  $LR \xrightarrow{d} \chi_r^2$  where  $r = K_2 - K_1$ . Let

$$B = \limsup_{n \rightarrow \infty} (c(n, K_1) - c(n, K_2)) < \infty.$$

Letting  $F_r(u)$  denote the  $\chi^2_r$  distribution function

$$\begin{aligned}\Pr(\widehat{\mathcal{M}} = \mathcal{M}_2) &= \Pr(LR > (c(n, K_2) - c(n, K_1))) \\ &\geq \Pr(LR > B) \\ &\longrightarrow \Pr(\chi^2_r > B) \\ &= 1 - F_r(B) \\ &> 0\end{aligned}$$

since  $\chi^2_r$  has support over the positive real line and  $B < \infty$ . This shows that the selection criterion asymptotically over-selects with positive probability. ■

**Proof of Theorem 24.9:** Since  $c(n, K) = o(n)$  the procedure is model selection consistent. Take two models  $\mathcal{M}_1, \mathcal{M}_2 \in \overline{\mathcal{M}}^*$  with  $K_1 < K_2$ . Since both models are true then  $LR = O_p(1)$ . Fix  $\varepsilon > 0$ . There is a  $B < \infty$  such that  $LR \leq B$  with probability exceeding  $1 - \varepsilon$ . By (24.19) there is an  $n$  sufficiently large such that  $(c(n, K_2) - c(n, K_1)) > B$ . Thus

$$\begin{aligned}\Pr(\widehat{\mathcal{M}} = \mathcal{M}_2) &\leq \Pr(LR > (c(n, K_2) - c(n, K_1))) \\ &\leq \Pr(LR > B) \\ &\leq \varepsilon.\end{aligned}$$

Since  $\varepsilon$  is arbitrary  $\Pr(\widehat{\mathcal{M}} = \mathcal{M}_2) \rightarrow 0$  as claimed. ■

**Proof of Theorem 24.10:** First, we examine  $R_n(K)$ . Write the predicted values in matrix notation as  $\hat{\mathbf{m}}_K = \mathbf{X}_K \hat{\boldsymbol{\beta}}_K = \mathbf{P}_K \mathbf{y}$  where  $\mathbf{P}_K = \mathbf{X}_K (\mathbf{X}'_K \mathbf{X}_K)^{-1} \mathbf{X}'_K$ . It is useful to observe that  $\mathbf{m} - \hat{\mathbf{m}}_K = \mathbf{M}_K \mathbf{m} - \mathbf{P}_K \mathbf{e}$  where  $\mathbf{M}_K = \mathbf{I}_K - \mathbf{P}_K$ . We find that the prediction risk equals

$$\begin{aligned}R_n(K) &= \mathbb{E}((\mathbf{m} - \hat{\mathbf{m}}_K)' (\mathbf{m} - \hat{\mathbf{m}}_K)) \\ &= \mathbb{E}((\mathbf{M}_K \mathbf{m} - \mathbf{P}_K \mathbf{e})' (\mathbf{M}_K \mathbf{m} - \mathbf{P}_K \mathbf{e})) \\ &= \mathbf{m}' \mathbf{M}_K \mathbf{m} + \mathbb{E}(\mathbf{e}' \mathbf{P}_K \mathbf{e}) \\ &= \mathbf{m}' \mathbf{M}_K \mathbf{m} + \sigma^2 K.\end{aligned}$$

The choice of regressors affects  $R_n(K)$  through the two terms in the final line. The first term  $\mathbf{m}' \mathbf{M}_K \mathbf{m}$  is the squared bias due to omitted variables. As  $K$  increases this term decreases reflecting reduced omitted variables bias. The second term  $\sigma^2 K$  is estimation variance. It is increasing in the number of regressors. Increasing the number of regressors affects the quality of out-of-sample prediction by reducing the bias but increasing the variance.

We next examine the adjusted Mallows criterion. We find that

$$\begin{aligned}C_n^*(K) &= \hat{\mathbf{e}}'_K \hat{\mathbf{e}}_K + 2\sigma^2 K - \mathbf{e}' \mathbf{e} \\ &= (\mathbf{m} + \mathbf{e})' \mathbf{M}_K (\mathbf{m} + \mathbf{e}) + 2\sigma^2 K - \mathbf{e}' \mathbf{e} \\ &= \mathbf{m}' \mathbf{M}_K \mathbf{m} + 2\mathbf{m}' \mathbf{M}_K \mathbf{e} - \mathbf{e}' \mathbf{P}_K \mathbf{e} + 2\sigma^2 K.\end{aligned}$$

The next step is to show that

$$\sup_K \left| \frac{C_n^*(K) - R_n(K)}{R_n(K)} \right| \xrightarrow{p} 0 \quad (24.43)$$

as  $n \rightarrow \infty$ . To establish (24.43), observe that

$$C_n^*(K) - R_n(K) = 2\mathbf{m}' \mathbf{M}_K \mathbf{e} - \mathbf{e}' \mathbf{P}_K \mathbf{e} + \sigma^2 K.$$

Pick  $\varepsilon > 0$  and some sequence  $B_n \rightarrow \infty$  such that  $B_n / (R_n^{\text{opt}})^r \rightarrow 0$ . (This is feasible by Assumption 24.1.5.) By Boole's inequality (B.24), Whittle's inequality (B.48), the facts that  $\mathbf{m}' \mathbf{M}_K \mathbf{m} \leq R_n(K)$  and  $R_n(K) \geq \sigma^2 K$ ,  $B_n / (R_n^{\text{opt}})^r \rightarrow 0$ , and  $\sum_{K=1}^{\infty} K^{-r} < \infty$

$$\begin{aligned} \mathbb{P}\left(\sup_K \left|\frac{\mathbf{m}' \mathbf{M}_K \mathbf{e}}{R_n(K)}\right| > \varepsilon\right) &\leq \sum_{K=1}^{\infty} \mathbb{P}\left(\left|\frac{\mathbf{m}' \mathbf{M}_K \mathbf{e}}{R_n(K)}\right| > \varepsilon\right) \\ &\leq \frac{C_{1r}}{\varepsilon^{2r}} \sum_{K=1}^{\infty} \frac{|\mathbf{m}' \mathbf{M}_K \mathbf{m}|^r}{R_n(K)^{2r}} \\ &\leq \frac{C_{1r}}{\varepsilon^{2r}} \sum_{K=1}^{\infty} \frac{1}{R_n(K)^r} \\ &= \frac{C_{1r}}{\varepsilon^{2r}} \sum_{K=1}^{B_n} \frac{1}{R_n(K)^r} + \frac{C_{1r}}{\varepsilon^{2r}} \sum_{K=B_n+1}^{\infty} \frac{1}{R_n(K)^r} \\ &\leq \frac{C_{1r}}{\varepsilon^{2r}} \frac{B_n}{(R_n^{\text{opt}})^r} + \frac{C_{1r}}{\varepsilon^{2r} \sigma^{2r}} \sum_{K=B_n+1}^{\infty} \frac{1}{K^r} \\ &\longrightarrow 0. \end{aligned}$$

By a similar argument but using Whittle's inequality (B.49),  $\text{tr}(\mathbf{P}_K \mathbf{P}_K) = \text{tr}(\mathbf{P}_K) = K$ , and  $K \leq \sigma^{-2} R_n(K)$

$$\begin{aligned} \mathbb{P}\left(\sup_K \left|\frac{\mathbf{e}' \mathbf{P}_K \mathbf{e} - \sigma^2 K}{R_n(K)}\right| > \varepsilon\right) &\leq \sum_{K=1}^{\infty} \mathbb{P}\left(\left|\frac{\mathbf{e}' \mathbf{P}_K \mathbf{e} - \mathbb{E}(\mathbf{e}' \mathbf{P}_K \mathbf{e})}{R_n(K)}\right| > \varepsilon\right) \\ &\leq \frac{C_{2r}}{\varepsilon^{2r}} \sum_{K=1}^{\infty} \frac{\text{tr}(\mathbf{P}_K \mathbf{P}_K)^r}{R_n(K)^{2r}} \\ &= \frac{C_{2r}}{\varepsilon^{2r}} \sum_{K=1}^{\infty} \frac{K^r}{R_n(K)^{2r}} \\ &\leq \frac{C_{1r}}{\varepsilon^{2r} \sigma^{2r}} \sum_{K=1}^{\infty} \frac{1}{R_n(K)^r} \\ &\longrightarrow 0. \end{aligned}$$

Together these imply (24.43).

Finally we show that (24.43) implies (24.20). The argument is similar to the standard consistency proof for nonlinear estimators. (24.43) states that  $C_n^*(K)$  converges uniformly in probability to  $R_n(K)$ . This implies that the minimizer of  $C_n^*(K)$  converges in probability to that of  $R_n(K)$ . Formally, since  $K_n^{\text{opt}}$  minimizes  $R_n(K)$

$$\begin{aligned} 0 &\leq \frac{R_n(\hat{K}_n) - R_n(K_n^{\text{opt}})}{R_n(\hat{K}_n)} \\ &= \frac{C_n^*(\hat{K}_n) - R_n(K_n^{\text{opt}})}{R_n(\hat{K}_n)} - \frac{C_n^*(\hat{K}_n) - R_n(\hat{K}_n)}{-R_n(\hat{K}_n)} \\ &\leq \frac{C_n^*(\hat{K}_n) - R_n(K_n^{\text{opt}})}{R_n(\hat{K}_n)} + o_p(1) \\ &\leq \frac{C_n^*(K_n^{\text{opt}}) - R_n(K_n^{\text{opt}})}{R_n(K_n^{\text{opt}})} + o_p(1) \\ &\leq o_p(1). \end{aligned}$$

The second inequality is (24.43). The following uses the facts that  $\hat{K}_n$  minimizes  $C_n^*(K)$  and  $K_n^{\text{opt}}$  minimizes  $R_n(K)$ . The final is (24.43). This is (24.20). ■

Before providing the proof of Theorem 24.11 we present two technical results.

**Theorem 24.23** The non-central chi-square density (5.3) obeys the recursive relationship

$$f_K(x, \lambda) = \frac{K}{x} f_{K+2}(x, \lambda) + \frac{\lambda}{x} f_{K+4}(x, \lambda).$$

The proof of Theorem 24.23 is a straightforward manipulation of the density function (5.3). The second technical result is from Bock (1975, Theorems A&B).

**Theorem 24.24** If  $\mathbf{x} \sim N(\boldsymbol{\theta}, \mathbf{I}_K)$  then for any function  $h(u)$

$$\mathbb{E}(\mathbf{x} h(\mathbf{x}' \mathbf{x})) = \boldsymbol{\theta} \mathbb{E}(h(Q_{K+2})) \quad (24.44)$$

$$\mathbb{E}(\mathbf{x}' \mathbf{x} h(\mathbf{x}' \mathbf{x})) = K \mathbb{E}(h(Q_{K+2})) + \lambda \mathbb{E}(h(Q_{K+4})) \quad (24.45)$$

where  $\lambda = \boldsymbol{\theta}' \boldsymbol{\theta}$  and  $Q_r \sim \chi_r^2(\lambda)$ , a non-central chi-square random variable with  $r$  degrees of freedom and non-centrality parameter  $\lambda$ .

**Proof of Theorem 24.24:** To show (24.44) we first show that for  $Z \sim N(\mu, 1)$  then for any function  $g(u)$

$$\mathbb{E}(Z g(Z^2)) = \mu \mathbb{E}(g(Q_3)). \quad (24.46)$$

Assume  $\mu > 0$ . Using the change-of-variables  $y = x^2$

$$\begin{aligned} \mathbb{E}(Z g(Z^2)) &= \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}} g(x^2) \exp\left(-\frac{1}{2}(x-\mu)^2\right) dx \\ &= \int_0^{\infty} \frac{y}{2\sqrt{2\pi}} e^{-(y+\mu^2)/2} \left(e^{\sqrt{y}\mu} - e^{-\sqrt{y}\mu}\right) g(y) dy. \end{aligned} \quad (24.47)$$

By expansion and Legendre's duplication formula

$$e^x - e^{-x} = 2 \sum_{i=0}^{\infty} \frac{x^{1+2i}}{(1+2i)!} = \sqrt{\pi} x \sum_{i=0}^{\infty} \frac{(x^2/2)^i}{2^i i! \Gamma(i+3/2)}.$$

Then (24.47) equals

$$\begin{aligned} \mu \int_0^{\infty} y e^{-(y+\mu^2)/2} \sum_{i=0}^{\infty} \frac{(\mu^2/2)^i y^{i+1/2}}{2^{3/2+i} i! \Gamma(i+3/2)} g(y) dy &= \mu \int_0^{\infty} y f_3(y, \mu^2) g(y) dy \\ &= \mu \mathbb{E}(g(Q_3)) \end{aligned}$$

where  $f_3(y, \lambda)$  is the non-central chi-square density (5.3) with 3 degrees of freedom. This is (24.46).

Take the  $j^{th}$  row of (24.44). Write  $\mathbf{x}' \mathbf{x} = x_j^2 + J$ , where  $x_j \sim N(\theta_j, 1)$  and  $J \sim \chi_{K-1}^2(\lambda - \theta_j^2)$  are independent. Setting  $g(u) = h(u + J)$  and using (24.47)

$$\begin{aligned} \mathbb{E}(x_j h(\mathbf{x}' \mathbf{x})) &= \mathbb{E}(x_j h(x_j^2 + J)) \\ &= \mathbb{E}(\mathbb{E}(x_j g(x_j^2) | J)) \\ &= \mathbb{E}(\theta_j \mathbb{E}(g(Q_3) | J)) \\ &= \theta_j \mathbb{E}(h(Q_3 + J)) \\ &= \theta_j \mathbb{E}(h(Q_{K+2})) \end{aligned}$$

which is (24.44). The final equality uses the fact that  $Q_3 + J \sim Q_{K+2}$ .

Observe that  $\mathbf{x}'\mathbf{x}$  has density  $f_K(x, \lambda)$ . Using Theorem 24.23

$$\begin{aligned}\mathbb{E}(\mathbf{x}'\mathbf{x}(\mathbf{x}'\mathbf{x})) &= \int_0^\infty x h(x) f_K(x, \lambda) dx \\ &= K \int_0^\infty h(x) f_{K+2}(x, \lambda) dx + \lambda \int_0^\infty h(x) f_{K+4}(x, \lambda) dx \\ &= K\mathbb{E}(h(Q_{K+2})) + \lambda\mathbb{E}(h(Q_{K+4}))\end{aligned}$$

which is (24.45). ■

**Proof of Theorem 24.11:** By the quadratic structure we can calculate that

$$\begin{aligned}\text{MSE}(\hat{\boldsymbol{\theta}}^*) &= \mathbb{E}\left(\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\mathbf{1}(\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}} \leq c)\right)' \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\mathbf{1}(\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}} \leq c)\right)\right) \\ &= \mathbb{E}\left((\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\right) - \mathbb{E}\left(\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}}\mathbf{1}(\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}} \leq c)\right) + 2\mathbb{E}\boldsymbol{\theta}'\hat{\boldsymbol{\theta}}\mathbf{1}(\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}} \leq c) \\ &= K - K\mathbb{E}(\mathbf{1}(Q_{K+2} \leq c)) - \lambda\mathbb{E}(\mathbf{1}(Q_{K+4} \leq c)) + 2\lambda\mathbb{E}(\mathbf{1}(Q_{K+2} \leq c)) \\ &= K + (2\lambda - K)F_{K+2}(c, \lambda) - \lambda F_{K+4}(c, \lambda).\end{aligned}$$

The third equality uses the two results from Theorem 24.24, setting  $h(u) = \mathbf{1}(u \leq c)$ . ■

Before providing the proof of Theorem 24.11 we present the following useful result.

**Theorem 24.25** If  $\phi(x)$  and  $\Phi(x)$  are the normal pdf and cdf functions, and  $b \geq 0$ , then

$$\int_{-\infty}^{\infty} \Phi(a + bx) \phi(x) dx = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$$

**Proof of Theorem 24.25:** Let  $X$  and  $Y$  be independent  $N(0, 1)$ . Note that  $Z = Y - bX \sim N(0, 1 + b^2)$ . Since the integral over  $\phi(x)$  can be written as an expectation over  $X$ , and  $\Phi(t) = \Pr(Y \leq t)$ ,

$$\begin{aligned}\int_{-\infty}^{\infty} \Phi(a + bx) \phi(x) dx &= \mathbb{E}(\Phi(a + bX)) \\ &= \Pr(Y \leq a + bX) \\ &= \Pr(Z \leq a) \\ &= \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)\end{aligned}$$

as stated. ■

**Proof of Theorem 24.12:** Without loss of generality we can set  $\mathbb{E}(y) = 0$  so  $\theta = -\beta\mu$ . Writing

$$T = \frac{\bar{y} - \beta\bar{x}\mathbf{1}(\bar{x}^2 > c) + \beta\mu}{\mathbf{1}(\bar{x}^2 \leq c) + \sqrt{1+\beta^2}\mathbf{1}(\bar{x}^2 > c)}$$

we have

$$\begin{aligned}
\Pr(T \leq t) &= \Pr\left(\bar{y} - \beta \bar{x} \mathbf{1}(\bar{x}^2 > c) + \beta \mu \leq t \mathbf{1}(\bar{x}^2 \leq c) + t \sqrt{1 + \beta^2} \mathbf{1}(\bar{x}^2 > c)\right) \\
&= \Pr\left(\bar{y} \leq (t - \beta \mu) \mathbf{1}(\bar{x}^2 \leq c) + \left(\beta(\bar{x} - \mu) + t \sqrt{1 + \beta^2}\right) \mathbf{1}(\bar{x}^2 > c)\right) \\
&= \mathbb{E}\left[\Phi\left((t - \beta \mu) \mathbf{1}(\bar{x}^2 \leq c) + \left(\beta(\bar{x} - \mu) + t \sqrt{1 + \beta^2}\right) \mathbf{1}(\bar{x}^2 > c)\right)\right] \\
&= \Phi(t - \beta \mu) \mathbb{E}[\mathbf{1}(\bar{x}^2 \leq c)] + \mathbb{E}\left[\Phi\left(\beta(\bar{x} - \mu) + t \sqrt{1 + \beta^2}\right) \mathbf{1}(\bar{x}^2 > c)\right]. \tag{24.48}
\end{aligned}$$

The first term in (24.48) is

$$\Phi(t - \beta \mu) \Pr(\bar{x}^2 \leq c) = \Phi(t - \beta \mu) (\Phi(\sqrt{c} - \mu) - \Phi(-\sqrt{c} - \mu)).$$

The second term in (24.48) is

$$\begin{aligned}
&\int_{x^2 > c} \Phi\left(\beta(r - \mu) + t \sqrt{1 + \beta^2}\right) \phi(r - \mu) dr \\
&= \int_{(x+\mu)^2 > c} \Phi\left(\beta s + t \sqrt{1 + \beta^2}\right) \phi(s) ds \\
&= \int_{-\infty}^{\infty} \Phi\left(\beta s + t \sqrt{1 + \beta^2}\right) \phi(s) ds - \int_{-\sqrt{c}-\mu}^{\sqrt{c}-\mu} \Phi\left(\beta s + t \sqrt{1 + \beta^2}\right) \phi(s) ds \\
&= \Phi(t) - \int_{-\sqrt{c}-\mu}^{\sqrt{c}-\mu} \Phi\left(\beta s + t \sqrt{1 + \beta^2}\right) \phi(s) ds
\end{aligned}$$

where the last equality uses Theorem 24.25. Adding we obtain the result. ■

**Proof of Theorem 24.18:** It will be convenient to denote  $Q_K = \hat{\boldsymbol{\theta}}' \mathbf{V}^{-1} \hat{\boldsymbol{\theta}} \sim \chi^2_K(\lambda)$  and  $\mathbf{1}_K = \mathbf{1}(Q_K < K - 2)$ . The estimator can be written as

$$\tilde{\boldsymbol{\theta}}^+ = \hat{\boldsymbol{\theta}} - \frac{(K-2)}{Q_K} \hat{\boldsymbol{\theta}} - \mathbf{h}(\hat{\boldsymbol{\theta}})$$

where

$$\mathbf{h}(\mathbf{x}) = \mathbf{x} \left(1 - \frac{(K-2)}{\mathbf{x}' \mathbf{V}^{-1} \mathbf{x}}\right) \mathbf{1}(\mathbf{x}' \mathbf{V}^{-1} \mathbf{x} < K-2).$$

Observe that

$$\begin{aligned}
\text{tr}\left(\frac{\partial}{\partial \mathbf{x}} \mathbf{h}(\mathbf{x})'\right) &= \text{tr}\left(\mathbf{I}_K \left(1 - \frac{K-2}{\mathbf{x}' \mathbf{V}^{-1} \mathbf{x}}\right) - 2 \left(\frac{K-2}{(\mathbf{x}' \mathbf{V}^{-1} \mathbf{x})^2}\right) \mathbf{V}^{-1} \mathbf{x} \mathbf{x}'\right) \mathbf{1}(\mathbf{x}' \mathbf{V}^{-1} \mathbf{x} < K-2) \\
&= \left(K - \frac{(K-2)^2}{\mathbf{x}' \mathbf{V}^{-1} \mathbf{x}}\right) \mathbf{1}(\mathbf{x}' \mathbf{V}^{-1} \mathbf{x} < K-2).
\end{aligned}$$

Thus by Stein's Lemma

$$\mathbb{E}(\mathbf{h}(\hat{\boldsymbol{\theta}})' \mathbf{V}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})) = \mathbb{E}\left(\left(K - \frac{(K-2)^2}{Q_K}\right) \mathbf{1}_K\right).$$

Using these expressions, the definition of  $\tilde{\boldsymbol{\theta}}$ , and expanding the quadratic,

$$\begin{aligned}
\text{wmse}(\tilde{\boldsymbol{\theta}}^+) &= \mathbb{E}(\tilde{\boldsymbol{\theta}}^+ - \boldsymbol{\theta})' \mathbf{V}^{-1} (\tilde{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}) + \mathbb{E}(\mathbf{h}(\tilde{\boldsymbol{\theta}})' \mathbf{h}(\tilde{\boldsymbol{\theta}})) \\
&\quad + \frac{2(K-2)}{Q_K} \hat{\boldsymbol{\theta}}' \mathbf{V}^{-1} \mathbf{h}(\hat{\boldsymbol{\theta}}) - 2\mathbb{E}(\mathbf{h}(\hat{\boldsymbol{\theta}})' \mathbf{V}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})) \\
&= \text{wmse}(\tilde{\boldsymbol{\theta}}) - \mathbb{E}\left(\left(2K - Q_K - \frac{(K-2)^2}{Q_K}\right) \mathbf{1}_K\right) \\
&= \text{wmse}(\tilde{\boldsymbol{\theta}}) - 2KF_K(K-2, \lambda K) + KF_{K+2}(K-2, \lambda) \\
&\quad + \lambda F_{K+4}(K-2, \lambda) + (K-2)^2 J_K(K-2, \lambda)
\end{aligned} \tag{24.49}$$

using Theorem 24.17 and (24.45). This is (24.32).

To show (24.31), examine (24.49). It is sufficient to show that the second term is strictly positive. Using the region of integration, Theorem 24.17, and then the fact that the distribution of  $Q$  is monotone in the degrees of freedom

$$\begin{aligned}\mathbb{E} \left( \left( 2K - Q_K - \frac{(K-2)^2}{Q_K} \right) \mathbf{1}_K \right) &\geq \mathbb{E} \left( \left( K + 2 - \frac{(K-2)^2}{Q_K} \right) \mathbf{1}_K \right) \\&= (K+2)F_K(K-2, \lambda) - (K-2)^2 \sum_{i=0}^{\infty} \frac{e^{-\lambda/2}}{i!} \frac{(\lambda/2)^i}{K+2i-2} F_{K+2i-2}(K-2, \lambda) \\&\geq (K+2)F_K(K-2, \lambda) - (K-2)^2 \sum_{i=0}^{\infty} \frac{e^{-\lambda/2}}{i!} \frac{(\lambda/2)^i}{K-2} F_K(K-2, \lambda) \\&= 4F_K(K-2, \lambda) > 0\end{aligned}$$

as claimed. ■

**Proof of Theorem 24.21:** In the following calculations we condition on the regressor matrix  $\mathbf{X}$  to simplify notation. In the homoskedastic regression model the bias variance of the ridge estimator is

$$\text{bias}(\hat{\boldsymbol{\beta}}_{\text{ridge}}) = \left( (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1} \mathbf{X}'\mathbf{X} - \mathbf{I}_k \right) \boldsymbol{\beta} = -\lambda (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1} \boldsymbol{\beta}.$$

Its variance matrix is

$$\text{var}(\hat{\boldsymbol{\beta}}_{\text{ridge}}) = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1} (\sigma^2 \mathbf{X}'\mathbf{X}) (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1}.$$

Hence

$$\begin{aligned}\text{mse}(\hat{\boldsymbol{\beta}}_{\text{ridge}}) &= \text{var}(\hat{\boldsymbol{\beta}}_{\text{ridge}}) + \text{bias}(\hat{\boldsymbol{\beta}}_{\text{ridge}}) \text{bias}(\hat{\boldsymbol{\beta}}_{\text{ridge}})' \\&= (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1} (\sigma^2 \mathbf{X}'\mathbf{X}) (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1} + \lambda^2 (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1} \boldsymbol{\beta} \boldsymbol{\beta}' (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1} \\&= (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1} (\sigma^2 \mathbf{X}'\mathbf{X} + \lambda^2 \boldsymbol{\beta} \boldsymbol{\beta}') (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1}.\end{aligned}$$

The MSE of the least squares estimator is

$$\begin{aligned}\text{mse}(\hat{\boldsymbol{\beta}}_{\text{ols}}) &= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \\&= (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1} (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k) \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k) (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1} \\&= (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1} \left( \sigma^2 \mathbf{X}'\mathbf{X} + 2\sigma^2 \lambda \mathbf{I}_k + \sigma^2 \lambda^2 (\mathbf{X}'\mathbf{X})^{-1} \right) (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1}.\end{aligned}$$

Hence

$$\begin{aligned}\text{mse}(\hat{\boldsymbol{\beta}}_{\text{ols}}) - \text{mse}(\hat{\boldsymbol{\beta}}_{\text{ridge}}) &= (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1} \left( 2\sigma^2 \lambda \mathbf{I}_k + \sigma^2 \lambda^2 (\mathbf{X}'\mathbf{X})^{-1} - \lambda^2 \boldsymbol{\beta} \boldsymbol{\beta}' \right) (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1} \\&\geq (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1} (2\sigma^2 \lambda \mathbf{I}_k - \lambda^2 \boldsymbol{\beta} \boldsymbol{\beta}') (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1}.\end{aligned}$$

The final line is positive definite if and only if  $2\sigma^2 \mathbf{I}_k - \lambda \boldsymbol{\beta} \boldsymbol{\beta}' > 0$ , which holds since for any  $\boldsymbol{\alpha}' \boldsymbol{\alpha} = 1$ ,

$$\boldsymbol{\alpha}' (2\sigma^2 \mathbf{I}_k - \lambda \boldsymbol{\beta} \boldsymbol{\beta}') \boldsymbol{\alpha} = 2\sigma^2 - \lambda (\boldsymbol{\alpha}' \boldsymbol{\beta})^2 \geq 2\sigma^2 - \lambda \boldsymbol{\beta}' \boldsymbol{\beta} > 0$$

the final inequality by the assumption on  $\lambda$ . ■

## Exercises

**Exercise 24.1** Verify equations (24.1)-(24.2).

**Exercise 24.2** Prove Theorem 24.1. As part of your derivation, verify equations (24.3) and (24.4).

**Exercise 24.3** Find the Mallows criterion for the weighted least squares estimator of a linear regression  $y_i = \mathbf{x}'_i \boldsymbol{\beta} + e_i$ ;  $y_i = \mathbf{x}'_i \boldsymbol{\beta} + e_i$  (assume conditional homoskedasticity).

**Exercise 24.4** Backward Stepwise Regression. Verify the claim that for the case of AIC selection, step (b) of the algorithm can be implemented by calculating the classical (homoskedastic) t-ratio for each active regressor and find the regressor with the smallest absolute t-ratio.

Hint: Use the relationship between likelihood ratio and F statistics, and the equality between F and Wald statistics, to show that for tests on one coefficient, the smallest change in the AIC is identical to identifying the smallest squared t statistic.

**Exercise 24.5** Forward Stepwise Regression. Verify the claim that for the case of AIC selection, step (b) of the algorithm can be implemented by identifying the regressor in the inactive set with the greatest absolute correlation with the residual from step (a).

Hint: This is challenging. First show that the goal is to find the regressor which will most decrease  $\text{SSE} = \hat{\mathbf{e}}' \hat{\mathbf{e}} = \|\hat{\mathbf{e}}\|^2$ . Use a geometric argument to show that the regressor most parallel to  $\hat{\mathbf{e}}$  will most decreases  $\|\hat{\mathbf{e}}\|$ . Show that this regressor has the greatest absolute correlation with  $\hat{\mathbf{e}}$ .

**Exercise 24.6** An economist estimates several models, and reports their favorite specification, stating that “the other specifications had insignificant coefficients”. How should we interpret the reported parameter estimates and t-ratios?

**Exercise 24.7** Verify Theorem 24.13, including (24.24), (24.25), and (24.26).

**Exercise 24.8** Under the assumptions of Theorem 24.13, show that  $\hat{\lambda} = \hat{\boldsymbol{\theta}}' V^{-1} \hat{\boldsymbol{\theta}} - K$  is an unbiased estimator of  $\lambda = \boldsymbol{\theta}' V^{-1} \boldsymbol{\theta}$

**Exercise 24.9** Verify Theorem 24.14.

**Exercise 24.10** Generalize the proof of Theorem 24.16 to allow for general variance matrices  $V$ .

**Exercise 24.11** Using the CPS dataset perform an analysis similar to that presented in Section 24.17, but instead use the sub-sample of Hispanic women. This sample has 3003 observations. Which models are selected by BIC, AIC, CV and FIC? (The precise information criteria you examine may be limited depending on your software.) How do you interpret the results? Which model/estimate would you select as your preferred choice?

**Exercise 24.12** Prove Theorem 24.19 for the simpler case of the unadjusted (not positive part) Stein estimator  $\tilde{\boldsymbol{\theta}}$ ,  $V = I_K$  and  $\mathbf{r} = \mathbf{0}$ .

Extra challenge: Show under these assumptions that

$$\begin{aligned}\text{wmse}(\tilde{\boldsymbol{\theta}}) &= K - (q-2)^2 J_q(\lambda_R) \\ \lambda_R &= \boldsymbol{\theta}' \mathbf{R} (\mathbf{R}' \mathbf{R})^{-1} \mathbf{R}' \boldsymbol{\theta}\end{aligned}$$

**Exercise 24.13** Suppose you have two unbiased estimators  $\hat{\boldsymbol{\theta}}_1$  and  $\hat{\boldsymbol{\theta}}_2$  of a parameter vector  $\hat{\boldsymbol{\theta}}$  with covariance matrices  $V_1$  and  $V_2$ . Take the goal of minimizing the unweighted mean squared error, e.g.  $\text{tr } V_1$  for  $\hat{\boldsymbol{\theta}}_1$ . Assume that  $\hat{\boldsymbol{\theta}}_1$  and  $\hat{\boldsymbol{\theta}}_2$  are uncorrelated. [This is important.]

- (a) Show that the optimal weighted average estimator equals

$$\frac{\frac{1}{\text{tr } V_1} \hat{\boldsymbol{\theta}}_1 + \frac{1}{\text{tr } V_2} \hat{\boldsymbol{\theta}}_2}{\frac{1}{\text{tr } V_1} + \frac{1}{\text{tr } V_2}}.$$

- (b) Generalize to the case of  $M$  unbiased uncorrelated estimators.

- (c) Interpret the formulae.

**Exercise 24.14** You estimate  $M$  linear regressions  $y_i = \mathbf{x}'_{mi} \boldsymbol{\beta}_m + e_{mi}$  by least squares. Let  $\hat{y}_{mi} = \mathbf{x}'_{mi} \hat{\boldsymbol{\beta}}_m$  be the fitted values.

- (a) Show that the Mallows averaging criterion is the same as

$$\sum_{i=1}^n (y_i - w_1 \hat{y}_{1i} - w_2 \hat{y}_{2i} - \cdots - w_M \hat{y}_{Mi})^2 + 2\sigma^2 \sum_{m=1}^M w_m k_m.$$

- (b) Assume the models are nested with  $M$  the largest model. If the previous criterion were minimized over  $\mathbf{w}$  in the probability simplex but the penalty was omitted, what would be the solution? (What would be the minimizing weight vector?)

**Exercise 24.15** You estimate  $M$  linear regressions  $y_i = \mathbf{x}'_{mi} \boldsymbol{\beta}_m + e_{mi}$  by least squares. Let  $\tilde{y}_{mi} = \mathbf{x}'_{mi} \hat{\boldsymbol{\beta}}_{m(-i)}$  be the predicted values from the leave-one-out regressions. Show that the JMA criterion is the same as

$$\sum_{i=1}^n (y_i - w_1 \tilde{y}_{1i} - w_2 \tilde{y}_{2i} - \cdots - w_M \tilde{y}_{Mi})^2.$$

# **Appendices**

# Appendix A

## Matrix Algebra

### A.1 Notation

A **scalar**  $a$  is a single number.

A **vector**  $\mathbf{a}$  is a  $k \times 1$  list of numbers, typically arranged in a column. We write this as

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix}$$

Equivalently, a vector  $\mathbf{a}$  is an element of Euclidean  $k$  space, written as  $\mathbf{a} \in \mathbb{R}^k$ . If  $k = 1$  then  $\mathbf{a}$  is a scalar.

A **matrix**  $\mathbf{A}$  is a  $k \times r$  rectangular array of numbers, written as

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kr} \end{bmatrix}$$

By convention  $a_{ij}$  refers to the element in the  $i^{th}$  row and  $j^{th}$  column of  $\mathbf{A}$ . If  $r = 1$  then  $\mathbf{A}$  is a column vector. If  $k = 1$  then  $\mathbf{A}$  is a row vector. If  $r = k = 1$ , then  $\mathbf{A}$  is a scalar.

A standard convention (which we will follow in this text whenever possible) is to denote scalars by lower-case italics ( $a$ ), vectors by lower-case bold italics ( $\mathbf{a}$ ), and matrices by upper-case bold italics ( $\mathbf{A}$ ). Sometimes a matrix  $\mathbf{A}$  is denoted by the symbol  $(a_{ij})$ .

A matrix can be written as a set of column vectors or as a set of row vectors. That is,

$$\mathbf{A} = [ \mathbf{a}_1 \ \mathbf{a}_2 \ \cdots \ \mathbf{a}_r ] = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_k \end{bmatrix}$$

where

$$\mathbf{a}_i = \begin{bmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{ki} \end{bmatrix}$$

are column vectors and

$$\mathbf{a}_j = [ a_{j1} \ a_{j2} \ \cdots \ a_{jr} ]$$

are row vectors.

The **transpose** of a matrix  $\mathbf{A}$ , denoted  $\mathbf{A}'$ ,  $\mathbf{A}^\top$ , or  $\mathbf{A}^t$ , is obtained by flipping the matrix on its diagonal. (In most of the econometrics literature, and this textbook, we use  $\mathbf{A}'$ , but in the mathematics literature  $\mathbf{A}^\top$  is the convention.) Thus

$$\mathbf{A}' = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{k1} \\ a_{12} & a_{22} & \cdots & a_{k2} \\ \vdots & \vdots & & \vdots \\ a_{1r} & a_{2r} & \cdots & a_{kr} \end{bmatrix}$$

Alternatively, letting  $\mathbf{B} = \mathbf{A}'$ , then  $b_{ij} = a_{ji}$ . Note that if  $\mathbf{A}$  is  $k \times r$ , then  $\mathbf{A}'$  is  $r \times k$ . If  $\mathbf{a}$  is a  $k \times 1$  vector, then  $\mathbf{a}'$  is a  $1 \times k$  row vector.

A matrix is **square** if  $k = r$ . A square matrix is **symmetric** if  $\mathbf{A} = \mathbf{A}'$ , which requires  $a_{ij} = a_{ji}$ . A square matrix is **diagonal** if the off-diagonal elements are all zero, so that  $a_{ij} = 0$  if  $i \neq j$ . A square matrix is **upper (lower) diagonal** if all elements below (above) the diagonal equal zero.

An important diagonal matrix is the **identity matrix**, which has ones on the diagonal. The  $k \times k$  identity matrix is denoted as

$$\mathbf{I}_k = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

A **partitioned matrix** takes the form

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1r} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \cdots & \mathbf{A}_{2r} \\ \vdots & \vdots & & \vdots \\ \mathbf{A}_{k1} & \mathbf{A}_{k2} & \cdots & \mathbf{A}_{kr} \end{bmatrix}$$

where the  $\mathbf{A}_{ij}$  denote matrices, vectors and/or scalars.

## A.2 Complex Matrices\*

Scalars, vectors and matrices may contain real or complex numbers as entries. (However, most econometric applications exclusively use real matrices.) If all elements of a vector  $\mathbf{x}$  are real we say that  $\mathbf{x}$  is a real vector, and similarly for matrices.

Recall that a complex number can be written as  $x = a + bi$  where  $i = \sqrt{-1}$  and  $a$  and  $b$  are real numbers. Similarly a vector with complex elements can be written as  $\mathbf{x} = \mathbf{a} + \mathbf{bi}$  where  $\mathbf{a}$  and  $\mathbf{b}$  are real vectors, and a matrix with complex elements can be written as  $\mathbf{X} = \mathbf{A} + \mathbf{Bi}$  where  $\mathbf{A}$  and  $\mathbf{B}$  are real matrices.

Recall that the complex conjugate of  $x = a + bi$  is  $x^* = a - bi$ . For matrices, the analogous concept is the conjugate transpose. The conjugate transpose of  $\mathbf{X} = \mathbf{A} + \mathbf{Bi}$  is  $\mathbf{X}^* = \mathbf{A}' - \mathbf{Bi}$ . It is obtained by taking the transpose and taking the complex conjugate of each element.

## A.3 Matrix Addition

If the matrices  $\mathbf{A} = (a_{ij})$  and  $\mathbf{B} = (b_{ij})$  are of the same order, we define the sum

$$\mathbf{A} + \mathbf{B} = (a_{ij} + b_{ij}).$$

Matrix addition follows the commutative and associative laws:

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$$

$$\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}.$$

## A.4 Matrix Multiplication

If  $\mathbf{A}$  is  $k \times r$  and  $c$  is real, we define their product as

$$\mathbf{Ac} = c\mathbf{A} = (a_{ij}c).$$

If  $\mathbf{a}$  and  $\mathbf{b}$  are both  $k \times 1$ , then their inner product is

$$\mathbf{a}'\mathbf{b} = a_1 b_1 + a_2 b_2 + \cdots + a_k b_k = \sum_{j=1}^k a_j b_j.$$

Note that  $\mathbf{a}'\mathbf{b} = \mathbf{b}'\mathbf{a}$ . We say that two vectors  $\mathbf{a}$  and  $\mathbf{b}$  are **orthogonal** if  $\mathbf{a}'\mathbf{b} = 0$ .

If  $\mathbf{A}$  is  $k \times r$  and  $\mathbf{B}$  is  $r \times s$ , so that the number of columns of  $\mathbf{A}$  equals the number of rows of  $\mathbf{B}$ , we say that  $\mathbf{A}$  and  $\mathbf{B}$  are **conformable**. In this event the matrix product  $\mathbf{AB}$  is defined. Writing  $\mathbf{A}$  as a set of row vectors and  $\mathbf{B}$  as a set of column vectors (each of length  $r$ ), then the matrix product is defined as

$$\begin{aligned} \mathbf{AB} &= \left[ \begin{array}{c} \mathbf{a}'_1 \\ \mathbf{a}'_2 \\ \vdots \\ \mathbf{a}'_k \end{array} \right] \left[ \begin{array}{cccc} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_s \end{array} \right] \\ &= \left[ \begin{array}{cccc} \mathbf{a}'_1 \mathbf{b}_1 & \mathbf{a}'_1 \mathbf{b}_2 & \cdots & \mathbf{a}'_1 \mathbf{b}_s \\ \mathbf{a}'_2 \mathbf{b}_1 & \mathbf{a}'_2 \mathbf{b}_2 & \cdots & \mathbf{a}'_2 \mathbf{b}_s \\ \vdots & \vdots & & \vdots \\ \mathbf{a}'_k \mathbf{b}_1 & \mathbf{a}'_k \mathbf{b}_2 & \cdots & \mathbf{a}'_k \mathbf{b}_s \end{array} \right]. \end{aligned}$$

Matrix multiplication is not commutative: in general  $\mathbf{AB} \neq \mathbf{BA}$ . However, it is associative and distributive:

$$\begin{aligned} \mathbf{A}(\mathbf{BC}) &= (\mathbf{AB})\mathbf{C} \\ \mathbf{A}(\mathbf{B} + \mathbf{C}) &= \mathbf{AB} + \mathbf{AC}. \end{aligned}$$

An alternative way to write the matrix product is to use matrix partitions. For example,

$$\begin{aligned} \mathbf{AB} &= \left[ \begin{array}{cc} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{array} \right] \left[ \begin{array}{cc} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{array} \right] \\ &= \left[ \begin{array}{cc} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{array} \right]. \end{aligned}$$

As another example,

$$\begin{aligned} \mathbf{AB} &= \left[ \begin{array}{cccc} \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_r \end{array} \right] \left[ \begin{array}{c} \mathbf{B}_1 \\ \mathbf{B}_2 \\ \vdots \\ \mathbf{B}_r \end{array} \right] \\ &= \mathbf{A}_1\mathbf{B}_1 + \mathbf{A}_2\mathbf{B}_2 + \cdots + \mathbf{A}_r\mathbf{B}_r \\ &= \sum_{j=1}^r \mathbf{A}_j\mathbf{B}_j. \end{aligned}$$

An important property of the identity matrix is that if  $\mathbf{A}$  is  $k \times r$ , then  $\mathbf{AI}_r = \mathbf{A}$  and  $\mathbf{I}_k\mathbf{A} = \mathbf{A}$ .

We say two matrices  $\mathbf{A}$  and  $\mathbf{B}$  are **orthogonal** if  $\mathbf{A}'\mathbf{B} = \mathbf{0}$ . This means that all columns of  $\mathbf{A}$  are orthogonal with all columns of  $\mathbf{B}$ .

The  $k \times r$  matrix  $\mathbf{H}$ ,  $r \leq k$ , is called **orthonormal** if  $\mathbf{H}'\mathbf{H} = \mathbf{I}_r$ . This means that the columns of  $\mathbf{H}$  are mutually orthogonal, and each column is normalized to have unit length.

## A.5 Trace

The **trace** of a  $k \times k$  square matrix  $\mathbf{A}$  is the sum of its diagonal elements

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^k a_{ii}.$$

Some straightforward properties for square matrices  $\mathbf{A}$  and  $\mathbf{B}$  and real  $c$  are

$$\begin{aligned}\text{tr}(c\mathbf{A}) &= c \text{tr}(\mathbf{A}) \\ \text{tr}(\mathbf{A}') &= \text{tr}(\mathbf{A}) \\ \text{tr}(\mathbf{A} + \mathbf{B}) &= \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B}) \\ \text{tr}(\mathbf{I}_k) &= k.\end{aligned}$$

Also, for  $k \times r$   $\mathbf{A}$  and  $r \times k$   $\mathbf{B}$  we have

$$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA}). \quad (\text{A.1})$$

Indeed,

$$\begin{aligned}\text{tr}(\mathbf{AB}) &= \text{tr} \left[ \begin{array}{cccc} \mathbf{a}'_1 \mathbf{b}_1 & \mathbf{a}'_1 \mathbf{b}_2 & \cdots & \mathbf{a}'_1 \mathbf{b}_k \\ \mathbf{a}'_2 \mathbf{b}_1 & \mathbf{a}'_2 \mathbf{b}_2 & \cdots & \mathbf{a}'_2 \mathbf{b}_k \\ \vdots & \vdots & & \vdots \\ \mathbf{a}'_k \mathbf{b}_1 & \mathbf{a}'_k \mathbf{b}_2 & \cdots & \mathbf{a}'_k \mathbf{b}_k \end{array} \right] \\ &= \sum_{i=1}^k \mathbf{a}'_i \mathbf{b}_i \\ &= \sum_{i=1}^k \mathbf{b}'_i \mathbf{a}_i \\ &= \text{tr}(\mathbf{BA}).\end{aligned}$$

## A.6 Rank and Inverse

The rank of the  $k \times r$  matrix ( $r \leq k$ )

$$\mathbf{A} = [\mathbf{a}_1 \ \mathbf{a}_2 \ \cdots \ \mathbf{a}_r]$$

is the number of linearly independent columns  $\mathbf{a}_j$ , and is written as  $\text{rank}(\mathbf{A})$ . We say that  $\mathbf{A}$  has full rank if  $\text{rank}(\mathbf{A}) = r$ .

A square  $k \times k$  matrix  $\mathbf{A}$  is said to be **nonsingular** if it has full rank, e.g.  $\text{rank}(\mathbf{A}) = k$ . This means that there is no  $k \times 1$   $\mathbf{c} \neq \mathbf{0}$  such that  $\mathbf{Ac} = \mathbf{0}$ .

If a square  $k \times k$  matrix  $\mathbf{A}$  is nonsingular then there exists a unique matrix  $k \times k$  matrix  $\mathbf{A}^{-1}$  called the **inverse** of  $\mathbf{A}$  which satisfies

$$\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_k.$$

For non-singular  $\mathbf{A}$  and  $\mathbf{C}$ , some important properties include

$$\begin{aligned}\mathbf{AA}^{-1} &= \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_k \\ (\mathbf{A}^{-1})' &= (\mathbf{A}')^{-1} \\ (\mathbf{AC})^{-1} &= \mathbf{C}^{-1}\mathbf{A}^{-1} \\ (\mathbf{A} + \mathbf{C})^{-1} &= \mathbf{A}^{-1}(\mathbf{A}^{-1} + \mathbf{C}^{-1})^{-1}\mathbf{C}^{-1} \\ \mathbf{A}^{-1} - (\mathbf{A} + \mathbf{C})^{-1} &= \mathbf{A}^{-1}(\mathbf{A}^{-1} + \mathbf{C}^{-1})^{-1}\mathbf{A}^{-1}.\end{aligned}$$

If a  $k \times k$  matrix  $\mathbf{H}$  is orthonormal (so that  $\mathbf{H}'\mathbf{H} = \mathbf{I}_k$ ), then  $\mathbf{H}$  is nonsingular and  $\mathbf{H}^{-1} = \mathbf{H}'$ . Furthermore,  $\mathbf{H}\mathbf{H}' = \mathbf{I}_k$  and  $\mathbf{H}'^{-1} = \mathbf{H}$ .

Another useful result for non-singular  $\mathbf{A}$  is known as the **Woodbury matrix identity**

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\mathbf{C}(\mathbf{C} + \mathbf{CDA}^{-1}\mathbf{B}\mathbf{C})^{-1}\mathbf{C}\mathbf{D}\mathbf{A}^{-1}.$$

In particular, for  $\mathbf{C} = 1$ ,  $\mathbf{B} = \mathbf{b}$  and  $\mathbf{D} = \mathbf{b}'$  for vector  $\mathbf{b}$  we find what is known as the **Sherman–Morrison formula**

$$(\mathbf{A} + \mathbf{bb}')^{-1} = \mathbf{A}^{-1} - (1 + \mathbf{b}'\mathbf{A}^{-1}\mathbf{b})^{-1}\mathbf{A}^{-1}\mathbf{bb}'\mathbf{A}^{-1}.$$

and similarly using  $\mathbf{C} = -1$

$$(\mathbf{A} - \mathbf{bb}')^{-1} = \mathbf{A}^{-1} + (1 - \mathbf{b}'\mathbf{A}^{-1}\mathbf{b})^{-1}\mathbf{A}^{-1}\mathbf{bb}'\mathbf{A}^{-1}. \quad (\text{A.2})$$

The following fact about inverting partitioned matrices is quite useful.

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{A}_{22}^{-1} \end{bmatrix} \quad (\text{A.3})$$

where  $\mathbf{A}_{11.2} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$  and  $\mathbf{A}_{22.1} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ . There are alternative algebraic representations for the components. For example, using the Woodbury matrix identity you can show the following alternative expressions

$$\begin{aligned} \mathbf{A}^{11} &= \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22.1}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} \\ \mathbf{A}^{22} &= \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11.2}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{A}^{12} &= -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22.1}^{-1} \\ \mathbf{A}^{21} &= -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11.2}^{-1}. \end{aligned}$$

Even if a matrix  $\mathbf{A}$  does not possess an inverse, we can still define the **Moore–Penrose generalized inverse  $\mathbf{A}^-$**  as the matrix which satisfies

$$\begin{aligned} \mathbf{AA}^- &= \mathbf{A} \\ \mathbf{A}^-\mathbf{AA}^- &= \mathbf{A}^- \\ \mathbf{AA}^- &\text{ is symmetric} \\ \mathbf{A}^-\mathbf{A} &\text{ is symmetric.} \end{aligned}$$

For any matrix  $\mathbf{A}$ , the Moore–Penrose generalized inverse  $\mathbf{A}^-$  exists and is unique.

For example, if

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

and  $\mathbf{A}_{11}^{-1}$  exists then

$$\mathbf{A}^- = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

## A.7 Orthogonal and Orthonormal Matrices

We say that two  $k \times 1$  vectors  $\mathbf{h}_1$  and  $\mathbf{h}_2$  are **orthogonal** if  $\mathbf{h}_1'\mathbf{h}_2 = 0$ . This means that they are perpendicular.

We say that a  $k \times 1$  vector  $\mathbf{h}$  is a **unit vector** if  $\mathbf{h}'\mathbf{h} = 1$ . This means that it has unit length in  $\mathbb{R}^k$ .

We say that two  $k \times 1$  vectors  $\mathbf{h}_1$  and  $\mathbf{h}_2$  are **orthonormal** if they are orthogonal unit vectors.

We say that the  $k \times m_1$  and  $k \times m_2$  matrices  $\mathbf{H}_1$  and  $\mathbf{H}_2$  are **orthogonal** if  $\mathbf{H}_1'\mathbf{H}_2 = \mathbf{0}$ .

We say that the  $k \times m$  ( $k \geq m$ ) matrix  $\mathbf{H}$  is **orthonormal** if  $\mathbf{H}'\mathbf{H} = \mathbf{I}_m$ . This means that the columns of  $\mathbf{H}$  are orthonormal. Some call  $\mathbf{H}$  an orthogonal matrix.

Typically an orthonormal matrix is written as  $\mathbf{H}$ .

If  $\mathbf{H}$  is a  $k \times k$  orthogonal matrix then it has full rank  $k$ ,  $\mathbf{H}'\mathbf{H} = \mathbf{I}_k$ ,  $\mathbf{HH}' = \mathbf{I}_k$ , and  $\mathbf{H}^{-1} = \mathbf{H}'$ .

## A.8 Determinant

The **determinant** is a measure of the volume of a square matrix. It is written as  $\det \mathbf{A}$  or  $|\mathbf{A}|$ .

While the determinant is widely used, its precise definition is rarely needed. However, we present the definition here for completeness. Let  $\mathbf{A} = (a_{ij})$  be a  $k \times k$  matrix. Let  $\pi = (j_1, \dots, j_k)$  denote a permutation of  $(1, \dots, k)$ . There are  $k!$  such permutations. There is a unique count of the number of inversions of the indices of such permutations (relative to the natural order  $(1, \dots, k)$ ), and let  $\varepsilon_\pi = +1$  if this count is even and  $\varepsilon_\pi = -1$  if the count is odd. Then the determinant of  $\mathbf{A}$  is defined as

$$\det \mathbf{A} = \sum_{\pi} \varepsilon_\pi a_{1j_1} a_{2j_2} \cdots a_{kj_k}.$$

For example, if  $\mathbf{A}$  is  $2 \times 2$ , then the two permutations of  $(1, 2)$  are  $(1, 2)$  and  $(2, 1)$ , for which  $\varepsilon_{(1,2)} = 1$  and  $\varepsilon_{(2,1)} = -1$ . Thus

$$\begin{aligned} \det \mathbf{A} &= \varepsilon_{(1,2)} a_{11} a_{22} + \varepsilon_{(2,1)} a_{21} a_{12} \\ &= a_{11} a_{22} - a_{12} a_{21}. \end{aligned}$$

For a square matrix  $\mathbf{A}$ , the **minor**  $M_{ij}$  of the  $ij^{th}$  element  $a_{ij}$  is the determinant of the matrix obtained by removing the  $i^{th}$  row and  $j^{th}$  column of  $\mathbf{A}$ . The **cofactor** of the  $ij^{th}$  element is  $C_{ij} = (-1)^{i+j} M_{ij}$ . An important representation known as Laplace's expansion relates the determinant of  $\mathbf{A}$  to its cofactors:

$$\det \mathbf{A} = \sum_{j=1}^k a_{ij} C_{ij}.$$

This holds for all  $i = 1, \dots, k$ . This is often presented as a method for computation of a determinant.

### Theorem A.1 Properties of the determinant

1.  $\det(\mathbf{A}) = \det(\mathbf{A}')$
2.  $\det(c\mathbf{A}) = c^k \det \mathbf{A}$
3.  $\det(\mathbf{AB}) = \det(\mathbf{BA}) = (\det \mathbf{A})(\det \mathbf{B})$
4.  $\det(\mathbf{A}^{-1}) = (\det \mathbf{A})^{-1}$
5.  $\det \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = (\det \mathbf{D}) \det(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})$  if  $\det \mathbf{D} \neq 0$
6.  $\det \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} = \det(\mathbf{A})(\det \mathbf{D})$  and  $\det \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \det(\mathbf{A})(\det \mathbf{D})$
7. If  $\mathbf{A}$  is  $p \times q$  and  $\mathbf{B}$  is  $q \times p$  then  $\det(\mathbf{I}_p + \mathbf{AB}) = \det(\mathbf{I}_q + \mathbf{BA})$
8. If  $\mathbf{A}$  and  $\mathbf{D}$  are invertible then  $\det(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C}) = \frac{\det(\mathbf{A})}{\det(\mathbf{D})} \det(\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B})$
9.  $\det \mathbf{A} \neq 0$  if and only if  $\mathbf{A}$  is nonsingular
10. If  $\mathbf{A}$  is triangular (upper or lower), then  $\det \mathbf{A} = \prod_{i=1}^k a_{ii}$
11. If  $\mathbf{A}$  is orthonormal, then  $\det \mathbf{A} = \pm 1$
12.  $\mathbf{A}^{-1} = (\det \mathbf{A})^{-1} \mathbf{C}$  where  $\mathbf{C} = (C_{ij})$  is the matrix of cofactors

## A.9 Eigenvalues

The **characteristic equation** of a  $k \times k$  square matrix  $\mathbf{A}$  is

$$\det(\lambda \mathbf{I}_k - \mathbf{A}) = 0.$$

The left side is a polynomial of degree  $k$  in  $\lambda$  so it has exactly  $k$  roots, which are not necessarily distinct and may be real or complex. They are called the **latent roots**, **characteristic roots**, or **eigenvalues** of  $\mathbf{A}$ . If  $\lambda$  is an eigenvalue of  $\mathbf{A}$ , then  $\lambda \mathbf{I}_k - \mathbf{A}$  is singular so there exists a non-zero vector  $\mathbf{h}$  such that  $(\lambda \mathbf{I}_k - \mathbf{A}) \mathbf{h} = \mathbf{0}$  or

$$\mathbf{A}\mathbf{h} = \mathbf{h}\lambda.$$

The vector  $\mathbf{h}$  is called a **latent vector**, **characteristic vector**, or **eigenvector** of  $\mathbf{A}$  corresponding to  $\lambda$ . They are typically normalized so that  $\mathbf{h}'\mathbf{h} = 1$  and thus  $\lambda = \mathbf{h}'\mathbf{A}\mathbf{h}$ .

Set  $\mathbf{H} = [\mathbf{h}_1 \cdots \mathbf{h}_k]$  and  $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_k\}$ . A matrix expression is

$$\mathbf{AH} = \mathbf{H}\Lambda$$

We now state some useful properties.

**Theorem A.2** Properties of eigenvalues. Let  $\lambda_i$  and  $\mathbf{h}_i$ ,  $i = 1, \dots, k$ , denote the  $k$  eigenvalues and eigenvectors of a square matrix  $\mathbf{A}$ .

1.  $\det(\mathbf{A}) = \prod_{i=1}^k \lambda_i$
2.  $\text{tr}(\mathbf{A}) = \sum_{i=1}^k \lambda_i$
3.  $\mathbf{A}$  is non-singular if and only if all its eigenvalues are non-zero.
4. The non-zero eigenvalues of  $\mathbf{AB}$  and  $\mathbf{BA}$  are identical.
5. If  $\mathbf{B}$  is non-singular then  $\mathbf{A}$  and  $\mathbf{B}^{-1}\mathbf{AB}$  have the same eigenvalues.
6. If  $\mathbf{Ah} = \mathbf{h}\lambda$  then  $(\mathbf{I} - \mathbf{A})\mathbf{h} = \mathbf{h}(1 - \lambda)$ . So  $\mathbf{I} - \mathbf{A}$  has the eigenvalue  $1 - \lambda$  and associated eigenvector  $\mathbf{h}$ .

Most eigenvalue applications in econometrics concern the case where the matrix  $\mathbf{A}$  is real and symmetric. In this case all eigenvalues of  $\mathbf{A}$  are real and its eigenvectors are mutually orthogonal. Thus  $\mathbf{H}$  is orthonormal so  $\mathbf{H}'\mathbf{H} = \mathbf{I}_k$  and  $\mathbf{HH}' = \mathbf{I}_k$ . When the eigenvalues are all real it is conventional to write them in descending order  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k$ .

The following is a very important property of real symmetric matrices, which follows directly from the equations  $\mathbf{AH} = \mathbf{H}\Lambda$  and  $\mathbf{H}'\mathbf{H} = \mathbf{I}_k$ .

**Theorem A.3 (Spectral Decomposition)** If  $\mathbf{A}$  is a  $k \times k$  real symmetric matrix, then  $\mathbf{A} = \mathbf{H}\Lambda\mathbf{H}'$  where  $\mathbf{H}$  contains the eigenvectors and  $\Lambda$  is a diagonal matrix with the eigenvalues on the diagonal. The eigenvalues are all real and the eigenvector matrix satisfies  $\mathbf{H}'\mathbf{H} = \mathbf{I}_k$ .

The spectral decomposition can be alternatively written as  $\mathbf{H}'\mathbf{AH} = \Lambda$ .

If  $\mathbf{A}$  is real, symmetric, and invertible, then by the spectral decomposition and the properties of orthonormal matrices,  $\mathbf{A}^{-1} = \mathbf{H}'^{-1}\Lambda^{-1}\mathbf{H}^{-1} = \mathbf{H}\Lambda^{-1}\mathbf{H}'$ . Thus the columns of  $\mathbf{H}$  are also the eigenvectors of  $\mathbf{A}^{-1}$ , and its eigenvalues are  $\lambda_1^{-1}, \lambda_2^{-1}, \dots, \lambda_k^{-1}$ .

## A.10 Positive Definite Matrices

We say that a  $k \times k$  real symmetric square matrix  $\mathbf{A}$  is **positive semi-definite** if for all  $\mathbf{c} \neq \mathbf{0}$ ,  $\mathbf{c}'\mathbf{A}\mathbf{c} \geq 0$ . This is written as  $\mathbf{A} \geq 0$ . We say that  $\mathbf{A}$  is **positive definite** if for all  $\mathbf{c} \neq \mathbf{0}$ ,  $\mathbf{c}'\mathbf{A}\mathbf{c} > 0$ . This is written as  $\mathbf{A} > 0$ .

Some properties include:

**Theorem A.4** Properties of positive semi-definite matrices

1. If  $\mathbf{A} = \mathbf{G}'\mathbf{B}\mathbf{G}$  with  $\mathbf{B} \geq 0$  and some matrix  $\mathbf{G}$ , then  $\mathbf{A}$  is positive semi-definite. (For any  $\mathbf{c} \neq \mathbf{0}$ ,  $\mathbf{c}'\mathbf{A}\mathbf{c} = \mathbf{c}'\mathbf{B}\mathbf{c} \geq 0$  where  $\mathbf{c} = \mathbf{G}\mathbf{c}$ .) If  $\mathbf{G}$  has full column rank and  $\mathbf{B} > 0$ , then  $\mathbf{A}$  is positive definite.
2. If  $\mathbf{A}$  is positive definite, then  $\mathbf{A}$  is non-singular and  $\mathbf{A}^{-1}$  exists. Furthermore,  $\mathbf{A}^{-1} > 0$ .
3.  $\mathbf{A} > 0$  if and only if it is symmetric and all its eigenvalues are positive.
4. By the spectral decomposition,  $\mathbf{A} = \mathbf{H}\Lambda\mathbf{H}'$  where  $\mathbf{H}'\mathbf{H} = \mathbf{I}_k$  and  $\Lambda$  is diagonal with non-negative diagonal elements. All diagonal elements of  $\Lambda$  are strictly positive if (and only if)  $\mathbf{A} > 0$ .
5. The rank of  $\mathbf{A}$  equals the number of strictly positive eigenvalues.
6. If  $\mathbf{A} > 0$  then  $\mathbf{A}^{-1} = \mathbf{H}\Lambda^{-1}\mathbf{H}'$ .
7. If  $\mathbf{A} \geq 0$  and  $\text{rank}(\mathbf{A}) = r \leq k$  then the Moore-Penrose generalized inverse of  $\mathbf{A}$  is  $\mathbf{A}^- = \mathbf{H}\Lambda^-\mathbf{H}'$  where  $\Lambda^- = \text{diag}(\lambda_1^{-1}, \lambda_2^{-1}, \dots, \lambda_r^{-1}, 0, \dots, 0)$ .

## A.11 Idempotent Matrices

A  $k \times k$  square matrix  $\mathbf{A}$  is **idempotent** if  $\mathbf{AA} = \mathbf{A}$ . When  $k = 1$  the only idempotent numbers are 1 and 0. For  $k > 1$  there are many possibilities. For example, the following matrix is idempotent

$$\mathbf{A} = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix}.$$

If  $\mathbf{A}$  is idempotent and symmetric with rank  $r$ , then it has  $r$  eigenvalues which equal 1 and  $k - r$  eigenvalues which equal 0. To see this, by the spectral decomposition we can write  $\mathbf{A} = \mathbf{H}\Lambda\mathbf{H}'$  where  $\mathbf{H}$  is orthonormal and  $\Lambda$  contains the eigenvalues. Then

$$\mathbf{A} = \mathbf{AA} = \mathbf{H}\Lambda\mathbf{H}'\mathbf{H}\Lambda\mathbf{H}' = \mathbf{H}\Lambda^2\mathbf{H}'.$$

We deduce that  $\Lambda^2 = \Lambda$  and  $\lambda_i^2 = \lambda_i$  for  $i = 1, \dots, k$ . Hence each  $\lambda_i$  must equal either 0 or 1. Since the rank of  $\mathbf{A}$  is  $r$ , and the rank equals the number of positive eigenvalues, it follows that

$$\Lambda = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{k-r} \end{bmatrix}.$$

Thus the spectral decomposition of an idempotent matrix  $\mathbf{A}$  takes the form

$$\mathbf{A} = \mathbf{H} \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{k-r} \end{bmatrix} \mathbf{H}' \tag{A.4}$$

with  $\mathbf{H}'\mathbf{H} = \mathbf{I}_k$ . Additionally,  $\text{tr}(\mathbf{A}) = \text{rank}(\mathbf{A})$  and  $\mathbf{A}$  is positive semi-definite.

If  $\mathbf{A}$  is idempotent and symmetric with rank  $r < k$  then it does not possess an inverse, but its Moore-Penrose generalized inverse takes the simple form  $\mathbf{A}^- = \mathbf{A}$ . This can be verified by checking the conditions for the Moore-Penrose generalized inverse, for example  $\mathbf{AA}^- \mathbf{A} = \mathbf{AAA} = \mathbf{A}$ .

If  $\mathbf{A}$  is idempotent then  $\mathbf{I} - \mathbf{A}$  is also idempotent.

One useful fact is that if  $\mathbf{A}$  is idempotent then for any conformable vector  $\mathbf{c}$ ,

$$\mathbf{c}' \mathbf{A} \mathbf{c} \leq \mathbf{c}' \mathbf{c} \quad (\text{A.5})$$

$$\mathbf{c}' (\mathbf{I} - \mathbf{A}) \mathbf{c} \leq \mathbf{c}' \mathbf{c} \quad (\text{A.6})$$

To see this, note that

$$\mathbf{c}' \mathbf{c} = \mathbf{c}' \mathbf{A} \mathbf{c} + \mathbf{c}' (\mathbf{I} - \mathbf{A}) \mathbf{c}.$$

Since  $\mathbf{A}$  and  $\mathbf{I} - \mathbf{A}$  are idempotent, they are both positive semi-definite, so both  $\mathbf{c}' \mathbf{A} \mathbf{c}$  and  $\mathbf{c}' (\mathbf{I} - \mathbf{A}) \mathbf{c}$  are non-negative. Thus they must satisfy (A.5)-(A.6).

## A.12 Singular Values

The singular values of a  $k \times r$  real matrix  $\mathbf{A}$  are the positive square roots of the eigenvalues of  $\mathbf{A}' \mathbf{A}$ . Thus for  $j = 1, \dots, r$

$$s_j = \sqrt{\lambda_j(\mathbf{A}' \mathbf{A})}.$$

Since  $\mathbf{A}' \mathbf{A}$  is positive semi-definite, its eigenvalues are non-negative. Thus singular values are always real and non-negative.

The non-zero singular values of  $\mathbf{A}$  and  $\mathbf{A}'$  are the same.

When  $\mathbf{A}$  is positive semi-definite then the singular values of  $\mathbf{A}$  correspond to its eigenvalues.

It is convention to write the singular values in descending order  $s_1 \geq s_2 \geq \dots \geq s_r$ .

## A.13 Matrix Decompositions

There are several useful ways to decompose a matrix into the products of simpler matrices. We have already introduced the spectral decomposition, which we repeat here for completeness. The following apply to real matrices  $\mathbf{A}$ .

**Spectral Decomposition:** If  $\mathbf{A}$  is  $k \times k$  and symmetric then  $\mathbf{A} = \mathbf{H} \Lambda \mathbf{H}'$  where  $\mathbf{H}' \mathbf{H} = \mathbf{I}_k$  and  $\Lambda$  is a diagonal matrix with the (real) eigenvalues on the diagonal.

**Eigendecomposition:** If  $\mathbf{A}$  is  $k \times k$  and has distinct eigenvalues there exists a nonsingular matrix  $\mathbf{P}$  such that  $\mathbf{A} = \mathbf{P} \Lambda \mathbf{P}^{-1}$  and  $\mathbf{P}^{-1} \mathbf{A} \mathbf{P} = \Lambda$ . The columns of  $\mathbf{P}$  are the eigenvectors.  $\Lambda$  is diagonal with the eigenvalues on the diagonal.

**Matrix Square Root:** If  $\mathbf{A}$  is  $k \times k$  and positive definite we can find a matrix  $\mathbf{B}$  such that  $\mathbf{A} = \mathbf{B} \mathbf{B}'$ . We call  $\mathbf{B}$  a **matrix square root** of  $\mathbf{A}$  and is typically written as  $\mathbf{B} = \mathbf{A}^{1/2}$ .

The matrix  $\mathbf{B}$  need not be unique. One matrix square root is obtained using the spectral decomposition  $\mathbf{A} = \mathbf{H} \Lambda \mathbf{H}'$ . Then  $\mathbf{B} = \mathbf{H} \Lambda^{1/2} \mathbf{H}'$  is itself symmetric and positive definite and satisfies  $\mathbf{A} = \mathbf{B} \mathbf{B}'$ . Another matrix square root is the Cholesky decomposition, described in Section A.16.

**Singular Value Decomposition:** If  $\mathbf{A}$  is  $k \times r$  then  $\mathbf{A} = \mathbf{U} \Lambda \mathbf{V}'$  where  $\mathbf{U}$  is  $k \times k$ ,  $\Lambda$  is  $k \times r$  and  $\mathbf{V}$  is  $r \times r$ .  $\mathbf{U}$  and  $\mathbf{V}$  are orthonormal ( $\mathbf{U}' \mathbf{U} = \mathbf{I}_k$  and  $\mathbf{V}' \mathbf{V} = \mathbf{I}_r$ ).  $\Lambda$  is a diagonal matrix with the singular values of  $\mathbf{A}$  on the diagonal.

**Cholesky Decomposition:** If  $\mathbf{A}$  is  $k \times k$  and positive definite the  $\mathbf{A} = \mathbf{L} \mathbf{L}'$  where  $\mathbf{L}$  is **lower triangular** and full rank. See Section A.16.

**QR Decomposition:** If  $\mathbf{A}$  is  $k \times r$  with  $k \geq r$  and rank  $r$  then  $\mathbf{A} = \mathbf{Q} \mathbf{R}$ .  $\mathbf{Q}$  is a  $k \times r$  and orthonormal matrix ( $\mathbf{Q}' \mathbf{Q} = \mathbf{I}_r$ ).  $\mathbf{R}$  is a  $r \times r$  full rank **upper triangular** matrix. See Section A.17.

**Jordan Matrix Decomposition:** If  $A$  is  $k \times k$  with  $r$  unique eigenvalues then  $A = PJP^{-1}$  where  $J$  takes the **Jordan normal form**. The latter is a block diagonal matrix  $J = \text{diag}\{J_1, \dots, J_r\}$ . The **Jordan blocks**  $J_i$  are  $m_i \times m_i$  where  $m_i$  is the multiplicity of  $\lambda_i$  (number of eigenvalues equalling  $\lambda_i$ ) and take the form

$$J_i = \begin{bmatrix} \lambda_i & 1 & 0 \\ 0 & \lambda_i & 1 \\ 0 & 0 & \lambda_i \end{bmatrix}$$

illustrated here for  $m_i = 3$ .

## A.14 Generalized Eigenvalues

Let  $A$  and  $B$  be  $k \times k$  matrices. The generalized characteristic equation is

$$\det(\mu B - A) = 0.$$

The solutions  $\mu$  are known as **generalized eigenvalues** of  $A$  with respect to  $B$ . Associated with each generalized eigenvalue  $\mu$  is a **generalized eigenvector**  $v$  which satisfies

$$Av = Bv\mu.$$

They are typically normalized so that  $v' B v = 1$  and thus  $\mu = v' A v$ .

A matrix expression is

$$AV = BVM$$

where  $M = \text{diag}\{\mu_1, \dots, \mu_k\}$ .

If  $A$  and  $B$  are real and symmetric then the generalized eigenvalues are real.

Suppose in addition that  $B$  is invertible. Then the generalized eigenvalues of  $A$  with respect to  $B$  are equal to the eigenvalues of  $B^{-1/2}AB^{-1/2}$ . The generalized eigenvectors  $V$  of  $A$  with respect to  $B$  are related to the eigenvectors  $H$  of  $B^{-1/2}AB^{-1/2}$  by the relationship  $V = B^{-1/2}H$ . This implies  $V'BV = I_k$ . Thus the generalized eigenvectors are orthogonalized with respect to the matrix  $B$ .

If  $Av = Bv\mu$  then  $(B - A)v = Bv(1 - \mu)$ . So a generalized eigenvalue of  $B - A$  with respect to  $B$  is  $1 - \mu$  with associated eigenvector  $v$ .

Generalized eigenvalue equations have an interesting dual property. The following is based on Lemma A.9 of Johansen (1995).

**Theorem A.5** Suppose that  $B$  and  $C$  are invertible  $p \times p$  and  $r \times r$  matrices, respectively, and  $A$  is  $p \times r$ . Then the generalized eigenvalue problems

$$\det(\mu B - AC^{-1}A') = 0 \tag{A.7}$$

and

$$\det(\mu C - A'B^{-1}A) = 0 \tag{A.8}$$

have the same non-zero generalized eigenvalues. Furthermore, for any such generalized eigenvalue  $\mu$ , if  $v$  and  $w$  are the associated generalized eigenvectors of (A.7) and (A.8), then

$$w = \mu^{-1/2}C^{-1}A'v. \tag{A.9}$$

**Proof.** Let  $\mu \neq 0$  be an eigenvalue of (A.7). Then using Theorem A.1.8

$$\begin{aligned} 0 &= \det(\mu B - AC^{-1}A') \\ &= \frac{\det(\mu B)}{\det(C)} \det(C - A'(\mu B)^{-1}A) \\ &= \frac{\det(B)}{\det(C)} \det(\mu C - A'B^{-1}A). \end{aligned}$$

Since  $\det(\mathbf{B}) / \det(\mathbf{C}) \neq 0$  this implies (A.9) holds. Hence  $\mu$  is an eigenvalue of (A.8), as claimed.

We next show that (A.9) is an eigenvector of (A.8). Note that the solutions to (A.7) and (A.8) satisfy

$$\mathbf{B}\mathbf{v}\mu = \mathbf{AC}^{-1}\mathbf{A}'\mathbf{v} \quad (\text{A.10})$$

and

$$\mathbf{C}\mathbf{w}\mu = \mathbf{A}'\mathbf{B}^{-1}\mathbf{Aw} \quad (\text{A.11})$$

and are normalized so that  $\mathbf{v}'\mathbf{B}\mathbf{v} = 1$  and  $\mathbf{w}'\mathbf{C}\mathbf{w} = 1$ . We show that (A.9) satisfies (A.11). Using (A.9), we find that the left-side of (A.11) equals

$$\mathbf{C}(\mu^{-1/2}\mathbf{C}^{-1}\mathbf{A}')\mu = \mathbf{A}'\mu^{1/2} = \mathbf{A}'\mathbf{B}^{-1}\mathbf{B}\mathbf{v}\mu^{1/2} = \mathbf{A}'\mathbf{B}^{-1}\mathbf{AC}^{-1}\mathbf{A}'\mathbf{v}\mu^{-1/2} = \mathbf{A}'\mathbf{B}^{-1}\mathbf{Aw}$$

The third equality is (A.10) and the final is (A.9). This shows that (A.11) holds and thus (A.9) is an eigenvector of (A.8) as stated. ■

## A.15 Extrema of Quadratic Forms

The extrema of quadratic forms in real symmetric matrices can be conveniently be written in terms of eigenvalues and eigenvectors.

Let  $\mathbf{A}$  denote a  $k \times k$  real symmetric matrix. Let  $\lambda_1 \geq \dots \geq \lambda_k$  be the ordered eigenvalues of  $\mathbf{A}$  and  $\mathbf{h}_1, \dots, \mathbf{h}_k$  the associated ordered eigenvectors.

We start with results for the extrema of  $\mathbf{x}'\mathbf{Ax}$ . Throughout this Section, when we refer to the “solution” of an extremum problem, it is the solution to the normalized expression.

- $\max_{\mathbf{x}'\mathbf{x}=1} \mathbf{x}'\mathbf{Ax} = \max_{\mathbf{x}} \frac{\mathbf{x}'\mathbf{Ax}}{\mathbf{x}'\mathbf{x}} = \lambda_1$ . The solution is  $\mathbf{x} = \mathbf{h}_1$ . (That is, the maximizer of  $\mathbf{x}'\mathbf{Ax}$  over  $\mathbf{x}'\mathbf{x} = 1$ .)
- $\min_{\mathbf{x}'\mathbf{x}=1} \mathbf{x}'\mathbf{Ax} = \min_{\mathbf{x}} \frac{\mathbf{x}'\mathbf{Ax}}{\mathbf{x}'\mathbf{x}} = \lambda_k$ . The solution is  $\mathbf{x} = \mathbf{h}_k$ .

Multivariate generalizations can involve either the trace or the determinant.

- $\max_{\mathbf{X}'\mathbf{X}=\mathbf{I}_\ell} \text{tr}(\mathbf{X}'\mathbf{AX}) = \max_{\mathbf{X}} \text{tr}\left((\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{AX})\right) = \sum_{i=1}^{\ell} \lambda_i$ .

The solution is  $\mathbf{X} = [\mathbf{h}_1, \dots, \mathbf{h}_\ell]$ .

- $\min_{\mathbf{X}'\mathbf{X}=\mathbf{I}_\ell} \text{tr}(\mathbf{X}'\mathbf{AX}) = \min_{\mathbf{X}} \text{tr}\left((\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{AX})\right) = \sum_{i=1}^{\ell} \lambda_{k-i+1}$ .

The solution is  $\mathbf{X} = [\mathbf{h}_{k-\ell+1}, \dots, \mathbf{h}_k]$ .

For a proof, see Theorem 11.13 of Magnus and Neudecker (1988).

Suppose as well that  $\mathbf{A} > 0$  with ordered eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k$  and eigenvectors  $[\mathbf{h}_1, \dots, \mathbf{h}_k]$

- $\max_{\mathbf{X}'\mathbf{X}=\mathbf{I}_\ell} \det(\mathbf{X}'\mathbf{AX}) = \max_{\mathbf{X}} \frac{\det(\mathbf{X}'\mathbf{AX})}{\det(\mathbf{X}'\mathbf{X})} = \prod_{i=1}^{\ell} \lambda_i$ . The solution is  $\mathbf{X} = [\mathbf{h}_1, \dots, \mathbf{h}_\ell]$ .
- $\min_{\mathbf{X}'\mathbf{X}=\mathbf{I}_\ell} \det(\mathbf{X}'\mathbf{AX}) = \min_{\mathbf{X}} \frac{\det(\mathbf{X}'\mathbf{AX})}{\det(\mathbf{X}'\mathbf{X})} = \prod_{i=1}^{\ell} \lambda_{k-i+1}$ . The solution is  $\mathbf{X} = [\mathbf{h}_{k-\ell+1}, \dots, \mathbf{h}_k]$ .
- $\max_{\mathbf{X}'\mathbf{X}=\mathbf{I}_\ell} \det(\mathbf{X}'(\mathbf{I}-\mathbf{A})\mathbf{X}) = \max_{\mathbf{X}} \frac{\det(\mathbf{X}'(\mathbf{I}-\mathbf{A})\mathbf{X})}{\det(\mathbf{X}'\mathbf{X})} = \prod_{i=1}^{\ell} (1 - \lambda_{k-i+1})$ . The solution is  $\mathbf{X} = [\mathbf{h}_{k-\ell+1}, \dots, \mathbf{h}_k]$ .

- $\min_{X'X=I_\ell} \det(X'(\mathbf{I} - \mathbf{A})X) = \min_X \frac{\det(X'(\mathbf{I} - \mathbf{A})X)}{\det(X'X)} = \prod_{i=1}^{\ell} (1 - \lambda_i)$ . The solution is  $X = [\mathbf{h}_1, \dots, \mathbf{h}_\ell]$ .

For a proof, see Theorem 11.15 of Magnus and Neudecker (1988).

We can extend the above results to incorporate generalized eigenvalue equations.

Let  $\mathbf{A}$  and  $\mathbf{B}$  be  $k \times k$  real symmetric matrices with  $\mathbf{B} > 0$ . Let  $\mu_1 \geq \dots \geq \mu_k$  be the ordered generalized eigenvalues of  $\mathbf{A}$  with respect to  $\mathbf{B}$  and  $\mathbf{v}_1, \dots, \mathbf{v}_k$  the associated ordered eigenvectors.

- $\max_{x'Bx=1} x'\mathbf{A}x = \max_x \frac{x'\mathbf{A}x}{x'\mathbf{B}x} = \mu_1$ . The solution is  $x = \mathbf{v}_1$ .
- $\min_{x'Bx=1} x'\mathbf{A}x = \min_x \frac{x'\mathbf{A}x}{x'\mathbf{B}x} = \mu_k$ . The solution is  $x = \mathbf{v}_k$ .
- $\max_{X'BX=I_\ell} \text{tr}(X'\mathbf{A}X) = \max_X \text{tr}\left((X'\mathbf{B}X)^{-1}(X'\mathbf{A}X)\right) = \sum_{i=1}^{\ell} \mu_i$ .  
The solution is  $X = [\mathbf{v}_1, \dots, \mathbf{v}_\ell]$ .
- $\min_{X'BX=I_\ell} \text{tr}(X'\mathbf{A}X) = \min_X \text{tr}\left((X'\mathbf{B}X)^{-1}(X'\mathbf{A}X)\right) = \sum_{i=1}^{\ell} \mu_{k-i+1}$ .  
The solution is  $X = [\mathbf{v}_{k-\ell+1}, \dots, \mathbf{v}_k]$ .

Suppose as well that  $\mathbf{A} > 0$ .

- $\max_{X'BX=I_\ell} \det(X'\mathbf{A}X) = \max_X \frac{\det(X'\mathbf{A}X)}{\det(X'\mathbf{B}X)} = \prod_{i=1}^{\ell} \mu_i$ .  
The solution is  $X = [\mathbf{v}_1, \dots, \mathbf{v}_\ell]$ .
- $\min_{X'BX=I_\ell} \det(X'\mathbf{A}X) = \min_X \frac{\det(X'\mathbf{A}X)}{\det(X'\mathbf{B}X)} = \prod_{i=1}^{\ell} \mu_{k-i+1}$ .  
The solution is  $X = [\mathbf{v}_{k-\ell+1}, \dots, \mathbf{v}_k]$ .
- $\max_{X'BX=I_\ell} \det(X'(\mathbf{I} - \mathbf{A})X) = \max_X \frac{\det(X'(\mathbf{I} - \mathbf{A})X)}{\det(X'\mathbf{B}X)} = \prod_{i=1}^{\ell} (1 - \mu_{k-i+1})$ .  
The solution is  $X = [\mathbf{v}_{k-\ell+1}, \dots, \mathbf{v}_k]$ .
- $\min_{X'BX=I_\ell} \det(X'(\mathbf{I} - \mathbf{A})X) = \min_X \frac{\det(X'(\mathbf{I} - \mathbf{A})X)}{\det(X'\mathbf{B}X)} = \prod_{i=1}^{\ell} (1 - \mu_i)$ .  
The solution is  $X = [\mathbf{v}_1, \dots, \mathbf{v}_\ell]$ .

By change-of-variables, we can re-express one eigenvalue problem in terms of another. For example, let  $\mathbf{A} > 0$ ,  $\mathbf{B} > 0$ , and  $\mathbf{C} > 0$ . Then

$$\max_X \frac{\det(X'\mathbf{CAC}X)}{\det(X'\mathbf{CBC}X)} = \max_X \frac{\det(X'\mathbf{AX})}{\det(X'\mathbf{BX})}$$

and

$$\min_X \frac{\det(X'\mathbf{CAC}X)}{\det(X'\mathbf{CBC}X)} = \min_X \frac{\det(X'\mathbf{AX})}{\det(X'\mathbf{BX})}$$

## A.16 Cholesky Decomposition

For a  $k \times k$  positive definite matrix  $\mathbf{A}$ , its **Cholesky decomposition** takes the form

$$\mathbf{A} = \mathbf{LL}'$$

where  $\mathbf{L}$  is **lower triangular** and full rank. A lower triangular matrix (also known as a **left triangular** matrix) takes the form

$$\mathbf{L} = \begin{bmatrix} L_{11} & 0 & \cdots & 0 \\ L_{21} & L_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ L_{k1} & L_{k2} & \cdots & L_{kk} \end{bmatrix}.$$

The diagonal elements of  $\mathbf{L}$  are all strictly positive. The Cholesky decomposition is unique (for positive definite  $\mathbf{A}$ ).

The decomposition is very useful for a range of computations, especially when a matrix square root is required. Algorithms for computation are available in standard packages (for example, `chol` in either MATLAB or R).

Lower triangular matrices such as  $\mathbf{L}$  have special properties. One is that its determinant equals the product of the diagonal elements.

Proofs of uniqueness of the Cholesky decomposition (as well as computation) are algorithmic. Here are the details for the case  $k = 3$ . Write out

$$\begin{bmatrix} A_{11} & A_{21} & A_{31} \\ A_{21} & A_{22} & A_{32} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} = \mathbf{A} = \mathbf{LL}' = \begin{bmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{bmatrix} \begin{bmatrix} L_{11} & L_{21} & L_{31} \\ 0 & L_{22} & L_{32} \\ 0 & 0 & L_{33} \end{bmatrix} \\ = \begin{bmatrix} L_{11}^2 & L_{11}L_{21} & L_{11}L_{31} \\ L_{11}L_{21} & L_{21}^2 + L_{22}^2 & L_{31}L_{21} + L_{32}L_{22} \\ L_{11}L_{31} & L_{31}L_{21} + L_{32}L_{22} & L_{31}^2 + L_{32}^2 + L_{33}^2 \end{bmatrix}.$$

There are six equations, six knowns (the elements of  $\mathbf{A}$ ) and six unknowns (the elements of  $\mathbf{L}$ ). We can solve for the latter by starting with the first column, moving from top to bottom. The first element has the simple solution

$$L_{11} = \sqrt{A_{11}}.$$

This has a real solution since  $A_{11} > 0$ . Moving down, since  $L_{11}$  is known, for the entries beneath  $L_{11}$  we solve and find

$$L_{21} = \frac{A_{21}}{L_{11}} = \frac{A_{21}}{\sqrt{A_{11}}}$$

$$L_{31} = \frac{A_{31}}{L_{11}} = \frac{A_{31}}{\sqrt{A_{11}}}.$$

Next we move to the second column. We observe that  $L_{21}$  is known. Then we solve for  $L_{22}$

$$L_{22} = \sqrt{A_{22} - L_{21}^2} = \sqrt{A_{22} - \frac{A_{21}^2}{A_{11}}}.$$

This has a real solution since  $\mathbf{A} > 0$ . Then since  $L_{22}$  is known we can move down the column to find

$$L_{32} = \frac{A_{32} - L_{31}L_{21}}{L_{22}} = \frac{A_{32} - \frac{A_{31}A_{21}}{A_{11}}}{\sqrt{A_{22} - \frac{A_{21}^2}{A_{11}}}}.$$

Finally we take the third column. All elements except  $L_{33}$  are known. So we solve to find

$$L_{33} = \sqrt{A_{33} - L_{31}^2 - L_{32}^2} = \sqrt{A_{33} - \frac{A_{31}^2}{A_{11}} - \frac{\left(A_{32} - \frac{A_{31}A_{21}}{A_{11}}\right)^2}{A_{22} - \frac{A_{21}^2}{A_{11}}}}.$$

## A.17 QR Decomposition

The QR decomposition is widely used for numerical problems such as matrix inversion and solving systems of linear equations.

Let  $\mathbf{A}$  be an  $k \times r$  matrix, with  $k \geq r$  and rank  $r$ . The QR decomposition of  $\mathbf{A}$  is

$$\mathbf{A} = \mathbf{Q}\mathbf{R}$$

where  $\mathbf{Q}$  is a  $k \times r$  orthonormal matrix and  $\mathbf{R}$  is a  $r \times r$  full rank **upper triangular** matrix (also known as a **right triangular** matrix).

To show that the QR decomposition exists, observe that  $\mathbf{A}'\mathbf{A}$  is  $r \times r$  and positive definite. Apply the Cholesky decomposition to find

$$\mathbf{A}'\mathbf{A} = \mathbf{L}\mathbf{L}'$$

where  $\mathbf{L}$  is lower triangular and full rank. We then set

$$\begin{aligned}\mathbf{Q} &= \mathbf{A}(\mathbf{L}')^{-1} \\ \mathbf{R} &= \mathbf{L}'.\end{aligned}$$

The matrix  $\mathbf{R}$  is upper triangular by construction. Also,

$$\begin{aligned}\mathbf{Q}'\mathbf{Q} &= (\mathbf{L}')^{-1'} \mathbf{A}'\mathbf{A} (\mathbf{L}')^{-1} \\ &= \mathbf{L}^{-1} \mathbf{L} \mathbf{L}' (\mathbf{L}')^{-1} \\ &= \mathbf{I}_k\end{aligned}$$

so  $\mathbf{Q}$  is orthonormal as claimed.

Numerical computation of the QR decomposition does not use the above matrix operations. Instead it is done algorithmically. Standard methods include the **Gram-Schmidt** and **Householder** algorithms. The Gram-Schmidt is simple to describe and implement, but the Householder is numerically more stable and is therefore the standard implementation. Since the algorithm is involved we do not describe it here.

## A.18 Solving Linear Systems

A linear system of  $k$  equations with  $k$  unknowns is

$$\begin{aligned}a_{11}b_1 + a_{12}b_2 + \cdots + a_{1k}b_k &= c_1 \\ a_{21}b_1 + a_{22}b_2 + \cdots + a_{2k}b_k &= c_2 \\ &\vdots \\ a_{k1}b_1 + a_{k2}b_2 + \cdots + a_{kk}b_k &= c_k\end{aligned}$$

or in matrix notation

$$\mathbf{Ab} = \mathbf{c} \tag{A.12}$$

where  $\mathbf{A}$  is  $k \times k$ , and  $\mathbf{b}$  and  $\mathbf{c}$  are  $k \times 1$ . If  $\mathbf{A}$  is full rank then the solution  $\mathbf{b} = \mathbf{A}^{-1}\mathbf{c}$  is unique. In this section we describe three algorithms to numerically find the solution  $\mathbf{b}$ . The first uses Gaussian elimination, the second uses the QR decomposition, and the third uses the Cholesky decomposition for positive definite  $\mathbf{A}$ .

### (1) Solving by Gaussian elimination

The solution  $\mathbf{b}$  in (A.12) is invariant to row operations; including multiplying an equation by non-zero numbers, and adding and subtracting equations from one another. To exploit this insight combine the known constants  $\mathbf{A}$  and  $\mathbf{c}$  into a  $k \times (k+1)$  augmented matrix

$$[\mathbf{A}|\mathbf{c}] . \tag{A.13}$$

The row operations described above are the same as multiplying rows of  $[A|c]$  by non-zero numbers, and adding and subtracting rows of  $[A|c]$  from one another. Such operations do not change the solution  $\mathbf{b}$ . Gaussian elimination works by applying row operations to  $[A|c]$  until the left section equals the identity matrix  $I_k$  and thus equals

$$[I_k|\mathbf{d}]. \quad (\text{A.14})$$

Since row operations do not alter the solution, this means that the solution  $\mathbf{b}$  in (A.12) also satisfies  $I_k\mathbf{b} = \mathbf{d}$  which implies  $\mathbf{b} = \mathbf{d}$ . Thus the solution  $\mathbf{b}$  can be found as the right-most vector  $\mathbf{d}$  in (A.14).

The Gauss-Jordan algorithm implements a sequence of row operations which obtains the solution for any pair (A.13) such that  $A$  is full rank. The algorithm is as follows.

For  $r = 1, \dots, k$ :

1. Divide the elements of row  $r$  by  $a_{rr}$ . Thus make the changes

- (a)  $a_{ri} \mapsto a_{ri}/a_{rr}$  for  $i = 1, \dots, k$
- (b)  $c_r \mapsto c_r/a_{rr}$

2. For rows  $j \neq r$ , subtract  $a_{jr}$  times row  $r$  from row  $j$ . Thus make the changes

- (a)  $a_{ji} \mapsto a_{ji} - a_{jr}a_{ri}$  for  $i = 1, \dots, k$
- (b)  $c_j \mapsto c_j - a_{jr}c_r$

Each pair of operations transforms a column of the matrix  $A$  into a column of the identity matrix  $I_k$ , starting with the first column and working sequentially to the right. The first operation (dividing by  $a_{rr}$ ) normalizes the  $r^{\text{th}}$  diagonal element to unity. The second set of operations makes row operations to transform the remaining elements of the  $r^{\text{th}}$  column to equal zero. Since the previous columns are unit vectors they are unaffected by these operations.

## (2) Solving by QR Decomposition

First, compute the QR decomposition

$$A = QR$$

where  $Q$  is a  $k \times k$  orthogonal matrix, and  $R$  is  $k \times k$  and upper triangular. This is done numerically (typically by the Householder algorithm) as described in Section A.17. This means that (A.12) can be written as

$$QR\mathbf{b} = \mathbf{c}.$$

Premultiplying by  $Q'$  and observing  $Q'Q = I_k$  we obtain

$$R\mathbf{b} = Q'\mathbf{c} \stackrel{\text{def}}{=} \mathbf{d}.$$

This system can be written as

$$\begin{aligned} r_{11}b_1 + r_{12}b_2 + \cdots + r_{1,k-1}b_{k-2} + r_{1k}b_k &= d_1 \\ r_{22}b_2 + \cdots + r_{2,k-1}b_{k-2} + r_{2k}b_k &= d_2 \\ &\vdots \\ r_{k-1,k-1}b_{k-2} + r_{k-1,k}b_k &= d_{k-1} \\ r_{kk}b_k &= d_k. \end{aligned}$$

This can be solved by **backwards recursion**

$$\begin{aligned} b_k &= d_k/r_{kk} \\ b_{k-1} &= (d_{k-1} - r_{k-1,k}b_k)/r_{k-1,k-1} \\ &\vdots \\ b_1 &= (d_1 - r_{12}b_2 - \cdots - r_{1k}b_k)/r_{11}. \end{aligned}$$

To summarize, the **QR solution method** is

1. Numerically compute the QR decomposition  $\mathbf{A} = \mathbf{Q}\mathbf{R}$ .
2. Calculate  $\mathbf{d} = \mathbf{Q}'\mathbf{c}$ .
3. Solve for  $\mathbf{b}$  by backward recursion.

(3) Solving by **Cholesky Decomposition** for positive definite  $\mathbf{A}$

First, compute the Cholesky decomposition

$$\mathbf{A} = \mathbf{L}\mathbf{R}$$

where  $\mathbf{L}$  is  $k \times k$  and lower triangular, and  $\mathbf{R} = \mathbf{L}'$  is upper triangular. This is done numerically as described in Section A.16. This means that (A.12) can be written as

$$\mathbf{L}\mathbf{R}\mathbf{b} = \mathbf{c}.$$

or

$$\mathbf{L}\mathbf{d} = \mathbf{c}$$

where  $\mathbf{d} = \mathbf{R}\mathbf{b}$ . The vector  $\mathbf{d}$  can be solved from  $\mathbf{L}$  and  $\mathbf{c}$  using **forward recursion**. The equation

$$\mathbf{R}\mathbf{b} = \mathbf{d}$$

can then be solved for  $\mathbf{b}$  by backwards recursion.

We have described three algorithms. Which should be used in practice? For positive definite  $\mathbf{A}$ , solving by the Cholesky decomposition is the preferred method as it is numerically most efficient and stable. When  $\mathbf{A}$  is not positive definite, solving by the QR decomposition is the preferred method as it is numerically most stable. The advantage of the Gauss-Jordan algorithm is that it is the simplest to program.

## A.19 Algorithmic Matrix Inversion

Numerical methods for solving linear systems can be used to calculate the inverse of a full-rank  $k \times k$  matrix  $\mathbf{A}$ . Let  $\mathbf{B} = \mathbf{A}^{-1}$  be the inverse of  $\mathbf{A}$ . The matrices satisfy

$$\mathbf{AB} = \mathbf{I}_k$$

which is a matrix generalization of (A.12). The goal is to solve this system to obtain  $\mathbf{B}$ .

(1) Solving by Gaussian elimination

Replace  $\mathbf{c}$  in (A.13) with  $\mathbf{I}_k$  and apply the Gauss-Jordan elimination algorithm. The solution is  $\mathbf{B}$ .

(2) Solving by QR decomposition

Numerically compute the QR decomposition

$$\mathbf{A} = \mathbf{Q}\mathbf{R}.$$

This implies

$$\mathbf{Q}\mathbf{RB} = \mathbf{I}_k.$$

Premultiplying by  $\mathbf{Q}'$  and observing  $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_k$  we obtain

$$\mathbf{RB} = \mathbf{Q}'.$$

Write  $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_k]$  and  $\mathbf{Q}' = [\mathbf{q}_1, \dots, \mathbf{q}_k]$ . For  $j = 1, \dots, k$

$$\mathbf{RB}_j = \mathbf{q}_j.$$

Since  $\mathbf{R}$  is upper triangular the vector  $\mathbf{b}_j$  can be found by backwards recursion as described in Section A.18.

(3) Solving by Cholesky decomposition for positive definite  $\mathbf{A}$

Compute the Cholesky decomposition

$$\mathbf{A} = \mathbf{L}\mathbf{R}$$

where  $\mathbf{L}$  is  $k \times k$  and lower triangular and  $\mathbf{R} = \mathbf{L}'$  is upper triangular. This implies

$$\mathbf{L}\mathbf{R}\mathbf{B} = \mathbf{I}_k$$

or

$$\mathbf{L}\mathbf{C} = \mathbf{I}_k$$

where  $\mathbf{C} = \mathbf{R}\mathbf{B}$ . Applying forward recursion one column at a time we can solve for  $\mathbf{C}$ . We then have

$$\mathbf{R}\mathbf{B} = \mathbf{C}.$$

Applying backwards recursion one column at a time we can solve for  $\mathbf{B}$ .

## A.20 Matrix Calculus

Let  $\mathbf{x} = (x_1, \dots, x_k)'$  be  $k \times 1$  and  $g(\mathbf{x}) = g(x_1, \dots, x_k) : \mathbb{R}^k \rightarrow \mathbb{R}$ . The vector derivative is

$$\frac{\partial}{\partial \mathbf{x}} g(\mathbf{x}) = \begin{pmatrix} \frac{\partial}{\partial x_1} g(\mathbf{x}) \\ \vdots \\ \frac{\partial}{\partial x_k} g(\mathbf{x}) \end{pmatrix}$$

and

$$\frac{\partial}{\partial \mathbf{x}'} g(\mathbf{x}) = \begin{pmatrix} \frac{\partial}{\partial x_1} g(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_k} g(\mathbf{x}) \end{pmatrix}.$$

Some properties are now summarized.

**Theorem A.6** Properties of matrix derivatives

1.  $\frac{\partial}{\partial \mathbf{x}} (\mathbf{a}' \mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} (\mathbf{x}' \mathbf{a}) = \mathbf{a}$
2.  $\frac{\partial}{\partial \mathbf{x}} (\mathbf{x}' \mathbf{A}) = \mathbf{A}$  and  $\frac{\partial}{\partial \mathbf{x}'} (\mathbf{A} \mathbf{x}) = \mathbf{A}$
3.  $\frac{\partial}{\partial \mathbf{x}} (\mathbf{x}' \mathbf{A} \mathbf{x}) = (\mathbf{A} + \mathbf{A}') \mathbf{x}$
4.  $\frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}'} (\mathbf{x}' \mathbf{A} \mathbf{x}) = \mathbf{A} + \mathbf{A}'$
5.  $\frac{\partial}{\partial \mathbf{A}} \text{tr}(\mathbf{B} \mathbf{A}) = \mathbf{B}'$
6.  $\frac{\partial}{\partial \mathbf{A}} \log \det(\mathbf{A}) = (\mathbf{A}^-)'$

To show part 1, note that

$$\frac{\partial}{\partial x_j} (\mathbf{a}' \mathbf{x}) = \frac{\partial}{\partial x_j} (a_1 x_1 + \cdots + a_k x_k) = a_j.$$

Thus

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{a}' \mathbf{x}) = \begin{pmatrix} a_1 \\ \vdots \\ a_k \end{pmatrix} = \mathbf{a}$$

as claimed.

For part 2, write  $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_m]$  so that

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{x}' \mathbf{A}) = \frac{\partial}{\partial \mathbf{x}} [\mathbf{x}' \mathbf{a}_1, \dots, \mathbf{x}' \mathbf{a}_m] = \left[ \frac{\partial}{\partial \mathbf{x}} (\mathbf{x}' \mathbf{a}_1), \dots, \frac{\partial}{\partial \mathbf{x}} (\mathbf{x}' \mathbf{a}_m) \right] = [\mathbf{a}_1, \dots, \mathbf{a}_m] = \mathbf{A}$$

using part 1.  $\frac{\partial}{\partial \mathbf{x}'} (\mathbf{A} \mathbf{x}) = \mathbf{A}$  follows by taking the transpose.

For part 3, notice  $\mathbf{x}' \mathbf{A} \mathbf{x} = \mathbf{x}' \mathbf{A}' \mathbf{x}$  and apply the product rule and then part 2,

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{x}' \mathbf{A} \mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} (\mathbf{x}' \mathbf{I}_k) \mathbf{A} \mathbf{x} + \frac{\partial}{\partial \mathbf{x}} (\mathbf{x}' \mathbf{A}') \mathbf{x} = \mathbf{I}_k \mathbf{A} \mathbf{x} + \mathbf{A}' \mathbf{x} = (\mathbf{A} + \mathbf{A}') \mathbf{x}.$$

For part 4, applying part 3 we find

$$\frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}'} (\mathbf{x}' \mathbf{A} \mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} \frac{\partial}{\partial \mathbf{x}'} (\mathbf{x}' \mathbf{A} \mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} \mathbf{x}' (\mathbf{A} + \mathbf{A}') = \mathbf{A} + \mathbf{A}'.$$

For part 5, recall from Section A.5 that we can write out explicitly

$$\text{tr}(\mathbf{B} \mathbf{A}) = \sum_i \sum_j a_{ij} b_{ji}.$$

Thus if we take the derivative with respect to  $a_{ij}$  we find

$$\frac{\partial}{\partial a_{ij}} \text{tr}(\mathbf{B} \mathbf{A}) = b_{ji}.$$

which is the  $ij^{th}$  element of  $\mathbf{B}'$ , establishing part 5.

For part 6, recall Laplace's expansion

$$\det \mathbf{A} = \sum_{j=1}^k a_{ij} C_{ij}.$$

where  $C_{ij}$  is the  $ij^{th}$  cofactor of  $\mathbf{A}$ . Set  $\mathbf{C} = (C_{ij})$ . Observe that  $C_{ij}$  for  $j = 1, \dots, k$  are not functions of  $a_{ij}$ . Thus the derivative with respect to  $a_{ij}$  is

$$\frac{\partial}{\partial a_{ij}} \log \det(\mathbf{A}) = (\det \mathbf{A})^{-1} \frac{\partial}{\partial a_{ij}} \det \mathbf{A} = (\det \mathbf{A})^{-1} C_{ij}.$$

Together this implies

$$\frac{\partial}{\partial \mathbf{A}} \log \det(\mathbf{A}) = (\det \mathbf{A})^{-1} \mathbf{C} = \mathbf{A}^{-1}$$

where the second equality is Theorem A.1.12.

## A.21 Kronecker Products and the Vec Operator

Let  $\mathbf{A} = [\mathbf{a}_1 \ \mathbf{a}_2 \ \cdots \ \mathbf{a}_n]$  be  $m \times n$ . The **vec** of  $\mathbf{A}$ , denoted by  $\text{vec}(\mathbf{A})$ , is the  $mn \times 1$  vector

$$\text{vec}(\mathbf{A}) = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{pmatrix}.$$

Let  $\mathbf{A} = (a_{ij})$  be an  $m \times n$  matrix and let  $\mathbf{B}$  be any matrix. The **Kronecker product** of  $\mathbf{A}$  and  $\mathbf{B}$ , denoted  $\mathbf{A} \otimes \mathbf{B}$ , is the matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}.$$

Some important properties are now summarized. These results hold for matrices for which all matrix multiplications are conformable.

### Theorem A.7 Properties of the Kronecker product

1.  $(\mathbf{A} + \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes \mathbf{C} + \mathbf{B} \otimes \mathbf{C}$
2.  $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$
3.  $\mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C}) = (\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C}$
4.  $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$
5.  $\text{tr}(\mathbf{A} \otimes \mathbf{B}) = \text{tr}(\mathbf{A}) \text{tr}(\mathbf{B})$
6. If  $\mathbf{A}$  is  $m \times m$  and  $\mathbf{B}$  is  $n \times n$ ,  $\det(\mathbf{A} \otimes \mathbf{B}) = (\det(\mathbf{A}))^n (\det(\mathbf{B}))^m$
7.  $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$
8. If  $\mathbf{A} > 0$  and  $\mathbf{B} > 0$  then  $\mathbf{A} \otimes \mathbf{B} > 0$
9.  $\text{vec}(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A}) \text{vec}(\mathbf{B})$
10.  $\text{tr}(\mathbf{ABCD}) = \text{vec}(\mathbf{D}')' (\mathbf{C}' \otimes \mathbf{A}) \text{vec}(\mathbf{B})$

## A.22 Vector Norms

Given any vector space  $V$  (such as Euclidean space  $\mathbb{R}^m$ ) a **norm** on  $V$  is a function  $\rho : V \rightarrow \mathbb{R}$  with the properties

1.  $\rho(c\mathbf{a}) = |c| \rho(\mathbf{a})$  for any complex number  $c$  and  $\mathbf{a} \in V$
2.  $\rho(\mathbf{a} + \mathbf{b}) \leq \rho(\mathbf{a}) + \rho(\mathbf{b})$
3. If  $\rho(\mathbf{a}) = 0$  then  $\mathbf{a} = \mathbf{0}$ .

A seminorm on  $V$  is a function which satisfies the first two properties. The second property is known as the triangle inequality, and it is the one property which typically needs a careful demonstration (as the other two properties typically hold by inspection).

The typical norm used for Euclidean space  $\mathbb{R}^m$  is the **Euclidean norm**

$$\|\mathbf{a}\| = (\mathbf{a}' \mathbf{a})^{1/2} = \left( \sum_{i=1}^m a_i^2 \right)^{1/2}.$$

An alternative norm is the  $p$ -norm (for  $p \geq 1$ )

$$\|\mathbf{a}\|_p = \left( \sum_{i=1}^m |a_i|^p \right)^{1/p}.$$

Special cases include the Euclidean norm ( $p = 2$ ), the 1–norm

$$\|\mathbf{a}\|_1 = \sum_{i=1}^m |a_i|$$

and the sup-norm

$$\|\mathbf{a}\|_\infty = \max(|a_1|, \dots, |a_m|).$$

For real numbers ( $m = 1$ ) these norms coincide.

## A.23 Matrix Norms

Two common norms used for matrix spaces are the **Frobenius norm** and the **spectral norm**. We can write either as  $\|\mathbf{A}\|$ , but may write  $\|\mathbf{A}\|_F$  or  $\|\mathbf{A}\|_2$  when we want to be specific.

The **Frobenius norm** of an  $m \times k$  matrix  $\mathbf{A}$  is the Euclidean norm applied to its elements

$$\begin{aligned} \|\mathbf{A}\|_F &= \|\text{vec}(\mathbf{A})\| \\ &= (\text{tr}(\mathbf{A}'\mathbf{A}))^{1/2} \\ &= \left( \sum_{i=1}^m \sum_{j=1}^k a_{ij}^2 \right)^{1/2}. \end{aligned}$$

When  $m \times m \mathbf{A}$  is real symmetric then

$$\|\mathbf{A}\|_F = \left( \sum_{\ell=1}^m \lambda_\ell^2 \right)^{1/2}$$

where  $\lambda_\ell$ ,  $\ell = 1, \dots, m$  are the eigenvalues of  $\mathbf{A}$ . To see this, by the spectral decomposition  $\mathbf{A} = \mathbf{H}\Lambda\mathbf{H}'$  with  $\mathbf{H}'\mathbf{H} = \mathbf{I}$  and  $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_m\}$ , so

$$\|\mathbf{A}\|_F = (\text{tr}(\mathbf{H}\Lambda\mathbf{H}'\mathbf{H}\Lambda\mathbf{H}'))^{1/2} = (\text{tr}(\Lambda\Lambda))^{1/2} = \left( \sum_{\ell=1}^m \lambda_\ell^2 \right)^{1/2}. \quad (\text{A.15})$$

A useful calculation is for any  $m \times 1$  vectors  $\mathbf{a}$  and  $\mathbf{b}$ , using (A.1),

$$\|\mathbf{ab}'\|_F = \text{tr}(\mathbf{ba}'\mathbf{ab}')^{1/2} = (\mathbf{b}'\mathbf{ba}'\mathbf{a})^{1/2} = \|\mathbf{a}\| \|\mathbf{b}\|$$

and in particular

$$\|\mathbf{aa}'\|_F = \|\mathbf{a}\|^2. \quad (\text{A.16})$$

The **spectral norm** of an  $m \times k$  real matrix  $\mathbf{A}$  is its largest singular value

$$\|\mathbf{A}\|_2 = s_{\max}(\mathbf{A}) = (\lambda_{\max}(\mathbf{A}'\mathbf{A}))^{1/2}$$

where  $\lambda_{\max}(\mathbf{B})$  denotes the largest eigenvalue of the matrix  $\mathbf{B}$ . Notice that

$$\lambda_{\max}(\mathbf{A}'\mathbf{A}) = \|\mathbf{A}'\mathbf{A}\|_2$$

so

$$\|\mathbf{A}\|_2 = \|\mathbf{A}'\mathbf{A}\|_2^{1/2}.$$

If  $\mathbf{A}$  is  $m \times m$  and symmetric with eigenvalues  $\lambda_j$  then

$$\|\mathbf{A}\|_2 = \max_{j \leq m} |\lambda_j|.$$

The Frobenius and spectral norms are closely related. They are equivalent when applied to a matrix of rank 1, since  $\|\mathbf{ab}'\|_2 = \|\mathbf{a}\| \|\mathbf{b}\| = \|\mathbf{ab}'\|_F$ . In general, for  $m \times k$  matrix  $\mathbf{A}$  with rank  $r$

$$\|\mathbf{A}\|_2 = (\lambda_{\max}(\mathbf{A}'\mathbf{A}))^{1/2} \leq \left( \sum_{j=1}^k \lambda_j(\mathbf{A}'\mathbf{A}) \right)^{1/2} = \|\mathbf{A}\|_F. \quad (\text{A.17})$$

Since  $\mathbf{A}'\mathbf{A}$  also has rank at most  $r$ , it has at most  $r$  non-zero eigenvalues, and hence

$$\|\mathbf{A}\|_F = \left( \sum_{j=1}^k \lambda_j(\mathbf{A}'\mathbf{A}) \right)^{1/2} = \left( \sum_{j=1}^r \lambda_j(\mathbf{A}'\mathbf{A}) \right)^{1/2} \leq (r \lambda_{\max}(\mathbf{A}'\mathbf{A}))^{1/2} = \sqrt{r} \|\mathbf{A}\|_2. \quad (\text{A.18})$$

Given any vector norm  $\|\mathbf{a}\|$  the **induced matrix norm** is defined as

$$\|\mathbf{A}\| = \sup_{\mathbf{x}'\mathbf{x}=1} \|\mathbf{Ax}\| = \sup_{\mathbf{x} \neq 0} \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|}.$$

To see that this is a norm we need to check that it satisfies the triangle inequality. Indeed

$$\|\mathbf{A} + \mathbf{B}\| = \sup_{\mathbf{x}'\mathbf{x}=1} \|\mathbf{Ax} + \mathbf{Bx}\| \leq \sup_{\mathbf{x}'\mathbf{x}=1} \|\mathbf{Ax}\| + \sup_{\mathbf{x}'\mathbf{x}=1} \|\mathbf{Bx}\| = \|\mathbf{A}\| + \|\mathbf{B}\|.$$

For any vector  $\mathbf{x}$ , by the definition of the induced norm

$$\|\mathbf{Ax}\| \leq \|\mathbf{A}\| \|\mathbf{x}\|$$

a property which is called consistent norms.

Let  $\mathbf{A}$  and  $\mathbf{B}$  be conformable and  $\|\mathbf{A}\|$  an induced matrix norm. Then using the property of consistent norms

$$\|\mathbf{AB}\| = \sup_{\mathbf{x}'\mathbf{x}=1} \|\mathbf{ABx}\| \leq \sup_{\mathbf{x}'\mathbf{x}=1} \|\mathbf{A}\| \|\mathbf{Bx}\| = \|\mathbf{A}\| \|\mathbf{B}\|.$$

A matrix norm which satisfies this property is called a **sub-multiplicative norm**, and is a matrix form of the Schwarz inequality.

Of particular interest, the matrix norm induced by the Euclidean vector norm is the spectral norm. Indeed,

$$\sup_{\mathbf{x}'\mathbf{x}=1} \|\mathbf{Ax}\|^2 = \sup_{\mathbf{x}'\mathbf{x}=1} \mathbf{x}'\mathbf{A}'\mathbf{Ax} = \lambda_{\max}(\mathbf{A}'\mathbf{A}) = \|\mathbf{A}\|_2^2.$$

It follows that the spectral norm is consistent with the Euclidean norm, and is sub-multiplicative.

## Appendix B

# Useful Inequalities

In this Appendix, we list a set of inequalities and bounds which are used frequently in econometric theory, predominantly in asymptotic analysis. The proofs are presented in Section B.5.

### B.1 Inequalities for Real Numbers

**Triangle Inequality.** For any real numbers  $x_j$

$$\left| \sum_{j=1}^m x_j \right| \leq \sum_{j=1}^m |x_j|. \quad (\text{B.1})$$

**Jensen's Inequality.** If  $g(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$  is convex, then for any non-negative weights  $a_j$  such that  $\sum_{j=1}^m a_j = 1$ , and any real numbers  $x_j$

$$g\left(\sum_{j=1}^m a_j x_j\right) \leq \sum_{j=1}^m a_j g(x_j). \quad (\text{B.2})$$

In particular, setting  $a_j = 1/m$ , then

$$g\left(\frac{1}{m} \sum_{j=1}^m x_j\right) \leq \frac{1}{m} \sum_{j=1}^m g(x_j). \quad (\text{B.3})$$

If  $g(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$  is concave then the inequalities in (B.2) and (B.3) are reversed.

**Geometric Mean Inequality.** For any non-negative real weights  $a_j$  such that  $\sum_{j=1}^m a_j = 1$ , and any non-negative real numbers  $x_j$

$$x_1^{a_1} x_2^{a_2} \cdots x_m^{a_m} \leq \sum_{j=1}^m a_j x_j. \quad (\text{B.4})$$

**Loève's  $c_r$  Inequality.** For any real numbers  $x_j$ , if  $0 < r \leq 1$

$$\left| \sum_{j=1}^m x_j \right|^r \leq \sum_{j=1}^m |x_j|^r \quad (\text{B.5})$$

and if  $r \geq 1$

$$\left| \sum_{j=1}^m x_j \right|^r \leq m^{r-1} \sum_{j=1}^m |x_j|^r. \quad (\text{B.6})$$

For the important special case  $m = 2$  we can combine these two inequalities as

$$|a + b|^r \leq C_r (|a|^r + |b|^r) \quad (\text{B.7})$$

where  $C_r = \max[1, 2^{r-1}]$ .

**Norm Monotonicity.** If  $0 < t \leq s$ , and any real numbers  $x_j$

$$\left| \sum_{j=1}^m |x_j|^s \right|^{1/s} \leq \left| \sum_{j=1}^m |x_j|^t \right|^{1/t}. \quad (\text{B.8})$$

## B.2 Inequalities for Vectors

**Triangle Inequality.** If  $\mathbf{a} = (a_1, \dots, a_m)'$

$$\|\mathbf{a}\| \leq \sum_{j=1}^m |a_j|. \quad (\text{B.9})$$

**$c_2$  Inequality.** For any  $m \times 1$  vectors  $\mathbf{a}$  and  $\mathbf{b}$ ,

$$(\mathbf{a} + \mathbf{b})' (\mathbf{a} + \mathbf{b}) \leq 2\mathbf{a}' \mathbf{a} + 2\mathbf{b}' \mathbf{b}. \quad (\text{B.10})$$

**Hölder's Inequality.** If  $p > 1$ ,  $q > 1$ , and  $1/p + 1/q = 1$ , then for any  $m \times 1$  vectors  $\mathbf{a}$  and  $\mathbf{b}$ ,

$$|\mathbf{a}' \mathbf{b}| \leq \|\mathbf{a}\|_p \|\mathbf{b}\|_q. \quad (\text{B.11})$$

**Schwarz Inequality.** For any  $m \times 1$  vectors  $\mathbf{a}$  and  $\mathbf{b}$ ,

$$|\mathbf{a}' \mathbf{b}| \leq \|\mathbf{a}\| \|\mathbf{b}\|. \quad (\text{B.12})$$

**Minkowski's Inequality.** For any  $m \times 1$  vectors  $\mathbf{a}$  and  $\mathbf{b}$ , if  $p \geq 1$ , then

$$\|\mathbf{a} + \mathbf{b}\|_p \leq \|\mathbf{a}\|_p + \|\mathbf{b}\|_p. \quad (\text{B.13})$$

**Triangle Inequality.** For any  $m \times 1$  vectors  $\mathbf{a}$  and  $\mathbf{b}$ ,

$$\|\mathbf{a} + \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|. \quad (\text{B.14})$$

## B.3 Inequalities for Matrices

**Schwarz Matrix Inequality:** For any  $m \times k$  and  $k \times m$  matrices  $\mathbf{A}$  and  $\mathbf{B}$ , and either the Frobenius or spectral norm,

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\| \|\mathbf{B}\|. \quad (\text{B.15})$$

**Triangle Inequality:** For any  $m \times k$  matrices  $\mathbf{A}$  and  $\mathbf{B}$ , and either the Frobenius or spectral norm,

$$\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|. \quad (\text{B.16})$$

**Trace Inequality.** For any  $m \times m$  matrices  $\mathbf{A}$  and  $\mathbf{B}$  such that  $\mathbf{A}$  is symmetric and  $\mathbf{B} \geq 0$

$$\text{tr}(\mathbf{AB}) \leq \|\mathbf{A}\|_2 \text{tr}(\mathbf{B}). \quad (\text{B.17})$$

**Quadratic Inequality.** For any  $m \times 1$   $\mathbf{b}$  and  $m \times m$  symmetric matrix  $\mathbf{A}$

$$\mathbf{b}' \mathbf{Ab} \leq \|\mathbf{A}\|_2 \mathbf{b}' \mathbf{b}. \quad (\text{B.18})$$

**Strong Schwarz Matrix Inequality.** For any conformable matrices  $\mathbf{A}$  and  $\mathbf{B}$

$$\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_F. \quad (\text{B.19})$$

**Norm Equivalence.** For any  $m \times k$  matrix  $\mathbf{A}$  of rank  $r$

$$\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F \leq \sqrt{r} \|\mathbf{A}\|_2. \quad (\text{B.20})$$

**Eigenvalue Product Inequality.** For any  $m \times m$  real symmetric matrices  $\mathbf{A} \geq 0$  and  $\mathbf{B} \geq 0$ , the eigenvalues  $\lambda_\ell(\mathbf{AB})$  are real and satisfy

$$\lambda_{\min}(\mathbf{A}) \lambda_{\min}(\mathbf{B}) \leq \lambda_\ell(\mathbf{AB}) \leq \lambda_{\max}(\mathbf{A}) \lambda_{\max}(\mathbf{B}). \quad (\text{B.21})$$

(Zhang and Zhang, 2006, Corollary 11).

## B.4 Probability Inequalities

**Monotone Probability Inequality.** For any events  $A$  and  $B$  such that  $A \subset B$ ,

$$\mathbb{P}(A) \leq \mathbb{P}(B). \quad (\text{B.22})$$

**Union Equality.** For any events  $A$  and  $B$ ,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B). \quad (\text{B.23})$$

**Boole's Inequality (Union Bound).** For any events  $A$  and  $B$ ,

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B). \quad (\text{B.24})$$

**Bonferroni's Inequality.** For any events  $A$  and  $B$ ,

$$\mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1. \quad (\text{B.25})$$

**Jensen's Inequality.** If  $g(\cdot) : \mathbb{R}^m \rightarrow \mathbb{R}$  is convex, then for any random vector  $\mathbf{x}$  for which  $\mathbb{E}\|\mathbf{x}\| < \infty$  and  $\mathbb{E}|g(\mathbf{x})| < \infty$ ,

$$g(\mathbb{E}(\mathbf{x})) \leq \mathbb{E}(g(\mathbf{x})). \quad (\text{B.26})$$

If  $g(\cdot)$  is concave the inequality is reversed.

**Conditional Jensen's Inequality.** If  $g(\cdot) : \mathbb{R}^m \rightarrow \mathbb{R}$  is convex, then for any random vectors  $(\mathbf{y}, \mathbf{x})$  for which  $\mathbb{E}\|\mathbf{y}\| < \infty$  and  $\mathbb{E}\|g(\mathbf{y})\| < \infty$ ,

$$g(\mathbb{E}(\mathbf{y} | \mathbf{x})) \leq \mathbb{E}(g(\mathbf{y}) | \mathbf{x}). \quad (\text{B.27})$$

If  $g(\cdot)$  is concave the inequality is reversed.

**Conditional Expectation Inequality.** For any  $r \geq 1$  such that  $\mathbb{E}|y|^r < \infty$ , then

$$\mathbb{E}(|\mathbb{E}(y | \mathbf{x})|^r) \leq \mathbb{E}|y|^r < \infty. \quad (\text{B.28})$$

**Expectation Inequality.** If  $\mathbb{E}\|\mathbf{Y}\| < \infty$  then

$$\|\mathbb{E}(\mathbf{Y})\| \leq \mathbb{E}\|\mathbf{Y}\|. \quad (\text{B.29})$$

**Hölder's Inequality.** If  $p > 1$  and  $q > 1$  and  $\frac{1}{p} + \frac{1}{q} = 1$ , then for any random  $m \times n$  matrices  $\mathbf{X}$  and  $\mathbf{Y}$ ,

$$\mathbb{E}\|\mathbf{X}'\mathbf{Y}\| \leq (\mathbb{E}(\|\mathbf{X}\|^p))^{1/p} (\mathbb{E}(\|\mathbf{Y}\|^q))^{1/q}. \quad (\text{B.30})$$

**Cauchy-Schwarz Inequality.** For any random  $m \times n$  matrices  $\mathbf{X}$  and  $\mathbf{Y}$ ,

$$\mathbb{E}\|\mathbf{X}'\mathbf{Y}\| \leq (\mathbb{E}(\|\mathbf{X}\|^2))^{1/2} (\mathbb{E}(\|\mathbf{Y}\|^2))^{1/2}. \quad (\text{B.31})$$

**Matrix Cauchy-Schwarz Inequality.** Tripathi (1999). For any random  $\mathbf{x} \in \mathbb{R}^m$  and  $\mathbf{y} \in \mathbb{R}^l$ ,

$$\mathbb{E}(\mathbf{y}\mathbf{x}') (\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-} \mathbb{E}(\mathbf{x}\mathbf{y}') \leq \mathbb{E}(\mathbf{y}\mathbf{y}'). \quad (\text{B.32})$$

**Minkowski's Inequality.** For any random  $m \times n$  matrices  $\mathbf{X}$  and  $\mathbf{Y}$ ,

$$(\mathbb{E}(\|\mathbf{X} + \mathbf{Y}\|^p))^{1/p} \leq (\mathbb{E}(\|\mathbf{X}\|^p))^{1/p} + (\mathbb{E}(\|\mathbf{Y}\|^p))^{1/p}. \quad (\text{B.33})$$

**Liapunov's Inequality.** For any random  $m \times n$  matrix  $\mathbf{X}$  and  $0 < r \leq p$ ,

$$(\mathbb{E}(\|\mathbf{X}\|^r))^{1/r} \leq (\mathbb{E}(\|\mathbf{X}\|^p))^{1/p}. \quad (\text{B.34})$$

**Markov's Inequality (standard form).** For any random vector  $\mathbf{x}$ , non-negative function  $g(\mathbf{x}) \geq 0$ , and  $\varepsilon > 0$

$$\mathbb{P}(g(\mathbf{x}) > \varepsilon) \leq \varepsilon^{-1} \mathbb{E}(g(\mathbf{x})). \quad (\text{B.35})$$

**Markov's Inequality (strong form).** For any random vector  $\mathbf{x}$ , non-negative function  $g(\mathbf{x}) \geq 0$ , and  $\varepsilon > 0$

$$\mathbb{P}(g(\mathbf{x}) > \varepsilon) \leq \varepsilon^{-1} \mathbb{E}(g(\mathbf{x}) \mathbf{1}(g(\mathbf{x}) > \varepsilon)). \quad (\text{B.36})$$

**Chebyshev's Inequality.** For any random variable  $x$  and  $\varepsilon > 0$

$$\mathbb{P}(|x - \mathbb{E}(x)| > \varepsilon) \leq \varepsilon^{-2} \text{var}(x). \quad (\text{B.37})$$

**Bernstein's Inequality.** If  $x_i$  are independent random variables,  $\mathbb{E}(x_i) = 0$ ,  $\bar{\sigma}^2 = \sum_{i=1}^n \mathbb{E}(x_i^2)$  and  $|x_i| \leq M < \infty$ , then for all  $\varepsilon > 0$

$$\mathbb{P}\left(\left|\sum_{i=1}^n x_i\right| > \varepsilon\right) \leq 2 \exp\left(-\frac{\varepsilon^2}{2\bar{\sigma}^2 + 2M\varepsilon/3}\right). \quad (\text{B.38})$$

**Bernstein's Inequality for vectors.** If  $\mathbf{x}_i = (x_{1i}, \dots, x_{mi})'$  are independent random vectors,  $\mathbb{E}(\mathbf{x}_i) = 0$ ,  $\bar{\sigma}^2 = \max_j \sum_{i=1}^n \mathbb{E}(x_{ji}^2)$  and  $|x_{ji}| \leq M < \infty$ , then for all  $\varepsilon > 0$

$$\mathbb{P}\left(\left|\sum_{i=1}^n \mathbf{x}_i\right| > \varepsilon\right) \leq 2m \exp\left(-\frac{\varepsilon^2}{2m^2\bar{\sigma}^2 + 2mM\varepsilon/3}\right). \quad (\text{B.39})$$

**Sub-Gaussian Inequality.** If  $x_i$  are independent random variables,  $\mathbb{E}(x_i) = 0$ , and  $|x_i| \leq a_i$ , then for any  $\lambda > 0$

$$\mathbb{E}\left(\exp\left(\lambda \sum_{i=1}^n x_i\right)\right) \leq \exp\left(\frac{\lambda^2 \sum_{i=1}^n a_i^2}{2}\right). \quad (\text{B.40})$$

**Hoeffding's Inequality.** If  $x_i$  are independent random variables,  $\mathbb{E}(x_i) = 0$ , and  $|x_i| \leq a_i$ , then for any  $\varepsilon > 0$

$$\mathbb{P}\left(\sum_{i=1}^n x_i > \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2}{2\sum_{i=1}^n a_i^2}\right), \quad (\text{B.41})$$

**Simple Inequality.** If  $X$  and  $Y$  are independent and mean zero then

$$\mathbb{E}|X + Y|^r \leq B_r (\mathbb{E}|X|^r + \mathbb{E}|Y|^r). \quad (\text{B.42})$$

where  $B_r = 1$  for  $0 < r \leq 1$ ,  $B_r = 2^{r-1}$  for  $1 < r \leq 2$ ,  $B_r = 2$  for  $2 \leq r \leq 3$ , and  $B_r = 2^{r-2}$  for  $r > 3$ .

**Difference Inequality.** If  $X$  and  $Y$  are independent and identically distributed then for any  $0 < r \leq 2$

$$\mathbb{E}|X - Y|^r \leq 2\mathbb{E}|X|^r. \quad (\text{B.43})$$

**Bahr-Esseen Inequality.** If  $x_i$  are independent and  $\mathbb{E}(x_i) = 0$ , then for any  $0 < r \leq 2$

$$\mathbb{E} \left| \sum_{i=1}^n x_i \right|^r \leq 2 \sum_{i=1}^n \mathbb{E} |x_i|^r. \quad (\text{B.44})$$

Some of the following inequalities make use of **Rademacher** random variables  $\varepsilon_i$  which are two-point discrete variables, independent of the observations, satisfying

$$\begin{aligned} \mathbb{P}(\varepsilon_i = 1) &= \frac{1}{2} \\ \mathbb{P}(\varepsilon_i = -1) &= \frac{1}{2}. \end{aligned}$$

**Symmetrization Inequality.** If  $x_i$  are independent and  $\mathbb{E}(x_i) = 0$ , then for any  $r \geq 1$

$$\mathbb{E} \left| \sum_{i=1}^n x_i \right|^r \leq D_r \mathbb{E} \left| \sum_{i=1}^n x_i \varepsilon_i \right|^r \quad (\text{B.45})$$

where  $\varepsilon_i$  are independent Rademacher random variables (independent of  $x_i$ ),  $D_r = 2$  for  $1 \leq r \leq 2$ , and  $D_r = 2^r$  for  $r > 2$ .

**Khintchine's Inequality.** If  $\varepsilon_i$  are independent Rademacher random variables, then for any  $r > 0$  and any real numbers  $a_i$

$$\mathbb{E} \left| \sum_{i=1}^n a_i \varepsilon_i \right|^r \leq K_r \left( \sum_{i=1}^n a_i^2 \right)^{r/2}. \quad (\text{B.46})$$

where  $K_r = 1$  for  $r \leq 2$  and  $K_r = 2^{r/2} \Gamma((r+1)/2) / \pi^{1/2}$  for  $r \geq 2$ .

**Marcinkiewicz-Zygmund Inequality.** If  $x_i$  are independent and  $\mathbb{E}(x_i) = 0$  then for any  $r \geq 1$

$$\mathbb{E} \left| \sum_{i=1}^n x_i \right|^r \leq M_r \mathbb{E} \left| \sum_{i=1}^n x_i^2 \right|^{r/2}. \quad (\text{B.47})$$

where  $M_r = D_r K_r$  with  $D_r$  and  $K_r$  from the symmetrization and Khintchine inequalities.

**Whittle's Inequalities.** (Whittle, 1960) If  $x_i$  are independent,  $\mathbb{E}(x_i) = 0$  and for some  $r \geq 2$ ,  $\mathbb{E}|x_i|^r \leq B_{1r} < \infty$ , then there is a constant  $C_{1r} < \infty$  such that for any real numbers  $a_i$

$$\mathbb{E} \left| \sum_{i=1}^n a_i x_i \right|^r \leq C_{1r} \left| \sum_{i=1}^n a_i^2 \right|^{r/2}. \quad (\text{B.48})$$

Furthermore, if  $\mathbb{E}|x_i|^{2r} \leq B_{2r} < \infty$ , then there is a constant  $C_{2r} < \infty$  such that for any real  $n \times n$  matrix  $A$  and writing  $\mathbf{x} = (x_1, \dots, x_n)'$ ,

$$\mathbb{E} |\mathbf{x}' A \mathbf{x} - \mathbb{E}(\mathbf{x}' A \mathbf{x})|^r \leq C_{2r} \operatorname{tr}(A' A)^{r/2}. \quad (\text{B.49})$$

Our proof shows that we can set  $C_{1r} = M_r B_{1r}$  and  $C_{2r} = 4^r K_r C_{1r}^{1/2} B_{2r}^{1/2}$  where  $M_r$  and  $K_r$  are from the Marcinkiewicz-Zygmund and Khinchine inequalities.

**Rosenthal's Inequality.** For any  $r > 0$  there is a constant  $R_r < \infty$  such that if  $x_i$  are independent and  $\mathbb{E}(x_i) = 0$  then

$$\mathbb{E} \left| \sum_{i=1}^n x_i \right|^r \leq R_r \left\{ \left( \sum_{i=1}^n \mathbb{E}(x_i^2) \right)^{r/2} + \sum_{i=1}^n \mathbb{E}|x_i|^r \right\}. \quad (\text{B.50})$$

Our proof establishes that (B.50) holds with  $R_r = 2^{r(r-2)/8} M_r$  where  $M_r$  is from the Marcinkiewicz-Zygmund inequality.

For a generalization of Rosenthal's inequality to the matrix case see B. Hansen (2015).

**Maximal Inequality.** For any  $r \geq 1$ , if  $x_i$  are independent,  $\mathbb{E}(x_i) = 0$ , and  $\mathbb{E}|x_i|^r < \infty$ , then for all  $\varepsilon > 0$

$$\mathbb{P}\left(\max_{1 \leq j \leq n} \left| \sum_{i=1}^j x_i \right| > \varepsilon\right) \leq \varepsilon^{-r} \mathbb{E}\left|\sum_{i=1}^n x_i\right|^r. \quad (\text{B.51})$$

**Kolmogorov's Inequality.** If  $x_i$  are independent,  $\mathbb{E}(x_i) = 0$ , and  $\mathbb{E}(x_i^2) < \infty$ , then for all  $\varepsilon > 0$

$$\mathbb{P}\left(\max_{1 \leq j \leq n} \left| \sum_{i=1}^j x_i \right| > \varepsilon\right) \leq \varepsilon^{-2} \sum_{i=1}^n \mathbb{E}(x_i^2). \quad (\text{B.52})$$

**Doob's Inequality.** For any  $r > 1$ , if  $x_i$  are independent,  $\mathbb{E}(x_i) = 0$ , and  $\mathbb{E}|x_i|^r < \infty$ , then

$$\mathbb{E}\left(\max_{1 \leq j \leq n} \left| \sum_{i=1}^j x_i \right|^r\right) \leq \left(\frac{r}{r-1}\right)^r \mathbb{E}\left|\sum_{i=1}^n x_i\right|^r. \quad (\text{B.53})$$

**Ottaviani's Inequality.** Set  $S_i = \sum_{j=1}^i x_j$ . If  $x_i$  are independent, then for any  $\varepsilon > 0$

$$\mathbb{P}\left(\max_{1 \leq i \leq n} |S_i| > \varepsilon\right) \leq \frac{\mathbb{P}(|S_n| > \varepsilon/2)}{1 - \max_{1 \leq i \leq n} \mathbb{P}(|S_n - S_i| > \varepsilon/2)}. \quad (\text{B.54})$$

## B.5 Proofs\*

**Proof of Triangle Inequality (B.1).** Take the case  $m = 2$ . Observe that

$$\begin{aligned} -|x_1| &\leq x_1 \leq |x_1| \\ -|x_2| &\leq x_2 \leq |x_2|. \end{aligned}$$

Adding, we find

$$-|x_1| - |x_2| \leq x_1 + x_2 \leq |x_1| + |x_2|$$

which is (B.1) for  $m = 2$ . For  $m > 2$ , we apply (B.1)  $m - 1$  times. ■

**Proof of Jensen's Inequality (B.2).** By the definition of convexity, for any  $\lambda \in [0, 1]$

$$g(\lambda x_1 + (1 - \lambda) x_2) \leq \lambda g(x_1) + (1 - \lambda) g(x_2). \quad (\text{B.55})$$

This implies

$$\begin{aligned} g\left(\sum_{j=1}^m a_j x_j\right) &= g\left(a_1 x_1 + (1 - a_1) \sum_{j=2}^m \frac{a_j}{1 - a_1} x_j\right) \\ &\leq a_1 g(x_1) + (1 - a_1) g\left(\sum_{j=2}^m b_j x_j\right) \end{aligned}$$

where  $b_j = a_j / (1 - a_1)$  and  $\sum_{j=2}^m b_j = 1$ . By another application of (B.55) this is bounded by

$$\begin{aligned} a_1 g(x_1) + (1 - a_1) &\left( b_2 g(x_2) + (1 - b_2) g\left(\sum_{j=2}^m c_j x_j\right) \right) \\ &= a_1 g(x_1) + a_2 g(x_2) + (1 - a_1)(1 - b_2) g\left(\sum_{j=2}^m c_j x_j\right) \end{aligned}$$

where  $c_j = b_j / (1 - b_2)$ . By repeated application of (B.55) we obtain (B.2). ■

**Proof of Geometric Mean Inequality (B.4).** Since the logarithm is strictly concave, by Jensen's inequality (B.2)

$$\log(x_1^{a_1} x_2^{a_2} \cdots x_m^{a_m}) = \sum_{j=1}^m a_j \log x_j \leq \log\left(\sum_{j=1}^m a_j x_j\right).$$

Applying the exponential yields (B.4). ■

**Proof of Loève's  $c_r$  Inequality (B.5).** For  $r \geq 1$  this is a rewriting of Jensen's inequality (B.3) with  $g(u) = u^r$ . For  $r < 1$ , define  $b_j = |x_j| / (\sum_{j=1}^m |x_j|)$ . The facts that  $0 \leq b_j \leq 1$  and  $r < 1$  imply  $b_j \leq b_j^r$  and thus

$$1 = \sum_{j=1}^m b_j \leq \sum_{j=1}^m b_j^r.$$

This implies

$$\left(\sum_{j=1}^m x_j\right)^r \leq \left(\sum_{j=1}^m |x_j|\right)^r \leq \sum_{j=1}^m |x_j|^r.$$

■

**Proof of Norm Monotonicity (B.8).** Set  $y_i = |x_j|^s$  and  $r = t/s \leq 1$ . The  $c_r$  inequality (B.5) implies  $\left|\sum_{j=1}^m y_j\right|^r \leq \sum_{j=1}^m y_j^r$  or

$$\left|\sum_{j=1}^m |x_j|^s\right|^{t/s} \leq \sum_{j=1}^m |x_j|^t.$$

Raising both sides to the power  $1/t$  yields (B.8). ■

**Proof of Triangle Inequality (B.9).** Apply the  $c_r$  inequality (B.5) with  $r = 1/2$  to find

$$\|\mathbf{a}\| = \left| \sum_{j=1}^m a_j^2 \right|^{1/2} \leq \sum_{j=1}^m |a_j|.$$

■

**Proof of  $c_2$  Inequality (B.10).** By the  $c_r$  inequality (B.6) with  $r = 2$ ,  $(a_j + b_j)^2 \leq 2a_j^2 + 2b_j^2$ . Thus

$$\begin{aligned} (\mathbf{a} + \mathbf{b})' (\mathbf{a} + \mathbf{b}) &= \sum_{j=1}^m (a_j + b_j)^2 \\ &\leq 2 \sum_{j=1}^m a_j^2 + 2 \sum_{j=1}^m b_j^2 \\ &= 2\mathbf{a}'\mathbf{a} + 2\mathbf{b}'\mathbf{b}. \end{aligned}$$

■

**Proof of Hölder's Inequality (B.11).** Set  $u_j = |a_j|^p / \|\mathbf{a}\|_p^p$  and  $v_j = |b_j|^q / \|\mathbf{b}\|_q^q$  and observe that  $\sum_{j=1}^m u_j = 1$  and  $\sum_{j=1}^m v_j = 1$ . By the weighted geometric mean inequality (B.4),

$$u_j^{1/p} v_j^{1/q} \leq \frac{u_j}{p} + \frac{v_j}{q}.$$

Then since  $\sum_{j=1}^m u_j = 1$ ,  $\sum_{j=1}^m v_j = 1$  and  $1/p + 1/q = 1$

$$\frac{\sum_{j=1}^m |a_j b_j|}{\|\mathbf{a}\|_p \|\mathbf{b}\|_q} = \sum_{j=1}^m u_j^{1/p} v_j^{1/q} \leq \sum_{j=1}^m \left( \frac{u_j}{p} + \frac{v_j}{q} \right) = 1. \quad (\text{B.56})$$

By the Triangle inequality (B.1) and then (B.56)

$$|\mathbf{a}'\mathbf{b}| = \left| \sum_{j=1}^m a_j b_j \right| \leq \sum_{j=1}^m |a_j b_j| \leq \|\mathbf{a}\|_p \|\mathbf{b}\|_q$$

which is (B.11). ■

**Proof of Schwarz Inequality (B.12).** This is a special case of Hölder's inequality (B.11) with  $p = q = 2$ .

$$|\mathbf{a}'\mathbf{b}| \leq \sum_{j=1}^m |a_j b_j| \leq \|\mathbf{a}\| \|\mathbf{b}\|.$$

■

**Proof of Minkowski's Inequality (B.13).** Set  $q = p/(p-1)$  so that  $1/p + 1/q = 1$ . Using the triangle inequality for real numbers (B.1) and two applications of Hölder's inequality (B.11)

$$\begin{aligned} \|\mathbf{a} + \mathbf{b}\|_p^p &= \sum_{j=1}^m |a_j + b_j|^p \\ &= \sum_{j=1}^m |a_j + b_j| |a_j + b_j|^{p-1} \\ &\leq \sum_{j=1}^m |a_j| |a_j + b_j|^{p-1} + \sum_{j=1}^m |b_j| |a_j + b_j|^{p-1} \\ &\leq \|\mathbf{a}\|_p \left( \sum_{j=1}^m |a_j + b_j|^{(p-1)q} \right)^{1/q} + \|\mathbf{b}\|_p \left( \sum_{j=1}^m |a_j + b_j|^{(p-1)q} \right)^{1/q} \\ &= (\|\mathbf{a}\|_p + \|\mathbf{b}\|_p) \|\mathbf{a} + \mathbf{b}\|_p^{p-1} \end{aligned}$$

Solving, we find (B.13). ■

**Proof of Triangle Inequality (B.14).** This is a special case of Minkowski's inequality (B.13) with  $p = 2$ .

■

**Proof of Schwarz Matrix Inequality (B.15).** The inequality holds for the spectral norm since it is an induced norm. Now consider the Frobenius norm. Partition  $\mathbf{A}' = [\mathbf{a}_1, \dots, \mathbf{a}_n]$  and  $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_n]$ . Then by partitioned matrix multiplication, the definition of the Frobenius norm and the Schwarz inequality (B.12)

$$\begin{aligned} \|\mathbf{AB}\|_F &= \left\| \begin{array}{ccc} \mathbf{a}'_1 \mathbf{b}_1 & \mathbf{a}'_1 \mathbf{b}_2 & \cdots \\ \mathbf{a}'_2 \mathbf{b}_1 & \mathbf{a}'_2 \mathbf{b}_2 & \cdots \\ \vdots & \vdots & \ddots \end{array} \right\|_F \\ &\leq \left\| \begin{array}{ccc} \|\mathbf{a}_1\| \|\mathbf{b}_1\| & \|\mathbf{a}_1\| \|\mathbf{b}_2\| & \cdots \\ \|\mathbf{a}_2\| \|\mathbf{b}_1\| & \|\mathbf{a}_2\| \|\mathbf{b}_2\| & \cdots \\ \vdots & \vdots & \ddots \end{array} \right\|_F \\ &= \left( \sum_{i=1}^m \sum_{j=1}^m \|\mathbf{a}_i\|^2 \|\mathbf{b}_j\|^2 \right)^{1/2} \\ &= \left( \sum_{i=1}^m \|\mathbf{a}_i\|^2 \right)^{1/2} \left( \sum_{i=1}^m \|\mathbf{b}_i\|^2 \right)^{1/2} \\ &= \left( \sum_{i=1}^k \sum_{j=1}^m \mathbf{a}_{ji}^2 \right)^{1/2} \left( \sum_{i=1}^m \sum_{j=1}^k \|\mathbf{b}_{ji}\|^2 \right)^{1/2} \\ &= \|\mathbf{A}\|_F \|\mathbf{B}\|_F. \end{aligned}$$

**Proof of Triangle Inequality (B.16).** The inequality holds for the spectral norm since it is an induced norm. Now consider the Frobenius norm. Let  $\mathbf{a} = \text{vec}(\mathbf{A})$  and  $\mathbf{b} = \text{vec}(\mathbf{B})$ . Then by the definition of the Frobenius norm and the triangle inequality (B.14)

$$\begin{aligned}\|\mathbf{A} + \mathbf{B}\|_F &= \|\text{vec}(\mathbf{A} + \mathbf{B})\|_F \\ &= \|\mathbf{a} + \mathbf{b}\| \\ &\leq \|\mathbf{a}\| + \|\mathbf{b}\| \\ &= \|\mathbf{A}\|_F + \|\mathbf{B}\|_F.\end{aligned}$$

**Proof of Trace Inequality (B.17).** By the spectral decomposition for symmetric matrices,  $\mathbf{A} = \mathbf{H}\Lambda\mathbf{H}'$  where  $\Lambda$  has the eigenvalues  $\lambda_j$  of  $\mathbf{A}$  on its diagonal and  $\mathbf{H}$  is orthonormal. Define  $\mathbf{C} = \mathbf{H}'\mathbf{B}\mathbf{H}$  which has non-negative diagonal elements  $C_{jj}$  since  $\mathbf{B}$  is positive semi-definite. Then

$$\text{tr}(\mathbf{AB}) = \text{tr}(\Lambda\mathbf{C}) = \sum_{j=1}^m \lambda_j C_{jj} \leq \max_j |\lambda_j| \sum_{j=1}^m C_{jj} = \|\mathbf{A}\|_2 \text{tr}(\mathbf{C})$$

where the inequality uses the fact that  $C_{jj} \geq 0$ . Note that

$$\text{tr}(\mathbf{C}) = \text{tr}(\mathbf{H}'\mathbf{B}\mathbf{H}) = \text{tr}(\mathbf{H}\mathbf{H}'\mathbf{B}) = \text{tr}(\mathbf{B})$$

since  $\mathbf{H}$  is orthonormal. Thus  $\text{tr}(\mathbf{AB}) \leq \|\mathbf{A}\|_2 \text{tr}(\mathbf{B})$  as stated. ■

**Proof of Quadratic Inequality (B.18).** In the trace inequality (B.17) set  $\mathbf{B} = \mathbf{b}\mathbf{b}'$  and note  $\text{tr}(\mathbf{AB}) = \mathbf{b}'\mathbf{Ab}$  and  $\text{tr}(\mathbf{B}) = \mathbf{b}'\mathbf{b}$ . ■

**Proof of Strong Schwarz Matrix Inequality (B.19).** By the definition of the Frobenius norm, the property of the trace, the trace inequality (B.17) (noting that both  $\mathbf{A}'\mathbf{A}$  and  $\mathbf{B}\mathbf{B}'$  are symmetric and positive semi-definite), and the Schwarz matrix inequality (B.15)

$$\begin{aligned}\|\mathbf{AB}\|_F &= (\text{tr}(\mathbf{B}'\mathbf{A}'\mathbf{AB}))^{1/2} \\ &= (\text{tr}(\mathbf{A}'\mathbf{ABB}'))^{1/2} \\ &\leq (\|\mathbf{A}'\mathbf{A}\|_2 \text{tr}(\mathbf{BB}'))^{1/2} \\ &= \|\mathbf{A}\|_2 \|\mathbf{B}\|_F.\end{aligned}$$

**Proof of Norm Equivalence (B.20).** The first inequality was established in (A.17), and the second in (A.18). ■

**Proof of Monotone Probability Inequality (B.22).** Since  $A \subset B$  then  $B = A \cup \{B \cap A^c\}$  where  $A^c$  is the complement of  $A$ . The sets  $A$  and  $\{B \cap A^c\}$  are disjoint. Thus

$$\mathbb{P}(B) = \mathbb{P}(A \cup \{B \cap A^c\}) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c) \geq \mathbb{P}(A)$$

since probabilities are non-negative. Thus  $\mathbb{P}(A) \leq \mathbb{P}(B)$  as claimed. ■

**Proof of Union Equality (B.23).**  $\{A \cup B\} = A \cup \{B \cap A^c\}$  where  $A$  and  $\{B \cap A^c\}$  are disjoint. Also  $B = \{B \cap A\} \cup \{B \cap A^c\}$  where  $\{B \cap A\}$  and  $\{B \cap A^c\}$  are disjoint. These two relationships imply

$$\begin{aligned}\mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(B \cap A^c) \\ \mathbb{P}(B) &= \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c).\end{aligned}$$

Substracting,

$$\mathbb{P}(A \cup B) - \mathbb{P}(B) = \mathbb{P}(A) - \mathbb{P}(B \cap A)$$

which is (B.23) upon rearrangement. ■

**Proof of Boole's Inequality (B.24).** From the union equality (B.23) and  $\mathbb{P}(A \cap B) \geq 0$ ,

$$\begin{aligned}\mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \\ &\leq \mathbb{P}(A) + \mathbb{P}(B)\end{aligned}$$

as claimed. ■

**Proof of Bonferroni's Inequality (B.25).** Rearranging the union equality (B.23) and using  $\mathbb{P}(A \cup B) \leq 1$

$$\begin{aligned}\mathbb{P}(A \cap B) &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B) \\ &\geq \mathbb{P}(A) + \mathbb{P}(B) - 1\end{aligned}$$

which is (B.25). ■

**Proof of Jensen's Inequality (B.26).** Since  $g(\mathbf{u})$  is convex, at any point  $\mathbf{u}$  there is a nonempty set of subderivatives (linear surfaces touching  $g(\mathbf{u})$  at  $\mathbf{u}$  but lying below  $g(\mathbf{u})$  for all  $\mathbf{u}$ ). Let  $a + \mathbf{b}'\mathbf{u}$  be a subderivative of  $g(\mathbf{u})$  at  $\mathbf{u} = \mathbb{E}(\mathbf{x})$ . Then for all  $\mathbf{u}$ ,  $g(\mathbf{u}) \geq a + \mathbf{b}'\mathbf{u}$  yet  $g(\mathbb{E}(\mathbf{x})) = a + \mathbf{b}'\mathbb{E}(\mathbf{x})$ . Applying expectations,  $\mathbb{E}(g(\mathbf{x})) \geq a + \mathbf{b}'\mathbb{E}(\mathbf{x}) = g(\mathbb{E}(\mathbf{x}))$ , as stated. ■

**Proof of Conditional Jensen's Inequality (B.27).** The same as the proof of Jensen's inequality (B.26), but using conditional expectations. The conditional expectations exist since  $\mathbb{E}\|y\| < \infty$  and  $\mathbb{E}\|g(y)\| < \infty$ . ■

**Proof of Conditional Expectation Inequality (B.28).** As the function  $|u|^r$  is convex for  $r \geq 1$ , the conditional Jensen's inequality (B.27) implies

$$|\mathbb{E}(y | \mathbf{x})|^r \leq \mathbb{E}(|y|^r | \mathbf{x}).$$

Taking unconditional expectations and the law of iterated expectations, we obtain

$$\mathbb{E}(|\mathbb{E}(y | \mathbf{x})|^r) \leq \mathbb{E}(\mathbb{E}(|y|^r | \mathbf{x})) = \mathbb{E}(|y|^r) < \infty$$

as required. ■

**Proof of Expectation Inequality (B.29).** By the triangle inequality (B.16), for  $\lambda \in [0, 1]$ ,

$$\|\lambda \mathbf{U}_1 + (1 - \lambda) \mathbf{U}_2\| \leq \lambda \|\mathbf{U}_1\| + (1 - \lambda) \|\mathbf{U}_2\|$$

which shows that the matrix norm  $g(\mathbf{U}) = \|\mathbf{U}\|$  is convex. Applying Jensen's inequality (B.26) we find (B.29). ■

**Proof of Hölder's Inequality (B.30).** Since  $\frac{1}{p} + \frac{1}{q} = 1$  an application of Jensen's inequality (B.2) shows that for any real  $a$  and  $b$

$$\exp\left[\frac{1}{p}a + \frac{1}{q}b\right] \leq \frac{1}{p}\exp(a) + \frac{1}{q}\exp(b).$$

Setting  $u = \exp(a)$  and  $v = \exp(b)$  this implies

$$u^{1/p}v^{1/q} \leq \frac{u}{p} + \frac{v}{q}$$

and this inequality holds for any  $u > 0$  and  $v > 0$ .

Set  $u = \|\mathbf{X}\|^p / \mathbb{E}(\|\mathbf{X}\|^p)$  and  $v = \|\mathbf{Y}\|^q / \mathbb{E}(\|\mathbf{Y}\|^q)$ . Note that  $\mathbb{E}(u) = \mathbb{E}(v) = 1$ . By the Schwarz matrix inequality (B.15),  $\|\mathbf{X}'\mathbf{Y}\| \leq \|\mathbf{X}\| \|\mathbf{Y}\|$ . Thus

$$\begin{aligned} \frac{\mathbb{E}\|\mathbf{X}'\mathbf{Y}\|}{(\mathbb{E}(\|\mathbf{X}\|^p))^{1/p} (\mathbb{E}(\|\mathbf{Y}\|^q))^{1/q}} &\leq \frac{\mathbb{E}(\|\mathbf{X}\| \|\mathbf{Y}\|)}{(\mathbb{E}(\|\mathbf{X}\|^p))^{1/p} (\mathbb{E}(\|\mathbf{Y}\|^q))^{1/q}} \\ &= \mathbb{E}(u^{1/p} v^{1/q}) \\ &\leq \mathbb{E}\left(\frac{u}{p} + \frac{v}{q}\right) \\ &= \frac{1}{p} + \frac{1}{q} \\ &= 1, \end{aligned}$$

which is (B.30). ■

**Proof of Cauchy-Schwarz Inequality (B.31).** Special case of Hölder's inequality (B.30) with  $p = q = 2$ .

**Proof of Matrix Cauchy-Schwarz Inequality (B.32).** Define  $\mathbf{e} = \mathbf{y} - (\mathbb{E}(\mathbf{y}\mathbf{x}')) (\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-} \mathbf{x}$ . Note that  $\mathbb{E}(\mathbf{e}\mathbf{e}') \geq 0$  is positive semi-definite. We can calculate that

$$\mathbb{E}(\mathbf{e}\mathbf{e}') = \mathbb{E}(\mathbf{y}\mathbf{y}') - (\mathbb{E}(\mathbf{y}\mathbf{x}')) (\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-} \mathbb{E}(\mathbf{x}\mathbf{y}').$$

Since the left-hand-side is positive semi-definite, so is the right-hand-side. This means  $\mathbb{E}(\mathbf{y}\mathbf{y}') \geq (\mathbb{E}(\mathbf{y}\mathbf{x}')) (\mathbb{E}(\mathbf{x}\mathbf{x}'))^{-} \mathbb{E}(\mathbf{x}\mathbf{y}')$  as stated. ■

**Proof of Minkowski's Inequality (B.33).** Note that by rewriting, using the triangle inequality (B.16), and then applying Hölder's inequality (B.30) to the two expectations

$$\begin{aligned} \mathbb{E}(\|\mathbf{X} + \mathbf{Y}\|^p) &= \mathbb{E}(\|\mathbf{X} + \mathbf{Y}\| \|\mathbf{X} + \mathbf{Y}\|^{p-1}) \\ &\leq \mathbb{E}(\|\mathbf{X}\| \|\mathbf{X} + \mathbf{Y}\|^{p-1}) + \mathbb{E}(\|\mathbf{Y}\| \|\mathbf{X} + \mathbf{Y}\|^{p-1}) \\ &\leq (\mathbb{E}(\|\mathbf{X}\|^p))^{1/p} \mathbb{E}\left(\left(\|\mathbf{X} + \mathbf{Y}\|^{q(p-1)}\right)^{1/q}\right) \\ &\quad + (\mathbb{E}(\|\mathbf{Y}\|^p))^{1/p} \mathbb{E}\left(\left(\|\mathbf{X} + \mathbf{Y}\|^{q(p-1)}\right)^{1/q}\right) \\ &= \left((\mathbb{E}(\|\mathbf{X}\|^p))^{1/p} + (\mathbb{E}(\|\mathbf{Y}\|^p))^{1/p}\right) \mathbb{E}\left(\left(\|\mathbf{X} + \mathbf{Y}\|^p\right)^{(p-1)/p}\right) \end{aligned}$$

where the second inequality picks  $q$  to satisfy  $1/p + 1/q = 1$ , and the final equality uses this fact to make the substitution  $q = p/(p-1)$  and then collects terms. Dividing both sides by  $\mathbb{E}((\|\mathbf{X} + \mathbf{Y}\|^p)^{(p-1)/p})$ , we obtain (B.33). ■

**Proof of Liapunov's Inequality (B.34).** The function  $g(u) = u^{p/r}$  is convex for  $u > 0$  since  $p \geq r$ . Set  $u = \|\mathbf{X}\|^r$ . By Jensen's inequality (B.26),  $\mathbb{E}(g(u)) \leq g(\mathbb{E}(u))$  or

$$(\mathbb{E}(\|\mathbf{X}\|^r))^{p/r} \leq \mathbb{E}\left(\left(\|\mathbf{X}\|^r\right)^{p/r}\right) = \mathbb{E}(\|\mathbf{X}\|^p).$$

Raising both sides to the power  $1/p$  yields  $(\mathbb{E}(\|\mathbf{X}\|^r))^{1/r} \leq (\mathbb{E}(\|\mathbf{X}\|^p))^{1/p}$  as claimed. ■

**Proof of Markov's Inequality (B.35) and (B.36).** Let  $F$  denote the distribution function of  $\mathbf{x}$ . Then

$$\begin{aligned} \mathbb{P}(g(\mathbf{x}) \geq \varepsilon) &= \int_{\{g(\mathbf{u}) \geq \varepsilon\}} dF(\mathbf{u}) \\ &\leq \int_{\{g(\mathbf{u}) \geq \varepsilon\}} \frac{g(\mathbf{u})}{\varepsilon} dF(\mathbf{u}) \\ &= \varepsilon^{-1} \int \mathbf{1}(g(\mathbf{u}) > \varepsilon) g(\mathbf{u}) dF(\mathbf{u}) \\ &= \varepsilon^{-1} \mathbb{E}(g(\mathbf{x}) \mathbf{1}(g(\mathbf{x}) > \varepsilon)) \end{aligned}$$

the inequality using the region of integration  $\{g(\mathbf{x}) > \varepsilon\}$ . This establishes the strong form (B.36). Since  $\mathbf{1}(g(\mathbf{x}) > \varepsilon) \leq 1$ , the final expression is less than  $\varepsilon^{-1}\mathbb{E}(g(\mathbf{x}))$ , establishing the standard form (B.35). ■

**Proof of Chebyshev's Inequality (B.37).** Define  $y = (x - \mathbb{E}(x))^2$  and note that  $\mathbb{E}(y) = \text{var}(x)$ . The events  $\{|x - \mathbb{E}(x)| > \varepsilon\}$  and  $\{y > \varepsilon^2\}$  are equal, so by an application of Markov's inequality (B.35) we find

$$\mathbb{P}(|x - \mathbb{E}(x)| > \varepsilon) = \mathbb{P}(y > \varepsilon^2) \leq \varepsilon^{-2}\mathbb{E}(y) = \varepsilon^{-2}\text{var}(x)$$

as stated. ■

**Proof of Bernsteins's Inequality (B.38).** We first show

$$\mathbb{P}\left(\sum_{i=1}^n x_i > \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2}{2\bar{\sigma}^2 + 2M\varepsilon/3}\right). \quad (\text{B.57})$$

Set  $t = \varepsilon / (\bar{\sigma}^2 + M\varepsilon/3) > 0$ . Using Markov's inequality (B.35) the left side of (B.57) equals

$$\mathbb{P}\left(\exp\left(t \sum_{i=1}^n x_i\right) > \exp(t\varepsilon)\right) \leq e^{-t\varepsilon} \mathbb{E}\left(\exp\left(t \sum_{i=1}^n x_i\right)\right) = e^{-t\varepsilon} \prod_{i=1}^n \mathbb{E}(\exp(tx_i)). \quad (\text{B.58})$$

The exponential function equals

$$\exp(x) = 1 + x + \sum_{k=2}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2} g(x)$$

where  $g(x) = 2 \sum_{k=2}^{\infty} \frac{x^{k-2}}{k!}$ . Notice for  $x < 3$  the fact  $k! \geq 2 \cdot 3^{k-2}$  implies

$$g(x) = 2 \sum_{k=2}^{\infty} \frac{x^{k-2}}{k!} \leq \sum_{k=2}^{\infty} \frac{x^{k-2}}{3^{k-2}} = \frac{1}{1-x/3}.$$

Then since  $tx_i \leq tM < 3$ ,

$$\mathbb{E}(\exp(tx_i)) = 1 + \mathbb{E}\left(\frac{t^2 x_i^2}{2} g(tx_i)\right) \leq 1 + \frac{t^2 \mathbb{E}(x_i^2)}{2(1-tM/3)} \leq \exp\left(\frac{t^2 \mathbb{E}(x_i^2)}{2(1-tM/3)}\right).$$

This means that the right side of (B.58) is less than

$$\exp\left(-t\varepsilon + \frac{t^2 \bar{\sigma}^2}{2(1-tM/3)}\right) = \exp\left(-\frac{\varepsilon^2}{2\bar{\sigma}^2 + 2M\varepsilon/3}\right)$$

the equality using the defintion  $t = \varepsilon / (\bar{\sigma}^2 + M\varepsilon/3)$ . This establishes (B.57).

Replacing  $x_i$  with  $-x_i$  in the above argument we obtain

$$\mathbb{P}\left(\sum_{i=1}^n x_i < -\varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2}{2\bar{\sigma}^2 + 2M\varepsilon/3}\right). \quad (\text{B.59})$$

Together, (B.57) and (B.59) establish (B.38). ■

**Proof of Bernsteins's Inequality for vectors (B.39).** By the triangle inequality (B.9), Boole's inequality (B.24), and then Bernstein's inequality (B.38)

$$\begin{aligned} \mathbb{P}\left(\left\|\sum_{i=1}^n \mathbf{x}_i\right\| > \varepsilon\right) &\leq \mathbb{P}\left(\sum_{j=1}^m \left|\sum_{i=1}^n x_{ji}\right| > \varepsilon\right) \\ &\leq \mathbb{P}\left(\bigcup_{j=1}^m \mathbf{1}\left(\left|\sum_{i=1}^n x_{ji}\right| > \varepsilon/m\right)\right) \\ &\leq \sum_{j=1}^m \mathbb{P}\left(\left|\sum_{i=1}^n x_{ji}\right| > \varepsilon/m\right) \\ &\leq 2m \exp\left(-\frac{(\varepsilon/m)^2}{2\bar{\sigma}^2 + 2M(\varepsilon/m)/3}\right) \end{aligned}$$

which is (B.39). ■

**Proof of Sub-Gaussian Inequality (B.40).** Using the definition of the exponential function and the fact  $(2k)! \geq 2^k k!$  we find that

$$\begin{aligned} \frac{\exp(-u) + \exp(u)}{2} &= \frac{1}{2} \sum_{k=0}^{\infty} \frac{1}{k!} ((-u)^k + u^k) \\ &= \sum_{k=0}^{\infty} \frac{u^{2k}}{(2k)!} \\ &\leq \sum_{k=0}^{\infty} \frac{(u^2/2)^k}{k!} \\ &= \exp\left(\frac{u^2}{2}\right). \end{aligned} \tag{B.60}$$

Set  $y_i = (a_i - x_i)/2a_i$ . Note that  $\lambda x_i = y_i(-\lambda a_i) + (1 - y_i)\lambda a_i$ . By the convexity of the exponential function

$$\begin{aligned} \exp(\lambda x_i) &= \exp(y_i(-\lambda a_i) + (1 - y_i)\lambda a_i) \\ &\leq y_i \exp(-\lambda a_i) + (1 - y_i) \exp(\lambda a_i). \end{aligned}$$

Taking expectations, using  $\mathbb{E}(y_i) = 1/2$ , and then (B.60), we find that

$$\begin{aligned} \mathbb{E}(\exp(\lambda x_i)) &\leq \mathbb{E}(y_i) \exp(-\lambda a_i) + (1 - \mathbb{E}(y_i)) \exp(\lambda a_i) \\ &= \frac{\exp(-\lambda a_i) + \exp(\lambda a_i)}{2} \\ &\leq \exp\left(\frac{\lambda^2 a_i^2}{2}\right). \end{aligned} \tag{B.61}$$

Since the  $x_i$  are independent, and then applying (B.61),

$$\mathbb{E}\left(\exp\left(\lambda \sum_{i=1}^n x_i\right)\right) = \prod_{i=1}^n \mathbb{E}(\exp(\lambda x_i)) \leq \prod_{i=1}^n \exp\left(\frac{\lambda^2 a_i^2}{2}\right) = \exp\left(\frac{\lambda^2 \sum_{i=1}^n a_i^2}{2}\right). \tag{B.62}$$

This is (B.40). ■

**Proof of Hoeffding's Inequality (B.41).** Set  $\lambda = \varepsilon / \sum_{i=1}^n a_i^2$ . Using Markov's inequality (B.35), the sub-Gaussian inequality (B.40), and the definition of  $\lambda$

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^n x_i > \varepsilon\right) &= \mathbb{P}\left(\exp\left(\lambda \sum_{i=1}^n x_i\right) > \exp(\lambda \varepsilon)\right) \\ &\leq \exp(-\lambda \varepsilon) \mathbb{E}\left(\exp\left(\lambda \sum_{i=1}^n x_i\right)\right) \\ &\leq \exp(-\lambda \varepsilon) \exp\left(\frac{\lambda^2 \sum_{i=1}^n a_i^2}{2}\right) \\ &= \exp\left(-\frac{\varepsilon^2}{2 \sum_{i=1}^n a_i^2}\right). \end{aligned}$$

This is (B.41). ■

**Proof of Simple Inequality (B.42).** For  $0 < r \leq 2$  (B.42) holds by an application of the  $c_r$  inequality (B.7). For  $2 < r \leq 3$  by the  $c_r$  inequality (B.5), the independence of  $X$  and  $Y$ , and the assumption that they are

mean zero

$$\begin{aligned}
\mathbb{E}|X+Y|^r &= \mathbb{E}(|X+Y|^{r-2}|X+Y|^2) \\
&\leq \mathbb{E}((|X|^{r-2} + |Y|^{r-2})(X^2 + 2XY + Y^2)) \\
&= \mathbb{E}|X|^r + 2\mathbb{E}(|X|^{r-1}Y) + \mathbb{E}(|X|^{r-2}Y^2) \\
&\quad + \mathbb{E}|Y|^r + 2\mathbb{E}(|Y|^{r-1}X) + \mathbb{E}(|Y|^{r-2}X^2) \\
&= \mathbb{E}|X|^r + \mathbb{E}|Y|^r + \mathbb{E}(|X|^{r-2}Y^2) + \mathbb{E}(|Y|^{r-2}X^2).
\end{aligned} \tag{B.63}$$

Using the geometric mean inequality (B.4) and Liapunov's inequality (B.34)

$$\mathbb{E}(|X|^{r-2}Y^2) \leq \frac{r-2}{r} (\mathbb{E}|X|^{r-2})^{r/(r-2)} + \left(\frac{2}{r}\mathbb{E}(Y^2)\right)^{r/2}$$

and similarly

$$\mathbb{E}(|Y|^{r-2}X^2) \leq \frac{r-2}{r} (\mathbb{E}|Y|^{r-2})^{r/(r-2)} + \left(\frac{2}{r}\mathbb{E}(X^2)\right)^{r/2}.$$

Hence (B.63) is bounded by  $2(\mathbb{E}|X|^r + \mathbb{E}|Y|^r)$  as claimed.

For  $r > 3$  instead of (B.5) use (B.5) which increases the bound by the factor  $2^{r-3}$  so that the constant is  $2^{r-2}$ . ■

**Proof of Difference Inequality (B.43).** For  $r = 2$  (B.43) holds by direct calculation so we assume  $0 < r < 2$ . We start with a characterization of an absolute moment. Define the generalized cosine integral

$$K(r) = \int_0^\infty t^{-(r+1)} (1 - \cos(t)) dt \tag{B.64}$$

which is finite for  $0 < r < 2$ . Making the change of variable  $t = sx$  and rearranging, we obtain

$$|x|^r = K(r)^{-1} \int_0^\infty t^{-(r+1)} (1 - \cos(tx)) dt.$$

Thus for any random variable  $X$  and  $0 < r < 2$

$$\mathbb{E}|X|^r = K(r)^{-1} \int_0^\infty t^{-(r+1)} (1 - R(t)) dt \tag{B.65}$$

where  $R(t) = \mathbb{E}(\cos(tx))$ , the real part of the characteristic function of  $X$ .

Let  $C(t)$  denote the characteristic function of  $X$  and let  $R(t)$  denote its real part. Since  $X$  and  $Y$  are independent and have the same distribution, the characteristic function of  $X - Y$  is  $C(t)C(-t) = |C(t)|^2$ . The facts  $|C(t)|^2 \geq |R(t)|^2$  and  $|R(t)| \leq 1$  imply

$$1 - |C(t)|^2 \leq 1 - R(t)^2 = (1 - R(t))(1 + R(t)) \leq 2(1 - R(t)) \tag{B.66}$$

Using the characterization (B.65), (B.66), and then (B.65) again we find that

$$\begin{aligned}
\mathbb{E}|X-Y|^r &= K(r)^{-1} \int_0^\infty t^{-(r+1)} (1 - |C(t)|^2) dt \\
&\leq 2 \sum_{i=1}^n K(r)^{-1} \int_0^\infty t^{-(r+1)} (1 - R(t)) dt \\
&= 2\mathbb{E}|X|^r
\end{aligned}$$

which is (B.43). ■

**Proof of Bahr-Esseen Inequality (B.44).** Our proof is taken from von Bahr and Esseen (1965). For  $0 < r \leq 1$  (B.44) holds by the  $c_r$  inequality (B.5). For  $r = 2$ , (B.44) holds by independence and direct calculation. We thus focus on the case  $1 < r < 2$ .

Let  $y_i$  be an independent copy of  $x_i$  and let  $\mathbb{E}_y$  denote expectation over  $y_i$ . Since  $x_i$  and  $y_i$  have the same distribution,  $\mathbb{E}_y(y_i) = 0$ . Thus using Jensen's inequality (B.26) since  $|u|^r$  is convex for  $r \geq 1$ ,

$$\begin{aligned} \mathbb{E} \left| \sum_{i=1}^n x_i \right|^r &= \mathbb{E} \left| \mathbb{E}_y \sum_{i=1}^n (x_i - y_i) \right|^r \\ &\leq \mathbb{E} \mathbb{E}_y \left| \sum_{i=1}^n (x_i - y_i) \right|^r \\ &= \mathbb{E} \left| \sum_{i=1}^n z_i \right|^r \end{aligned} \tag{B.67}$$

where the final equality sets  $z_i = x_i - y_i$ .

Let  $C_i(t)$  denote the characteristic function of  $x_i$ , and let  $R_i(t)$  denote its real part. Since  $x_i$  and  $y_i$  are independent and have the same distribution, the characteristic function of  $z_i$  is  $C_i(t)C_i(-t) = |C_i(t)|^2$ . Since the  $z_i$  are mutually independent, the characteristic function of  $\sum_{i=1}^n z_i$  is  $\prod_{i=1}^n |C_i(t)|^2$ , which is real. We next employ the following inequality. If  $|a_i| \leq 1$  then

$$1 - \prod_{i=1}^n a_i \leq \sum_{i=1}^n (1 - a_i). \tag{B.68}$$

To establish (B.68) first do so for  $n = 2$ , and then apply induction. The inequality follows from  $0 \leq (1 - a_1)(1 - a_2) = 1 - a_1 - a_2 + a_1 a_2$  and rearranging. Since  $|C_i(t)| \leq 1$  (B.68) implies

$$1 - \prod_{i=1}^n |C_i(t)|^2 \leq \sum_{i=1}^n (1 - |C_i(t)|^2). \tag{B.69}$$

Recalling the characterization (B.65) and the fact that the characteristic function of  $\sum_{i=1}^n z_i$  is  $\prod_{i=1}^n |C_i(t)|^2$ , applying (B.69), and then using (B.65) again

$$\begin{aligned} \mathbb{E} \left| \sum_{i=1}^n z_i \right|^r &= K(r)^{-1} \int_0^\infty t^{-(r+1)} \left( 1 - \prod_{i=1}^n |C_i(t)|^2 \right) dt \\ &\leq \sum_{i=1}^n K(r)^{-1} \int_0^\infty t^{-(r+1)} (1 - |C_i(t)|^2) dt \\ &= \sum_{i=1}^n \mathbb{E} |z_i|^r \\ &\leq 2 \sum_{i=1}^n \mathbb{E} |z_i|^r. \end{aligned}$$

The final inequality is (B.43) since  $z_i = x_i - y_i$  where  $x_i$  and  $y_i$  are independent with the same distribution. Combined with (B.67) this is (B.44). ■

**Proof of Symmetrization Inequality (B.45).** Let  $y_i$  be an independent copy of  $x_i$ . As shown in (B.67)

$$\mathbb{E} \left| \sum_{i=1}^n x_i \right|^r \leq \mathbb{E} \left| \sum_{i=1}^n (x_i - y_i) \right|^r. \tag{B.70}$$

Let  $\varepsilon_i$  be an independent Rademacher random variable. Since  $x_i - y_i$  is symmetrically distributed about 0, it has the same distribution as  $\varepsilon_i(x_i - y_i)$ . Thus (B.70) equals

$$\mathbb{E} \left| \sum_{i=1}^n \varepsilon_i (x_i - y_i) \right|^r = \mathbb{E} \mathbb{E}_{xy} \left| \sum_{i=1}^n \varepsilon_i x_i - \sum_{i=1}^n \varepsilon_i y_i \right|^r \tag{B.71}$$

where  $\mathbb{E}_{xy}$  denotes expectations over  $x_i$  and  $y_i$  only. Conditional on  $\{\varepsilon_i\}$ , the two sums in (B.71) are independent and identically distributed. For  $r \leq 2$  we employ the difference inequality (B.43) and for  $r > 2$  the  $c_r$  inequality (B.6). This bounds (B.71) by

$$D_r \mathbb{E} \left| \sum_{i=1}^n \varepsilon_i x_i \right|^r$$

which is (B.45). ■

**Proof of Khintchine's Inequality (B.46).** For  $r \leq 2$  by Liapunov's inequality (B.34)

$$\left( \mathbb{E} \left| \sum_{i=1}^n a_i \varepsilon_i \right|^r \right)^{1/r} \leq \left( \mathbb{E} \left| \sum_{i=1}^n a_i \varepsilon_i \right|^2 \right)^{1/2} = \left( \sum_{i=1}^n \mathbb{E}(a_i^2 \varepsilon_i^2) \right)^{1/2} = \left( \sum_{i=1}^n a_i^2 \right)^{1/2}$$

which is (B.46) with  $K_r = 1$ .

Take  $r \geq 2$ . Let  $b_i = a_i / (\sum_{i=1}^n a_i^2)^{1/2}$  so  $\sum_{i=1}^n b_i^2 = 1$ . Then

$$\mathbb{E} \left| \sum_{i=1}^n a_i \varepsilon_i \right|^r = \left( \sum_{i=1}^n a_i^2 \right)^{r/2} \mathbb{E} \left| \sum_{i=1}^n b_i \varepsilon_i \right|^r. \quad (\text{B.72})$$

We show below that the expectation is bounded by replacing the  $b_i$  with the common value  $n^{-1/2}$ . Thus

$$\begin{aligned} \mathbb{E} \left| \sum_{i=1}^n b_i \varepsilon_i \right|^r &\leq \mathbb{E} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \right|^r \\ &\leq \limsup_{n \rightarrow \infty} \mathbb{E} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \right|^r \\ &= \mathbb{E} |Z|^r = 2^{r/2} \Gamma((r+1)/2) / \pi^{1/2} \end{aligned} \quad (\text{B.73})$$

The second-to-last equality follows from the central limit theorem and the fact that  $\varepsilon_i$  are bounded and thus uniformly integrable. The final equality is Theorem 5.2. Together with (B.72) this is (B.46) with  $K_r = 2^{r/2} \Gamma((r+1)/2) / \pi^{1/2}$ .

The proof is completed by showing (B.73). Without loss of generality assume  $b_i \geq 0$  and are ordered ascending, so that  $b_1$  is the smallest and  $b_n$  is the largest. The argument below shows that the left side of (B.73) is increased if we replace  $b_1$  and  $b_n$  by the common value  $\sqrt{(b_1^2 + b_n^2)/2}$  (which does not alter the sum  $\sum_{i=1}^n b_i^2$ ). Iteratively this implies (B.73).

Set  $S = \sum_{i=2}^{n-1} b_i \varepsilon_i$ . Then

$$\begin{aligned} \mathbb{E} \left[ \left| \sum_{i=1}^n b_i \varepsilon_i \right|^r \middle| S \right] &= \mathbb{E} [ |b_1 \varepsilon_1 + b_n \varepsilon_n + S|^r \mid S] \\ &= \frac{1}{4} [ |b_1 + b_n + S|^r + |b_1 - b_n + S|^r + |-b_1 + b_n + S|^r + |-b_1 - b_n + S|^r ] \\ &= g(u_1) + g(u_2) \end{aligned} \quad (\text{B.74})$$

where

$$\begin{aligned} g(u) &= \frac{1}{4} (|S + \sqrt{u}|^r + |S - \sqrt{u}|^r) \\ u_1 &= (b_1 + b_n)^2 \\ u_2 &= (b_n - b_1)^2. \end{aligned}$$

Notice that  $g(u)$  is convex on  $u \geq 0$  since  $S \geq 0$  and  $r \geq 2$ . (For a formal proof see Whittle (1960, Lemma 1.) Set  $c = 2(b_1^2 + b_n^2)$ . We find

$$\begin{aligned} g(u_1) + g(u_2) &= g\left(\frac{u_1}{c}w_1 + (1 - \frac{u_1}{c})0\right) + g\left(\frac{u_2}{c}c + (1 - \frac{u_2}{c})0\right) \\ &\leq \frac{u_1}{c}g(c) + (1 - \frac{u_1}{c})g(0) + \frac{u_2}{c}g(c) + (1 - \frac{u_2}{c})g(0) \\ &= g(c) + g(0). \end{aligned}$$

The inequality is two applications of Jensen's (B.2) and the final equality is  $u_1 + u_2 = c$ . Combined with (B.74) we have shown that

$$\mathbb{E}\left[\left|\sum_{i=1}^n b_i \varepsilon_i\right|^r \middle| S\right] \leq g(c) + g(0).$$

The right-hand-side is (B.74) when  $b_1 = b_n = \sqrt{c}/2$ .

This means that the left side of (B.73) can be increased by replacing  $b_1$  and  $b_n$  by the common value  $\sqrt{c}/2$  as described earlier. Iteratively, we replace the smallest and largest  $b_i$  by their common value, with each step increasing the expectation. In the limit we obtain (B.73). ■

**Proof of Marcinkiewicz-Zygmund Inequality (B.47).** Let  $\varepsilon_i$  be independent Rademacher random variables. Let  $\mathbb{E}_x$  and  $\mathbb{E}_\varepsilon$  denote expectations over  $x_i$  and  $\varepsilon_i$ , respectively. By the symmetrization inequality (B.45)

$$\mathbb{E}\left|\sum_{i=1}^n x_i\right|^r \leq D_r \mathbb{E}\left|\sum_{i=1}^n \varepsilon_i x_i\right|^r = D_r \mathbb{E}_x \mathbb{E}_\varepsilon \left|\sum_{i=1}^n \varepsilon_i x_i\right|^r. \quad (\text{B.75})$$

The expectation  $\mathbb{E}_\varepsilon$  treats  $x_i$  as fixed, so we can apply Khintchine's inequality (B.46). Thus (B.75) is bounded by

$$D_r K_r \mathbb{E}_x \left(\sum_{i=1}^n x_i^2\right)^{r/2} = D_r K_r \mathbb{E} \left(\sum_{i=1}^n x_i^2\right)^{r/2}.$$

This is (B.47). ■

**Proof of Whittle's Inequality (B.48).** By the Marcinkiewicz-Zygmund inequality (B.47), Minkowski's inequality (B.33) and  $\mathbb{E}|x_i|^r \leq B_{1r}$

$$\begin{aligned} \mathbb{E}\left|\sum_{i=1}^n a_i x_i\right|^r &\leq M_r \mathbb{E}\left|\sum_{i=1}^n a_i^2 x_i^2\right|^{r/2} \\ &\leq M_r \left(\sum_{i=1}^n a_i^2 (\mathbb{E}|x_i|^r)^{2/r}\right)^{r/2} \\ &\leq B_{1r} M_r \left(\sum_{i=1}^n a_i^2\right)^{r/2}. \end{aligned}$$

which is (B.48) with  $C_{1r} = B_{1r} M_r$  as claimed. ■

**Proof of Whittle's Inequality (B.49).** As shown in (B.67)

$$\mathbb{E}|x' A x - \mathbb{E}(x' A x)|^r \leq \mathbb{E}|x' A x - y' A y|^r \quad (\text{B.76})$$

where  $y = (y_1, \dots, y_n)'$  is an independent copy of  $x$ . We can write

$$(x + y)' A (x - y) = \xi' (x - y)$$

where  $\xi = A(x + y)$ .

Independence implies exchangeability, which implies the distribution of  $x_i - y_i$  conditional on  $x_i + y_i$  is symmetric about the origin. To see this formally, by exchangeability

$$\begin{aligned}\mathbb{P}(x_i - y_i \leq u | x_i + y_i = v) &= \mathbb{P}(y_i - x_i \leq u | y_i + x_i = v) \\ &= 1 - \mathbb{P}(y_i - x_i > u | y_i + x_i = v) \\ &= 1 - \mathbb{P}(x_i - y_i < -u | x_i + y_i = v) \\ &= \mathbb{P}(x_i - y_i \geq -u | x_i + y_i = v)\end{aligned}$$

which is the definition of a symmetric distribution. Thus we can write  $x_i - y_i = \eta_i \varepsilon_i$  where  $\eta_i = |x_i - y_i|$  and  $\varepsilon_i$  is an independent Rademacher random variable.

Denote the  $i^{th}$  element of  $\xi$  as  $\xi_i = \mathbf{a}'_i(\mathbf{x} + \mathbf{y})$  where  $\mathbf{a}'_i$  is the  $i^{th}$  column of  $\mathbf{A}$ . Then  $\xi'(\mathbf{x} - \mathbf{y}) = \sum_{i=1}^n \xi_i \eta_i \varepsilon_i$ . Applying Khintchine's inequality

$$\begin{aligned}\mathbb{E}|\mathbf{x}'\mathbf{A}\mathbf{x} - \mathbf{y}'\mathbf{A}\mathbf{y}|^r &= \mathbb{E}\left|\sum_{i=1}^n \xi_i \eta_i \varepsilon_i\right|^r \\ &\leq K_r \mathbb{E}\left(\sum_{i=1}^n \xi_i^2 \eta_i^2\right)^{r/2} \\ &\leq K_r \left(\sum_{i=1}^n (\mathbb{E}(|\xi_i|^r |\eta_i|^r))^{2/r}\right)^{r/2} \\ &\leq K_r \left(\sum_{i=1}^n (\mathbb{E}(\xi_i^{2r}))^{1/r} (\mathbb{E}(\eta_i^{2r}))^{1/r}\right)^{r/2}. \tag{B.77}\end{aligned}$$

The first inequality is Minkowski's (B.33), the second is Cauchy-Schwarz (B.31). Observe that

$$\mathbb{E}(\eta_i^{2r}) \leq 2^{2r} \mathbb{E}(x_i^{2r}) \leq 2^{2r} B_{2r}.$$

By Whittle's first inequality (B.48)

$$\mathbb{E}(\xi_i^{2r}) = \mathbb{E}\left|\sum_{j=1}^n a_{ji}(x_j + y_j)\right|^{2r} \leq 2^{2r} C_{1r} \left(\sum_{j=1}^n a_{ji}^2\right)^r.$$

Hence (B.77) is bounded by

$$4^r K_r C_{1r}^{1/2} B_{2r}^{1/2} \left(\sum_{i=1}^n \sum_{j=1}^n a_{ji}^2\right)^{r/2} = 4^r K_r C_{1r}^{1/2} B_{2r}^{1/2} (\text{tr}(\mathbf{A}'\mathbf{A})^{r/2})^{r/2}.$$

This is (B.49) with  $C_{2r} = 4^r K_r C_{1r}^{1/2} B_{2r}^{1/2}$ . ■

**Proof of Rosenthal's Inequality (B.50).** Define  $\mu_s = \sum_{i=1}^n \mathbb{E}|x_i|^s$  for any  $s > 0$ .

Take  $0 < r \leq 2$ . By Liapunov's inequality (B.34)

$$\left(\mathbb{E}\left|\sum_{i=1}^n x_i\right|^r\right)^{1/r} \leq \left(\mathbb{E}\left|\sum_{i=1}^n x_i\right|^2\right)^{1/2} = \mu_2^{1/2}.$$

Raising to the power  $r$  implies (B.50). For the remainder assume  $r > 2$ .

By the Marcinkiewicz-Zygmund inequality (B.47)

$$\mathbb{E}\left|\sum_{i=1}^n x_i\right|^r \leq M_r \mathbb{E}|S_n|^{r/2}. \tag{B.78}$$

where  $S_n = \sum_{i=1}^n x_i^2$ . For any  $i$ , using the  $c_r$  inequality (B.7)

$$|S_n|^{r/2-1} = \left| x_i + \sum_{j \neq i} x_j \right|^{r/2-1} \leq c_{r/2-1} \left( |x_i|^{r/2} + \left| \sum_{j \neq i} x_j \right|^{r/2-1} \right)$$

Thus

$$\begin{aligned} \mathbb{E}|S_n|^{r/2} &= \mathbb{E}(S_n |S_n|^{r/2-1}) \\ &= \sum_{i=1}^n \mathbb{E}(x_i^2 |S_n|^{r-2}) \\ &\leq c_{r/2-1} \sum_{i=1}^n \mathbb{E}\left(x_i^2 \left(|x_i|^{r/2} + \left| \sum_{j \neq i} x_j \right|^{r/2-1}\right)\right) \\ &\leq c_{r/2-1} \left( \sum_{i=1}^n \mathbb{E}|x_i|^r + \sum_{i=1}^n \mathbb{E}\left(x_i^2 \left| \sum_{j \neq i} x_j^2 \right|^{r/2-1}\right) \right) \\ &= c_{r/2-1} \left( \mu_r + \sum_{i=1}^n \mathbb{E}(x_i^2) \mathbb{E}\left| \sum_{j \neq i} x_j^2 \right|^{r/2-1} \right) \\ &\leq c_{r/2-1} (\mu_r + \mu_2 \mathbb{E}|S_n|^{r/2-1}). \end{aligned} \quad (\text{B.79})$$

The second-to-last line holds since  $x_i^2$  is independent of  $\sum_{j \neq i} x_j^2$ . The final inequality holds since  $\sum_{j \neq i} x_j^2 \leq S_n$  and  $\sum_{i=1}^n \mathbb{E}(x_i^2) = \mu_2$ .

Suppose  $2 \leq r \leq 4$ . Then  $r/2 - 1 \leq 1$ . By Jensen's inequality (B.26)  $\mathbb{E}|S_n|^{r/2-1} \leq (\mathbb{E}|S_n|)^{r/2-1} = \mu_2^{r/2-1}$ . Also,  $c_{r/2-1} = 1$ . Together, we can bound (B.79) by  $\sum_{i=1}^n \mathbb{E}|x_i|^r + \mu_2^{r/2}$ . This implies

$$\mathbb{E}|S_n|^{r/2} \leq \sum_{i=1}^n \mathbb{E}|x_i|^r + \mu_2^{r/2}. \quad (\text{B.80})$$

We now establish

$$\mathbb{E}|S_n|^{s/2} \leq 2^{s(s-2)/8} (\mu_s + \mu_2^{s/2}) \quad (\text{B.81})$$

for all  $s \geq 2$  by a recursive argument. (B.80) shows that (B.81) holds for  $2 \leq s \leq 4$ . We now show that (B.81) for  $s = r - 2$  implies (B.81) for  $s = r$ . Take  $r > 4$ . Using (B.79),  $c_{r/2-1} = 2^{r/2-2}$  and (B.81)

$$\begin{aligned} \mathbb{E}|S_n|^{r/2} &\leq 2^{r/2-2} (\mu_r + \mu_2 \mathbb{E}|S_n|^{r/2-1}) \\ &\leq 2^{r/2-2} \left( \mu_r + \mu_2 2^{(r-2)(r-4)/8} (\mu_{r-2} + \mu_2^{(r-2)/2}) \right) \\ &\leq 2^{r/2-2} 2^{(r-2)(r-4)/8} (\mu_r + \mu_2 \mu_{r-2} + \mu_2^{r/2}) \\ &= 2^{r(r-2)/8-1} (\mu_r + \mu_2 \mu_{r-2} + \mu_2^{r/2}). \end{aligned}$$

Using Hölder's inequality (B.30), Hölder's inequality for vectors (B.11), and the geometric mean inequality (B.4)

$$\begin{aligned} \mu_2 \mu_{r-2} &\leq \mu_2 \sum_{i=1}^n (\mathbb{E}(x_i^2))^{2/(r-2)} (\mathbb{E}|x_i|^r)^{(r-4)/(r-2)} \\ &\leq \mu_2 \left( \sum_{i=1}^n \mathbb{E}(x_i^2) \right)^{2/(r-2)} \left( \sum_{i=1}^n \mathbb{E}|x_i|^r \right)^{(r-4)/(r-2)} \\ &= \mu_2^{r/(r-2)} \mu_r^{(r-4)/(r-2)} \\ &\leq \frac{2}{r-2} \mu_2^{r/2} + \frac{r-4}{r-2} \mu_r \\ &\leq \mu_2^{r/2} + \mu_r. \end{aligned}$$

Together

$$\mathbb{E}|S_n|^{r/2} \leq 2^{r(r-2)/8} (\mu_r + \mu_2^{r/2})$$

which is (B.81) for  $s = r$ . This shows that (B.81) holds for all  $s \geq 2$ .

(B.78) and (B.81) imply (B.50) for  $r > 2$ . ■

**Proof of Maximal Inequality (B.51).** Set  $S_i = \sum_{j=1}^i x_j$ . Note that since the observations are independent and mean zero

$$\mathbb{E}(S_n | x_1, \dots, x_i) = \sum_{j=1}^i x_j = S_i.$$

By the conditional Jensen's inequality (B.27), since  $|u|^r$  is convex

$$|S_i|^r = |\mathbb{E}(S_n | x_1, \dots, x_i)|^r \leq \mathbb{E}(|S_n|^r | x_1, \dots, x_i). \quad (\text{B.82})$$

Let  $E_i$  be the event that  $|S_i|$  is the first  $|S_j|$  which strictly exceeds  $\varepsilon$ . Formally

$$E_i = \left\{ |S_i| > \varepsilon, \max_{j < i} |S_j| \leq \varepsilon \right\}. \quad (\text{B.83})$$

These events are disjoint. Their union is

$$E = \bigcup_{i=1}^n E_i = \left\{ \max_{i \leq n} |S_i| > \varepsilon \right\}.$$

By the same method as to prove Markov's inequality (B.35)

$$\begin{aligned} \varepsilon^r \mathbb{P}(E) &= \sum_{i=1}^n \varepsilon^r \mathbb{E}(\mathbf{1}(E_i)) \\ &\leq \sum_{i=1}^n \mathbb{E}(|S_i|^r \mathbf{1}(E_i)) \\ &\leq \sum_{i=1}^n \mathbb{E}(\mathbb{E}(|S_n|^r | x_1, \dots, x_i) \mathbf{1}(E_i)) \\ &= \sum_{i=1}^n \mathbb{E}(\mathbb{E}(|S_n|^r \mathbf{1}(E_i) | x_1, \dots, x_i)) \\ &= \sum_{i=1}^n \mathbb{E}(|S_n|^r \mathbf{1}(E_i)) \\ &= \mathbb{E}(|S_n|^r \mathbf{1}(E)) \\ &\leq \mathbb{E}|S_n|^r. \end{aligned} \quad (\text{B.84})$$

The second inequality is (B.82). The following equalities use the conditioning theorem, the law of iterated expectations, and the definition of the event  $E$ .

In our proof of Doob's inequality below we will also (B.84) which can be written as

$$\mathbb{P}\left(\max_{1 \leq j \leq n} \left| \sum_{i=1}^j x_i \right| > \varepsilon\right) \leq \varepsilon^{-r} \mathbb{E}\left(\left| \sum_{i=1}^n x_i \right|^r \mathbf{1}\left(\max_{1 \leq j \leq n} \left| \sum_{i=1}^j x_i \right| > \varepsilon\right)\right). \quad (\text{B.85})$$

■

**Proof of Kolmogorov's Inequality (B.52).** By the maximal inequality (B.51) and the independence assumption

$$\mathbb{P}\left(\max_{1 \leq j \leq n} \left| \sum_{i=1}^j x_i \right| > \varepsilon\right) \leq \varepsilon^{-2} \mathbb{E}\left(\sum_{i=1}^n x_i\right)^2 = \varepsilon^{-2} \sum_{i=1}^n \sigma_i^2.$$

■

**Proof of Doob's Inequality (B.53).** Define  $S_i = \sum_{j=1}^i x_j$  and  $R_n = \max_{1 \leq i \leq n} |S_i|$ . Using Theorem 2.12, the strong maximal inequality (B.85), and Hölder's Inequality (B.30)

$$\begin{aligned}\mathbb{E}(R_n^r) &= \int_0^\infty \mathbb{P}(R_n^r > u) du \\ &= \int_0^\infty \mathbb{P}(R_n > u^{1/r}) du \\ &\leq \int_0^\infty u^{-1/r} \mathbb{E}(|S_n| \mathbf{1}(R_n > u^{1/r})) du \\ &\leq \mathbb{E}\left(|S_n| \int_0^{R_n^r} u^{-1/r} du\right) \\ &= \frac{r}{r-1} \mathbb{E}(|S_n| R_n^{r-1}) \\ &\leq \frac{r}{r-1} (\mathbb{E}|S_n|^r)^{1/r} (\mathbb{E}R_n^r)^{(r-1)/r}.\end{aligned}$$

Solving, we find

$$\mathbb{E}(R_n^r) \leq \left(\frac{r}{r-1}\right)^r \mathbb{E}|S_n|^r$$

which is (B.53). ■

**Proof of Ottaviani's Inequality (B.54).** Define  $E_i$  as in (B.83). Then for any  $i$ ,

$$\begin{aligned}\left(1 - \max_{j \leq n} \mathbb{P}(S_n - S_j > \varepsilon/2)\right) \mathbb{P}(E_i) &= \min_{j \leq n} \mathbb{P}(|S_n - S_j| \leq \varepsilon/2) \mathbb{P}(E_i) \\ &\leq \mathbb{P}(|S_n - S_i| \leq \varepsilon/2) \mathbb{P}(E_i) \\ &= \mathbb{P}(|S_n - S_i| \leq \varepsilon/2, E_i) \\ &\leq \mathbb{P}(|S_n| \leq \varepsilon/2, E_i).\end{aligned}$$

The second equality holds since  $|S_n - S_i|$  is independent of  $E_i$ . The final inequality holds since  $|S_n - S_i| \leq \varepsilon/2$  and  $|S_i| > \varepsilon$  imply  $|S_n| > \varepsilon/2$ .

Summing over  $i$  we obtain

$$\left(1 - \max_{j \leq n} \mathbb{P}(S_n - S_j > \varepsilon/2)\right) \sum_{i=1}^n \mathbb{P}(E_i) \leq \sum_{i=1}^n \mathbb{P}(|S_n| \leq \varepsilon/2, E_i).$$

Since the events  $E_i$  are disjoint this implies

$$\begin{aligned}\left(1 - \max_{j \leq n} \mathbb{P}(S_n - S_j > \varepsilon/2)\right) \mathbb{P}\left(\max_{i \leq n} |S_i| > \varepsilon\right) &= \left(1 - \max_{j \leq n} \mathbb{P}(S_n - S_j > \varepsilon/2)\right) \mathbb{P}\left(\bigcup_{i=1}^n E_i\right) \\ &\leq \mathbb{P}\left(|S_n| \leq \varepsilon/2, \bigcup_{i=1}^n E_i\right) \\ &\leq \mathbb{P}(|S_n| \leq \varepsilon/2).\end{aligned}$$

This is (B.54). ■

# References

- [1] Abadir, Karim M. and Jan R. Magnus (2005): *Matrix Algebra*, Cambridge University Press.
- [2] Acemoglu, Daron, Simon Johnson, James A. Robinson (2001): “The Colonial Origins of Comparative Development: An Empirical Investigation,” *American Economic Review*, 91, 1369-1401.
- [3] Acemoglu, Daron, Simon Johnson, James A. Robinson (2012): “The Colonial Origins of Comparative Development: An Empirical Investigation: Reply,” *American Economic Review*, 102, 3077-3110.
- [4] Aitken, A.C. (1935): “On least squares and linear combinations of observations,” *Proceedings of the Royal Statistical Society*, 55, 42-48.
- [5] Akaike, H. (1969). “Fitting autoregressive models for prediction,” *Annals of the Institute of Statistical Mathematics*, 21, 243–247.
- [6] Akaike, H. (1973). “Information theory and an extension of the maximum likelihood principle,” in Petrov, B. and Csaki, F. (editors), *Second International Symposium on Information Theory*, 267-281. Akademiai Kiado, Budapest.
- [7] Amemiya, Takeshi (1971): “The estimation of the variances in a variance-components model,” *International Economic Review*, 12, 1-13.
- [8] Amemiya, Takeshi (1974): “The nonlinear two-stage least-squares estimator,” *Journal of Econometrics*, 2, 105-110.
- [9] Amemiya, Takeshi (1977): “The maximum likelihood and nonlinear three-stage least squares estimator in the general nonlinear simultaneous equations model,” *Econometrica*, 45, 955-968.
- [10] Amemiya, Takeshi (1985): *Advanced Econometrics*, Harvard University Press.
- [11] Amemiya, Takeshi (1994): *Introduction to Statistics and Econometrics*, Harvard University Press.
- [12] Amemiya, Takeshi. and Thomas E. MaCurdy (1986): “Instrumental-variable estimation of an error components model,” *Econometrica*, 54, 869-881.
- [13] Anderson, Theodore W. (1951): “Estimating linear restrictions on regression coefficients for multivariate normal distributions,” *Annals of Mathematical Statistics*, 22, 327-350.
- [14] Anderson, Theodore W. and Cheng Hsiao (1982): “Formulation and estimation of dynamic models using panel data,” *Journal of Econometrics*, 18, 47-82.
- [15] Anderson, Theodore W. and H. Rubin (1949): “Estimation of the parameters of a single equation in a complete system of stochastic equations,” *The Annals of Mathematical Statistics*, 20, 46-63.
- [16] Andrews, Donald W. K. (1984), “Non-strong mixing autoregressive processes,” *Journal of Applied Probability*, 21, 930-934.

- [17] Andrews, Donald W. K. (1991a), "Asymptotic normality of series estimators for nonparametric and semiparametric regression models," *Econometrica*, 59, 307-345.
- [18] Andrews, Donald W. K. (1991b), "Heteroskedasticity and autocorrelation consistent covariance matrix estimation," *Econometrica*, 59, 817-858.
- [19] Andrews, Donald W. K. (1991c): "Asymptotic optimality of generalized  $C_L$ , cross-validation, and generalized cross-validation in regression with heteroskedastic errors," *Journal of Econometrics*, 47, 359-377.
- [20] Andrews, Donald W. K. (1993), "Tests for parameter instability and structural change with unknown change point," *Econometrica*, 61, 821-8516.
- [21] Andrews, Donald W. K. (2017), "Examples of  $L^2$ -complete and boundedly-complete distributions," *Journal of Econometrics*, 199, 213-220.
- [22] Andrews, Donald W. K. and Werner Ploberger (1994): "Optimal tests when a nuisance parameter is present only under the alternative," *Econometrica*, 62, 1383-1414.
- [23] Angrist, Joshua D., Guido W. Imbens, and Alan B. Krueger (1991): "Jackknife instrumental variables estimation," *Journal of Applied Econometrics*, 14, 57-67.
- [24] Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996): "Identification of causal effects using instrumental variables," *Journal of the American Statistical Association*, 55, 650-659.
- [25] Angrist, Joshua D. and Alan B. Krueger (1991): "Does compulsory school attendance affect schooling and earnings?" *Quarterly Journal of Economics*, 91, 444-455.
- [26] Angrist, Joshua D. and Victor Lavy (1999): "Using Maimonides' rule to estimate the effect of class size on scholastic achievement," *Quarterly Journal of Economics*, 114, 533-575.
- [27] Angrist, Joshua D. and Jörn-Steffen Pischke (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.
- [28] Arellano, Manuel (1987): "Computing standard errors for robust within-groups estimators," *Oxford Bulletin of Economics and Statistics*, 49, 431-434.
- [29] Arellano, Manuel (2003): *Panel Data Econometrics*, Oxford University Press.
- [30] Arellano, Manuel and Stephen Bond (1991): "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations," *Review of Economic Studies*, 58, 277-297.
- [31] Arellano, Manuel and Olympia Bover (1995): "Another look at the instrumental variable estimation of error-components models," *Journal of Econometrics*, 68, 29-51.
- [32] Ash, Robert B. (1972): *Real Analysis and Probability*, Academic Press.
- [33] Bai, Jushan (2003): "Inferential theory for factor models of large dimensions," *Econometrica*, 71, 135-172.
- [34] Bai, Jushan and Serena Ng (2002): "Determining the number of factors in approximate factor models," *Econometrica*, 70, 191-221.
- [35] Bai, Jushan and Serena Ng (2006): "Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions," *Econometrica*, 74, 1133-1150.
- [36] Balestra, Pietro and Marc Nerlove (1966): "Pooling cross section and time series data in the estimation of a dynamic model: The demand for natural gas," *Econometrica*, 34, 585-612.

- [37] Baltagi, Badi H. (2013): *Econometric Analysis of Panel Data, 5<sup>th</sup> Edition*, Wiley.
- [38] Barro, Robert J. (1977): “Unanticipated money growth and unemployment in the United States,” *American Economic Review*, 67, 101–115
- [39] Basmann, R. L. (1957): “A generalized classical method of linear estimation of coefficients in a structural equation,” *Econometrica*, 25, 77-83.
- [40] Basmann, R. L. (1960): “On finite sample distributions of generalized classical linear identifiability test statistics,” *Journal of the American Statistical Association*, 55, 650-659.
- [41] Baum, Christopher F, Mark E. Schaffer, and Steven Stillman (2003): “Instrumental variables and GMM: Estimation and testing,” *The Stata Journal*, 3, 1-31.
- [42] Bekker, P.A. (1994): “Alternative approximations to the distributions of instrumental variable estimators, *Econometrica*, 62, 657-681.
- [43] Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato (2015): “Some new asymptotic theory for least squares series: Pointwise and uniform results,” *Journal of Econometrics*, 186, 345-366.
- [44] Bernheim, B. Douglas, Jonathan Meer and Neva K. Novarro (2016): “Do consumers exploit commitment opportunities? Evidence from natural experiments involving liquor consumption,” *American Economic Journal: Economic Policy*, 8, 41-69.
- [45] Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004): “How much should we trust differences-in-differences estimates?” *Quarterly Journal of Economics*, 119, 249-275.
- [46] Billingsley, Patrick (1968): *Convergence of Probability Measures*. New York: Wiley.
- [47] Billingsley, Patrick (1995): *Probability and Measure*, 3rd Edition, New York: Wiley.
- [48] Blanchard, Olivier Jean and Roberto Perotti (2002): “An empirical characterization of the dynamic effects of changes in government spending and taxes on output,” *Quarterly Journal of Economics*, 117, 1329-1368.
- [49] Blanchard, Olivier Jean and Danny Quah (1989): “The dynamic effects of aggregate demand and supply disturbances,” *American Economic Review*, 89, 655-673.
- [50] Blundell, Richard and Stephen Bond (1998): “Initial conditions and moment restrictions in dynamic panel data models,” *Journal of Econometrics*, 87, 115-143.
- [51] Bock, M.E. (1975): “Minimax estimators of the mean of a multivariate normal distribution,” *The Annals of Statistics*, 3, 209-218.
- [52] Box, George E. P. and Dennis R. Cox, (1964): “An analysis of transformations,” *Journal of the Royal Statistical Society, Series B*, 26, 211-252.
- [53] Breusch, Trevor S., Graham E. Mizon and Peter Schmidt (1989): “Efficient estimation using panel data,” *Econometrica*, 57, 695-700.
- [54] Brockwell, Peter J. and Richard A. Davis (1991): *Time Series: Theory and Methods, Second Edition*, Springer-Verlag.
- [55] Burnham, Kenneth P. and David R. Anderson (1998): *Model Selection and MultiModel Inference: A Practical Information-Theoretic Approach*, 2<sup>nd</sup> Edition, Springer.
- [56] Cameron, A. Colin and Pravin K. Trivedi (1998): *Regression Analysis of Count Data*, Cambridge University Press.

- [57] Cameron, A. Colin, Johan B. Gelbach, and Douglas L. Miller (2008): "Bootstrap-based improvements for inference with clustered errors," *Review of Economics and Statistics*, 90, 414-437.
- [58] Cameron, A. Colin and Pravin K. Trivedi (2005): *Microeometrics: Methods and Applications*, Cambridge University Press.
- [59] Canova, Fabio (1995): "Vector autoregressive models: Specification, estimation, inference, and forecasting," in *Handbook of Applied Econometrics, Volume 1: Macroeconomics*, edited by M. Hashem Pesaran and Michael R. Wickens, Blackwell.
- [60] Card, David (1995): "Using geographic variation in college proximity to estimate the return to schooling," in *Aspects of Labor Market Behavior: Essays in Honour of John Vanderkamp*, L.N. Christofides, E.K. Grant, and R. Swidinsky, editors. Toronto: University of Toronto Press.
- [61] Card, David and Alan B Krueger (1994): "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania," *American Economic Review*, 84, 772-793.
- [62] Card, David, David S. Lee, Zhuan Pei, and Andrea Weber (2015): "Inference on causal effects in a generalized regression kink design," *Econometrica*, 57, 695-700.
- [63] Casella, George and Roger L. Berger (2002): *Statistical Inference*, 2nd Edition, Duxbury Press.
- [64] Chamberlain, Gary (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305-334.
- [65] Chang, Pao Li and Shinichi Sakata (2007): "Estimation of impulse response functions using long autoregression," *Econometrics Journal*, 10, 453-469.
- [66] Chao, John C., Norman R. Swanson, Jerry A. Hausman, Whitney K. Newey, and Tiemen Woutersen (2012): "Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments," *Econometric Theory*, 28, 42-86.
- [67] Chen, Xiaohong (2007): "Large sample sieve estimation of semi-nonparametric models," in James J. Heckman and Edward E. Leamer, (eds.) *Handbook of Econometrics*, vol. VI, Part B, 5549-5632, North Holland: Amsterdam.
- [68] Chen, Xiaohong and Timothy M. Christensen (2015): "Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions," *Journal of Econometrics*, 188, 447-465.
- [69] Chen, Xiaohong and Timothy M. Christensen (2018): "Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression," *Quantitative Economics*, 9, 39-84.
- [70] Chen, Xiaohong, Zhipeng Liao, and Yixiao Sun (2012): "Sieve inference on semi-nonparametric time series models," Cowles Foundation Discussion Paper #1849.
- [71] Chen, Xiaohong and Demian Pouzo (2015): "Sieve Wald and QLR inferences on semi/nonparametric conditional moment models," *Econometrica*, 83, 1013-1079.
- [72] Claeskens, Gerda and Nils Lid Hjort (2003): "The focused information criterion," *Journal of the American Statistical Association*, 98, 900-945.
- [73] Claeskens, Gerda and Nils Lid Hjort (2008): *Model Selection and Model Averaging*, Cambridge University Press.
- [74] Conley, Timothy G. and Christopher R. Taber (2011): "Inference with 'difference in differences' with a small number of policy changes," *Review of Economics and Statistics*, 93, 113-125.

- [75] Cox, Donald, Bruce E. Hansen, and Emmanuel Jimenez (2004): "How responsive are private transfers to income? Evidence from a laissez-faire economy," *Journal of Public Economics*, 88, 2193-2219.
- [76] Cragg, John G. and Stephen G. Donald (1993): "Testing identifiability and specification in instrumental variable models," *Econometric Theory*, 9, 222-240.
- [77] Craven, Peter and Grace Wahba (1979): "Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation", *Numerische Mathematik*, 31, 377-403
- [78] Davidson, James (1994): *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford: Oxford University Press.
- [79] Davidson, James (2000): *Econometric Theory*, Blackwell Publishers.
- [80] Davidson, Russell and Emmanuel Flachaire (2008): "The wild bootstrap, tamed at last," *Journal of Econometrics*, 146, 162-169.
- [81] Davidson, Russell and James G. MacKinnon (1993): *Estimation and Inference in Econometrics*, Oxford University Press.
- [82] Davidson, Russell and James G. MacKinnon (2004): *Econometric Theory and Methods*, Oxford University Press.
- [83] Davison, A. C. and D. V. Hinkley (1997): *Bootstrap Methods and their Application*. Cambridge University Press.
- [84] De Luca, Giuseppe, Jan R. Magnus, and Franco Peracchi (2018): "Balanced variable addition in linear models" *Journal of Economic Surveys*, 31, 1183-1200.
- [85] de Moivre, Abraham (1733): "Approximatio ad Summam Terminorum Binomii  $(a+b)^n$  in Seriem expansi," manuscript.
- [86] Dickey, David A. and Wayne A. Fuller (1979): "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, 74, 427-431.
- [87] DiTella, Rafael and Ernesto Schargrodsky (2004): "Do police reduce crime? Estimates using the allocation of police forces after a terrorist attack," *American Economic Review*, 94, 115-138.
- [88] Donald, Stephen G. and Whitney K. Newey (2001): "Choosing the number of instruments," *Econometrica*, 69, 1161-1191.
- [89] Donohue, John J. III and Steven D. Levitt (2001): "The impact of legalized abortion on crime," *The Quarterly Journal of Economics*, 116, 379-420.
- [90] Duflo, Esther, Pascaline Dupas, and Michael Kremer (2011): "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya," *American Economic Review*, 101, 1739-1774.
- [91] Dufour, Jean-Marie (1997): "Some impossibility theorems in econometrics with applications to structural and dynamic models," *Econometrica*, 65, 1365-1387.
- [92] Durbin, James (1954): "Errors in variables," *Review of the International Statistical Institute*, 22, 23-32.
- [93] Durbin, James (1960): "The fitting of time-series models," *Revue de l'Institut International de Statistique*, 28, 233-44.

- [94] Efron, Bradley (1979): "Bootstrap methods: Another look at the jackknife," *Annals of Statistics*, 7, 1-26.
- [95] Efron, Bradley (1982): *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society for Industrial and Applied Mathematics.
- [96] Efron, Bradley (1987): "Better bootstrap confidence intervals (with discussion)", *Journal of the American Statistical Association*, 82, 171-200.
- [97] Efron, Bradley (2010): *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction*, Cambridge University Press.
- [98] Efron, Bradley and Trevor Hastie (2017): *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*, Cambridge University Press.
- [99] Efron, Bradley and Robert J. Tibshirani (1993): *An Introduction to the Bootstrap*, New York: Chapman-Hall.
- [100] Eichenbaum, Martin S., Lars Peter Hansen, and Kenneth J. Singleton (1988): "A time series analysis of representative agent models of consumption and leisure choice," *The Quarterly Journal of Economics*, 103, 51-78.
- [101] Eicker, F. (1963): "Asymptotic normality and consistency of the least squares estimators for families of linear regressions," *Annals of Mathematical Statistics*, 34, 447-456.
- [102] Elliott, Graham and Allan Timmermann (2016): *Economic Forecasting*, Princeton University Press.
- [103] Enders, Walter (2014): *Applied Economic Time Series, Fourth Edition*, Wiley.
- [104] Engle, Robert F and Clive W. J. Granger (1987): "Co-integration and error correction: Representation, estimation and testing," *Econometrica*, 55, 251-276.
- [105] Epanechnikov, V. I. (1969): "Non-parametric estimation of a multivariate probability density," *Theory of Probability and its Application*, 14, 153-158.
- [106] Fan, Jianqing (1992): "Design-adaptive nonparametric regression," *Journal of the American Statistical Association*, 87, 998-1004.
- [107] Fan, Jianqing (1993): "Local linear regression smoothers and their minimax efficiency," *Annals of Statistics*, 21, 196-216.
- [108] Fan, Jianqing and Irene Gijbels (1996): *Local Polynomial Modelling and Its Applications*, Chapman & Hall.
- [109] Fan, Jianqing and Qiwei Yao (1998): "Efficient estimation of conditional variance functions in stochastic regression," *Biometrika*, 85, 645-660.
- [110] Fan, Jianqing and Qiwei Yao (2003): *Nonlinear Time Series: Nonparametric and Parametric Methods*, New York: Springer-Verlag.
- [111] Foote, Christopher L. and Christopher F. Goetz (2008): "The impact of legalized abortion on crime: Comment," *The Quarterly Journal of Economics*, 123, 407-423.
- [112] Freyberger, Joachim (2017): "On completeness and consistency in nonparametric instrumental variable models," *Econometrica*, 85, 1629-1644.
- [113] Frisch, Ragnar (1933): "Editorial," *Econometrica*, 1, 1-4.

- [114] Frisch, Ragnar and Frederick V. Waugh (1933): "Partial time regressions as compared with individual trends," *Econometrica*, 1, 387-401.
- [115] Fuller, Wayne A. (1977): "Some properties of a modification of the limited information estimator," *Econometrica*, 45, 939-953.
- [116] Gallant, A. Ronald (1977): "Three-stage least-squares estimation for a system of simultaneous, nonlinear, implicit equations," *Journal of Econometrics*, 5, 71-88.
- [117] Gallant, A. Ronald (1997): *An Introduction to Econometric Theory*, Princeton University Press.
- [118] Gallant, A. Ronald and Dale W. Jorgenson (1979): "Statistical inference for a system of nonlinear, implicit equations in the context of instrumental variable estimation," *Journal of Econometrics*, 11, 275-302.
- [119] Galton, Francis (1886): "Regression Towards Mediocrity in Hereditary Stature," *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246-263.
- [120] Gardner, Robert (1998): "Unobservable individual effects in unbalanced panel data," *Economics Letters*, 58, 39-42.
- [121] Godambe, V. P. (1991): *Estimating Functions*, Oxford University Press, New York.
- [122] Goldberger, Arthur S. (1964): *Econometric Theory*, Wiley.
- [123] Goldberger, Arthur S. (1968): *Topics in Regression Analysis*, Macmillan.
- [124] Goldberger, Arthur S. (1991): *A Course in Econometrics*. Cambridge: Harvard University Press.
- [125] Goodnight, James H. (1979): "A tutorial on the SWEEP operator," *The American Statistician*, 33, 149-158.
- [126] Gosset, William S. (a.k.a. "Student") (1908): "The probable error of a mean," *Biometrika*, 6, 1-25.
- [127] Gauss, K. F. (1809): "Theoria motus corporum coelestium," in *Werke*, Vol. VII, 240-254.
- [128] Gourieroux, Christian (2000): *Econometrics of Qualitative Dependent Variables*, Cambridge University Press.
- [129] Granger, Clive W. J. (1969): "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, 37, 424-438.
- [130] Granger, Clive W. J. (1981): "Some properties of time series data and their use in econometric specification," *Journal of Econometrics*, 16, 121-130.
- [131] Granger, Clive W. J. (1989): *Forecasting in Business and Economics, Second Edition*, Academic Press.
- [132] Granger, Clive W. J. and Paul Newbold (1986): *Forecasting in Business and Economic Time Series, Second Edition*, Academic Press.
- [133] Greene, William H. (2017): *Econometric Analysis, Eighth Edition*, Pearson.
- [134] Gregory, A. and M. Veall (1985): "On formulating Wald tests of nonlinear restrictions," *Econometrica*, 53, 1465-1468.
- [135] Haagerup, Uffe (1982): "The best constants in the Khintchine inequality," *Studia Mathematica*, 70, 231-283.
- [136] Haavelmo, T. (1944): "The probability approach in econometrics," *Econometrica*, supplement, 12.

- [137] Hahn, Jinyong (1996): "A note on bootstrapping generalized method of moments estimators," *Econometric Theory*, 12, 187-197.
- [138] Hall, B. H. and R. E. Hall (1993): "The Value and Performance of U.S. Corporations" (1993) *Brookings Papers on Economic Activity*, 1-49.
- [139] Hall, Peter (1992): *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag.
- [140] Hall, Peter (1994): "Methodology and theory for the bootstrap," *Handbook of Econometrics, Vol. IV*, eds. R.F. Engle and D.L. McFadden. New York: Elsevier Science.
- [141] Hall, Peter, and C. C. Heyde (1980): *Martingale Limit Theory and Its Application*, Academic Press.
- [142] Hall, Peter and Joel L. Horowitz (1996): "Bootstrap critical values for tests based on generalized-method-of-moments estimation," *Econometrica*, 64, 891-916.
- [143] Hall, Robert E (1978): "Stochastic implications of the life cycle-permanent income hypothesis: theory and evidence," *Journal of Political Economy*, 86, 971-987.
- [144] Halmos, Paul R. (1956): *Lectures in Ergodic Theory*, Chelsea Publishing.
- [145] Hamilton, James D. (1994) *Time Series Analysis*, Princeton University Press.
- [146] Hansen, Bruce E. (1992): "Consistent covariance matrix estimation for dependent heterogenous processes," *Econometrica*, 60, 967-972.
- [147] Hansen, Bruce E. (1996): "Inference when a nuisance parameter is not identified under the null hypothesis," *Econometrica*, 64, 413-430.
- [148] Hansen, Bruce E. (2006): "Edgeworth expansions for the Wald and GMM statistics for nonlinear restrictions," *Econometric Theory and Practice: Frontiers of Analysis and Applied Research*, edited by Dean Corbae, Steven N. Durlauf and Bruce E. Hansen. Cambridge University Press.
- [149] Hansen, Bruce E. (2007): "Least squares model averaging," *Econometrica*, 75, 1175-1189.
- [150] Hansen, Bruce E. (2014): "Model averaging, asymptotic risk, and regressor groups," *Quantitative Economics*, 5, 495-530
- [151] Hansen, Bruce E. (2015): "The integrated mean squared error of series regression and a Rosenthal Hilbert-space inequality," *Econometric Theory*, 31, 337-361.
- [152] Hansen, Bruce E. and Seojeong Lee (2018): "Inference for iterated GMM under misspecification and clustering", working paper.
- [153] Hansen, Bruce E. and Jeffrey Racine (2012): "Jackknife model averaging," *Journal of Econometrics*, 167, 38-46.
- [154] Hansen, Christopher B. (2007): "Asymptotic properties of a robust variance matrix estimator for panel data when  $T$  is large," *Journal of Econometrics*, 141, 595-620.
- [155] Hansen, Lars Peter (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50, 1029-1054.
- [156] Hansen, Lars Peter and Robert J. Hodrick (1980): "Forward exchange rates as optimal predictors of future spot rates: An econometric analysis," *Journal of Political Economy*, 88, 829-853.
- [157] Hansen, Lars Peter, John Heaton, and A. Yaron (1996): "Finite sample properties of some alternative GMM estimators," *Journal of Business and Economic Statistics*, 14, 262-280.

- [158] Härdle, Wolfgang (1990): *Applied Nonparametric Regression*, Cambridge University Press.
- [159] Harvey, Andrew (1990): *The Econometric Analysis of Time Series, Second Edition*, MIT Press.
- [160] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2008): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.
- [161] Hausman, Jerry A. (1978): "Specification tests in econometrics," *Econometrica*, 46, 1251-1271.
- [162] Hausman, Jerry A., Whitney K. Newey, Tiemen Woutersen, John C. Chao, and Norman R. Swanson (2012): "Instrumental variable estimation with heteroskedasticity and many instruments," *Quantitative Economics*, 3, 211-255.
- [163] Hausman, Jerry A. and William E. Taylor (1981): "Panel data and unobservable individual effects," *Econometrica*, 49, 1377-1398.
- [164] Hayashi, Fumio (2000): *Econometrics*, Princeton University Press.
- [165] Heckman, James (1979): "Sample selection bias as a specification error," *Econometrica*, 47, 153-161.
- [166] Hinkley, D. V. (1977): "Jackknifing in unbalanced situations," *Technometrics*, 19, 285-292.
- [167] Hodges J. L. and E. L. Lehmann (1956): "The efficiency of some nonparametric competitors of the t-test," *Annals of Mathematical Statistics*, 27, 324-335.
- [168] Hoerl, A. E. and R. W. Kennard (1970): "Ridge regression: Biased estimation for non-orthogonal problems," *Technometrics*, 12, 55-67.
- [169] Holtz-Eakin, Douglas, Whitney Newey and Harvey S. Rosen (1988): "Estimating vector autoregressions with panel data," *Econometrica*, 56, 1371-1395.
- [170] Horn, S. D., R. A. Horn, and D. B. Duncan. (1975): "Estimating heteroscedastic variances in linear model," *Journal of the American Statistical Association*, 70, 380-385.
- [171] Horowitz, Joel (2001): "The Bootstrap," *Handbook of Econometrics*, Vol. 5, J.J. Heckman and E.E. Leamer, eds., Elsevier Science, 3159-3228.
- [172] Horowitz, Joel (2011): "Applied nonparametric instrumental variables estimation," *Econometrica*, 79, 347-394.
- [173] Hsiao, Cheng (2003): *Analysis of Panel Data, 2nd Edition*, Cambridge University Press.
- [174] Imbens, Guido W., and Joshua D. Angrist (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62, 467-476.
- [175] Jackson, D. (1912): "On the approximation by trigonometric sums and polynomials with positive coefficients," *TAMS*, 13, 491-515.
- [176] James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani (2013): *An Introduction to Statistical Learning: with Applications in R*, Springer.
- [177] James, W. and Charles M. Stein (1961): "Estimation with quadratic loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 361-380.
- [178] Johansen, Soren (1988): "Statistical analysis of cointegrating vectors," *Journal of Economic Dynamics and Control*, 12, 231-254.

- [179] Johansen, Soren (1991): "Estimation and hypothesis testing of cointegration vectors in the presence of linear trend," *Econometrica*, 59, 1551-1580.
- [180] Johansen, Soren (1995): *Likelihood-Based Inference in Cointegrated Vector Auto-Regressive Models*, Oxford University Press.
- [181] Johnston, Jack and John DiNardo (1997): *Econometric Methods: Fourth Edition*, McGraw-Hill.
- [182] Jones, M. C. and S. J. Sheather (1991): "Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives," *Statistics and Probability Letters*, 11, 511-514.
- [183] Jordà, Òscar (2005): "Estimation and inference of impulse responses by local projections," *American Economic Review*, 95, 161-182.
- [184] Judge, George G., W. E. Griffiths, R. Carter Hill, Helmut Lütkepohl, and Tsoung-Chao Lee (1985): *The Theory and Practice of Econometrics, Second Edition*, Wiley.
- [185] Keating, John W. (1992): "Structural approaches to vector autoregressions," *Federal Reserve Bank of St. Louis Review*, 74, 37-57.
- [186] Kilian, Lutz (2009): "Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market," *American Economic Review*, 99, 1053-1069.
- [187] Kilian, Lutz and Helmut Lütkepohl: (2017): *Structural Vector Autoregressive Analysis*, Cambridge University Press.
- [188] Kinal, Terrence W. (1980): "The existence of moments of k-class estimators," *Econometrica*, 48, 241-249.
- [189] Kleibergen, Frank and Richard Paap (2006): "Generalized reduced rank tests using the singular value decomposition," *Journal of Econometrics*, 133, 97-126.
- [190] Koenker, Roger (2005): *Quantile Regression*. Cambridge University Press.
- [191] Lafontaine, F. and K. J. White (1986): "Obtaining any Wald statistic you want," *Economics Letters*, 21, 35-40.
- [192] Legendre, Adrien-Marie (1805): *Nouvelles methodes pour la determination des orbites de cometes [New Methods for the Determination of the Orbits of Comets]*, Pris: F. Didot.
- [193] Lehmann, E. L. and George Casella (1998): *Theory of Point Estimation, 2<sup>nd</sup> Edition*, Springer.
- [194] Lehmann, E. L. and Joseph P. Romano (2005): *Testing Statistical Hypotheses, 3<sup>rd</sup> Edition*, Springer.
- [195] Lindeberg, Jarl Waldemar, (1922): "Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung," *Mathematische Zeitschrift*, 15, 211-225.
- [196] Li, Ker-Chau (1986): "Asymptotic optimality of  $C_L$  and generalized cross-validation in ridge regression with application to spline smoothing," *Annals of Statistics*, 14, pp. 1101-1112.
- [197] Li, Ker-Chau (1987): "Asymptotic optimality for  $C_p$ ,  $C_L$ , cross-validation and generalized cross-validation: Discrete Index Set," *Annals of Statistics*, 15, pp. 958-975.
- [198] Li, Qi and Jeffrey Racine (2007) *Nonparametric Econometrics*.
- [199] Linton, Oliver (2017): *Probability, Statistics, and Econometrics*, Academic Press.

- [200] Liu, R. Y. (1988): "Bootstrap procedures under some non-I.I.D. models," *Annals of Statistics*, 16, 1696-1708.
- [201] Lorentz, G. G. (1986): *Approximation of Functions*, Second Edition, New York: Chelsea.
- [202] Lovell, Michael C. (1963): "Seasonal adjustment of economic time series," *Journal of the American Statistical Association*, 58, 993-1010.
- [203] Lütkepohl, Helmut (2005): *New Introduction to Multiple Time Series Analysis*, Springer.
- [204] Lyapunov, Aleksandr (1901): *Nouvelle forme du théorème sur la limite de probabilité*.
- [205] MacKinnon, James G. and Halbert White (1985): "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, 29, 305-325.
- [206] Maddala, G. S. (1983): *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press.
- [207] Magnus, Jan R. (2017): *Introduction to the Theory of Econometrics*, VU University Press.
- [208] Magnus, Jan R., and H. Neudecker (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, New York: John Wiley and Sons.
- [209] Mallows, C. L. (1973). "Some comments on  $C_p$ ,". *Technometrics*, 15, 661-675.
- [210] Mammen, E. (1993): "Bootstrap and wild bootstrap for high dimensional linear models," *Annals of Statistics*, 21, 255-285.
- [211] Mankiw, N. Gregory, David Romer, and David N. Weil (1992): "A contribution to the empirics of economic growth," *The Quarterly Journal of Economics*, 107, 407-437.
- [212] Mann, H.B. and A. Wald (1943). "On stochastic limit and order relationships," *The Annals of Mathematical Statistics* 14, 217-226.
- [213] Mariano, R. S. (1982): "Analytical small-sample distribution theory in econometrics: the simultaneous equations case," *International Economic Review*, 23, 503-534.
- [214] McCracken, Michael W. and Serena Ng (2015): FRED-MD: "A monthly database for macroeconomic research," working paper 2015-012B, Federal Reserve Bank of St. Louis.
- [215] McCulloch, J. Huston (1985): "On heteros\*edasticity," *Econometrica*, 53, 483.
- [216] Mertens, Karel and Morten O. Ravn (2013): "The dynamic effects of personal and corporate income tax changes in the United States," *American Economic Review*, 103, 1212-1247.
- [217] Mhaskar, Hrushikesh N. (1996) Introduction to the theory of weighted polynomial approximation, World Scientific.
- [218] Moulton, Brent R. (1990): "An illustration of a pitfall in estimating the effects of aggregate variables on micro units," *Review of Economics and Statistics*, 72, 334-338.
- [219] Mundlak, Yair (1961): "Empirical production function free of management bias," *Journal of Farm Economics*, 43, 44-56.
- [220] Murphy, Kevin M. and Robert H. Topel (1985): "Estimation and inference in two-step econometric models," *Journal of Business and Economic Statistics*, 3, 370-379.
- [221] Nadaraya, E. A. (1964): "On estimating regression," *Theory of Probability and Its Applications*, 9, 141-142.

- [222] Nerlove, Marc (1963): "Returns to scale in electricity supply," Chapter 7 of *Measurement in Economics* (C. Christ, et al, eds.). Stanford: Stanford University Press, 167-198.
- [223] Newey, Whitney K. (1985): "Generalized method of moments specification testing," *Journal of Econometrics*, 29, 229-256.
- [224] Newey, Whitney K. (1990): "Semiparametric efficiency bounds," *Journal of Applied Econometrics*, 5, 99-135.
- [225] Newey, Whitney K. (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79, 147-168.
- [226] Newey, Whitney K. and Daniel L. McFadden (1994): "Large sample estimation and hypothesis testing," in Robert Engle and Daniel McFadden, (eds.) *Handbook of Econometrics*, vol. IV, 2111-2245, North Holland: Amsterdam.
- [227] Newey, Whitney K. and James L. Powell (2003): "Instrumental variable estimation of nonparametric models," *Econometrica*, 71, 1565-1578.
- [228] Newey, Whitney K. and Kenneth D. West (1987a): "Hypothesis testing with efficient method of moments estimation," *International Economic Review*, 28, 777-787.
- [229] Newey, Whitney K. and Kenneth D. West (1987b): "A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix," *Econometrica*, 55, 703-708.
- [230] Newey, Whitney K. and Kenneth D. West (1994): "Automatic lag selection in covariance matrix estimation," *Review of Economic Studies*, 631-654.
- [231] Nickell, Stephen (1981): "Biases in dynamic models with fixed effects," *Econometrica*, 49, 1417-1426.
- [232] Owen, Art B. (1988): "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, 75, 237-249.
- [233] Owen, Art B. (2001): *Empirical Likelihood*. New York: Chapman & Hall.
- [234] Pagan, Adrian (1984): "Econometric issues in the analysis of regressions with generated regressors," *International Economic Review*, 25, 221-247.
- [235] Pagan, Adrian (1986): "Two stage and related estimators and their applications," *Review of Economic Studies*, 53, 517-538.
- [236] Pagan, Adrian and Aman Ullah (1999): *Nonparametric Econometrics*, Cambridge University Press.
- [237] Park, Joon Y. and Peter C. B. Phillips (1988): "On the formulation of Wald tests of nonlinear restrictions," *Econometrica*, 56, 1065-1083.
- [238] Parzen, Emanuel (1962): "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, 33, 1065-1076.
- [239] Pham, Tuan D. and Lanh T. Tran (1985): "Some mixing properties of time series models," *Stochastic Processes and their Applications*, 19, 297-303.
- [240] Phillips, Alban W. (1958): "The relation between unemployment and the rate of change of money wage rates in the United Kingdom 1861–1957" *Economica*, 25, 283-299.
- [241] Phillips, G. D. A. and C. Hale (1977): "The bias of instrumental variable estimators of simultaneous equation systems," *International Economic Review*, 18, 219-228.

- [242] Phillips, Peter C. B. (1983): "Exact small sample theory in the simultaneous equations model," *Handbook of Econometrics, Volume 1*, edited by Z. Griliches and M. D. Intriligator, North-Holland.
- [243] Phillips, Peter C. B. and Sam Ouliaris (1990): "Asymptotic properties of residual based tests for cointegration," *Econometrica*, 58, 165-193.
- [244] Politis, Dimitris N. and Joseph P. Romano (1996): "The stationary bootstrap," *Journal of the American Statistical Association*, 89, 1303-1313.
- [245] Politis, Dimitris N., Joseph P. Romano, and Michael Wolf (1999): *Subsampling*, New York: Springer.
- [246] Quenouille, M. (1949): "Approximate tests of correlation in time series," *Journal of the Royal Statistical Society Series B*, 11, 18-84.
- [247] Ramanathan, Ramu (1993): *Statistical Methods in Econometrics*, Academic Press.
- [248] Ramey, Valerie A. (2016): "Macroeconomic shocks and their propagation," in *Handbook of Macroeconomics, Volume 2*, edited by John B. Taylor and Harald Uhlig, Elsevier.
- [249] Ramsey, James B. (1969): "Tests for specification errors in classical linear least-squares regression analysis," *Journal of the Royal Statistical Society Series B*, 31, 350-371.
- [250] Reiersøl, Olav (1945). *Confluence Analysis by Means of Instrumental Sets of Variables*.
- [251] Reinhart, Carmen M. and Kenneth S. Rogoff (2010): "Growth in a time of debt," *American Economic Review: Papers and Proceedings*, 573-578.
- [252] Rosenblatt, M. (1956): "Remarks on some non-parametric estimates of a density function," *Annals of Mathematical Statistics*, 27, 832-837.
- [253] Ruud, Paul A. (2000): *An Introduction to Classical Econometric Theory*, Oxford University Press.
- [254] Said, S. E. and D. A. Dickey (1984): "Testing for unit roots in autoregressive-moving average models of unknown order," *Biometrika*, 71, 599-608.
- [255] Samuelson, Paul A. (1939): "Interactions between the multiplier analysis and the principle of acceleration," *Review of Economic Statistics*, 21, 75-78.
- [256] Sargan, J. D. (1958): "The estimation of economic relationships using instrumental variables," *Econometrica*, 26, 393-415.
- [257] Schumaker, Larry L. (2007): *Spline Functions: Basic Theory, Third Edition*, Cambridge University Press.
- [258] Schwarz, G. (1978): "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461-464.
- [259] Scott, David W. (1992): *Multivariate Density Estimation*, Wiley.
- [260] Secrist, Horace (1933): *The Triumph of Mediocrity in Business*. Evanston: Northwestern University.
- [261] Shao, Jun (2003): *Mathematical Statistics*, 2nd edition, Springer.
- [262] Shao, Jun and D. Tu (1995): *The Jackknife and Bootstrap*. NY: Springer.
- [263] Shapiro, Matthew D. and Mark W. Watson (1988): "Sources of business cycle fluctuations," in Stanley Fischer, editor, *NBER Macroeconomics Annual*, MIT Press, 111-148.
- [264] Sheather, S. J. and M. C. Jones (1991): "A reliable data-based bandwidth selection method for kernel density estimation," *Journal of the Royal Statistical Society, Series B*, 53, 683-690.

- [265] Shibata, R. (1980): "Asymptotically efficient selection of the order of the model for estimating parameters of a linear process," *The Annals of Statistics*, 8, 147-164.
- [266] Silverman, B. W. (1986): *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
- [267] Sims, C. A. (1972): "Money, income and causality," *American Economic Review*, 62, 540-552.
- [268] Sims, C. A. (1980): "Macroeconomics and reality," *Econometrica*, 48, 1-48.
- [269] Snedecor, George W. (1934): *Calculation and Interpretation of Analysis of Variance and Covariance*, Collegiate Press.
- [270] Staiger, D. and James H. Stock (1997): "Instrumental variables regression with weak instruments," *Econometrica*, 65, 557-586.
- [271] Stock, James H. (1987): "Asymptotic properties of least squares estimators of cointegrating vectors," *Econometrica*, 55, 1035-1056.
- [272] Stock, James H. and Francesco Trebbi (2003). "Retrospectives: Who Invented Instrumental Variable Regression?", *Journal of Economic Perspectives*, 17, 177-194
- [273] Stock, James H. and Mark W. Watson (2007): "Why has U.S. inflation become harder to forecast?," *Journal of Money, Credit and Banking*, 39, 3-33.
- [274] Stock, James H. and Mark W. Watson (2008): "Heteroskedasticity-robust standard errors for fixed effects panel data regression, *Econometrica*, 76, 155-174.
- [275] Stock, James H. and Mark W. Watson (2012): "Disentangling the channels of the 2007-09 recession," *Brookings Papers on Economic Activity*, 81-135.
- [276] Stock, James H. and Mark W. Watson (2014): *Introduction to Econometrics*, 3<sup>rd</sup> edition, Pearson.
- [277] Stock, James H. and Mark W. Watson (2016): "Factor models and structural vector autoregressions in macroeconomics," in *Handbook of Macroeconomics, Volume 2*, edited by John B. Taylor and Harald Uhlig, Elsevier.
- [278] Stock, James H. and Jonathan H. Wright (2000): "GMM with weak identification," *Econometrica*, 68, 1055-1096.
- [279] Stock, James H. and Motohiro Yogo (2005): "Testing for weak instruments in linear IV regression," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg*, eds Donald W.K. Andrews and James H. Stock, Cambridge University Press, 80-108.
- [280] Stone, C. J. (1977): "Consistent nonparametric regression," *Annals of Statistics*, 5, 595-645.
- [281] Stone, Marshall H. (1948): "The Generalized Weierstrass Approximation Theorem," *Mathematics Magazine*, 21, 167-184.
- [282] Stout, William F. (1974): *Almost Sure Convergence*, Academic Press.
- [283] Theil, Henri. (1953): "Repeated least squares applied to complete equation systems," The Hague, Central Planning Bureau, mimeo.
- [284] Theil, Henri (1961): *Economic Forecasts and Policy*. Amsterdam: North Holland.
- [285] Theil, Henri. (1971): *Principles of Econometrics*, New York: Wiley.

- [286] Tikhonov, Andrey Nikolayevich (1943): "On the stability of inverse problems," *Doklady Akademii Nauk SSSR*, 39, 195-198.
- [287] Tobin, James (1958): "Estimation of relationships for limited dependent variables," *Econometrica*, 26, 24-36.
- [288] Tong, Howell (1990): *Non-linear Time Series: A Dynamical System Approach*, Oxford University Press.
- [289] Tukey, John (1958): "Bias and confidence in not quite large samples," *Annals of Mathematical Statistics*, 29, 614.
- [290] Tripathi, Gautam (1999): "A matrix extension of the Cauchy-Schwarz inequality," *Economics Letters*, 63, 1-3.
- [291] van der Vaart, A.W. (1998): *Asymptotic Statistics*, Cambridge University Press.
- [292] von Bahr, Bengt and Carl-Gustav Esseen (1965): "Inequalities for the  $r^{\text{th}}$  absolute moment of a sum of random variables,  $1 \leq r \leq 2$ ," *The Annals of Mathematical Statistics*, 36, 299-303.
- [293] Wald, Abraham. (1940): "The fitting of straight lines if both variables are subject to error," *The Annals of Mathematical Statistics*, 11, 283-300
- [294] Wald, Abraham. (1943): "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Transactions of the American Mathematical Society*, 54, 426-482.
- [295] Wansbeek, T. J. and A. Kapteyn (1989): "Estimation of the error components model with incomplete panels," *Journal of Econometrics*, 41, 341-349.
- [296] Wasserman, Larry (2006): *All of Nonparametric Statistics*, New York: Springer.
- [297] Watson, G. S. (1964): "Smooth regression analysis," *Sankya Series A*, 26, 359-372.
- [298] Watson, Mark W. (1994): "VARs and cointegration," in *Handbook of Econometrics, volume 4*, edited by Robert Engle and Daniel McFadden, North-Holland.
- [299] Weierstrass, K. (1885): "Über die analytische Darstellbarkeit sogenannter willkürlicher Funktionen einer reellen Veränderlichen," *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin*, 1885.
- [300] Windmeijer, Frank (2005): "A finite sample correction for the variance of linear efficient two-step GMM estimators", *Journal of Econometrics*, 126, 25-51.
- [301] White, Halbert (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 48, 817-838.
- [302] White, Halbert (1982): "Instrumental variables regression with independent observations," *Econometrica*, 50, 483-499.
- [303] White, Halbert (1984): *Asymptotic Theory for Econometricians*, Academic Press.
- [304] Whittle, P. (1960): "Bounds for the moments of linear and quadratic forms in independent variables," *Theory of Probability and Its Applications*, 5, 302-305.
- [305] Halbert White and Ian Domowitz (1984): "Nonlinear regression with dependent observations," *Econometrica*, 52, 143-162.

- [306] Wooldridge, Jeffrey M. (1995): "Score diagnostics for linear models estimated by two stage least squares," In *Advances in Econometrics and Quantitative Economics: Essays in honor of Professor C. R. Rao*, eds. G. S. Maddala, P.C.B. Phillipis, and T.N. Srinivasan, 66-87. Cambridge: Blackwell.
- [307] Wooldridge, Jeffrey M. (2010): *Econometric Analysis of Cross Section and Panel Data*, 2<sup>nd</sup> edition, MIT Press.
- [308] Wooldridge, Jeffrey M. (2015): *Introductory Econometrics: A Modern Approach*, 6<sup>th</sup> edition, South-western.
- [309] Working, Elmer J. (1927) "What Do Statistical 'Demand Curves' Show?" *Quarterly Journal of Economics*, 41, 212-35.
- [310] Wright, Philip G. (1915): "Moore's Economic Cycles," *Quarterly Journal of Economics*. 29, 631-641.
- [311] Wright, Philip G. (1928): *The Tariff on Animal and Vegetable Oils*, New York: MacMillan.
- [312] Wright, Sewell (1921): "Correlation and causation," *Journal of Agricultural Research*, 20, 557-585.
- [313] Wu, De-Min (1973): Alternative tests of independence between stochastic regressors and disturbances," *Econometrica*, 41, 733-750.
- [314] Zellner, Arnold. (1962): "An efficient method of estimating seemingly unrelated regressions, and tests for aggregation bias," *Journal of the American Statistical Association*, 57, 348-368.
- [315] Zhang, Fuzhen and Qingling Zhang (2006): "Eigenvalue inequalities for matrix product," *IEEE Transactions on Automatic Control*, 51, 1506-1509.