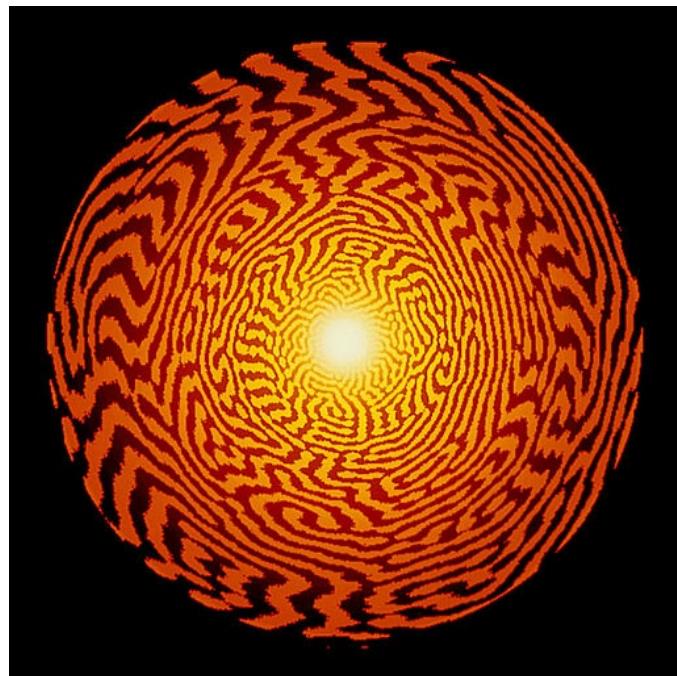


Forecasting  
*in Economics, Business, Finance and Beyond*



Francis X. Diebold  
University of Pennsylvania

Edition 2017  
Version Tuesday 1<sup>st</sup> August, 2017



Forecasting



# **Forecasting**

in Economics, Business, Finance and Beyond

**Francis X. Diebold**

Copyright © 2013-2017,  
by Francis X. Diebold.

This work is freely available for your use, but be warned: it is preliminary, incomplete, and evolving. It is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. (Briefly: I retain copyright, but you can use, copy and distribute non-commercially, so long as you give me attribution and do not modify. To view a copy of the license, go to <http://creativecommons.org/licenses/by-nc-nd/4.0/>.) In return I ask that you please cite the book whenever appropriate, as: “Diebold, F.X. (2017), *Forecasting*, Department of Economics, University of Pennsylvania, <http://www.ssc.upenn.edu/~fdiebold/Textbooks.html>.”

To three decades of wonderful students,  
who continue to inspire me





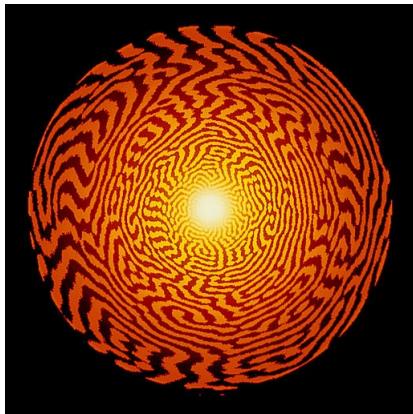
# About the Author



**Francis X. Diebold** is Paul F. and Warren S. Miller Professor of Economics, and Professor of Finance and Statistics, at the University of Pennsylvania and its Wharton School, as well as Faculty Research Associate at the National Bureau of Economic Research in Cambridge, Mass., and past President of the Society for Financial Econometrics. He has published nearly two hundred scientific papers in forecasting, econometrics, finance and macroeconomics, and he has served on the editorial boards of numerous scholarly and practitioner journals. He is an elected Fellow of the Econometric Society, the American Statistical Association, and the International Institute of Forecasters, and the recipient of Sloan, Guggenheim, and Humboldt fellowships. Diebold lectures actively, worldwide, and has received several prizes for outstanding teaching. He has held visiting appointments in Economics and Finance at Princeton University, Cambridge University, the University of Chicago, the London School of Economics, Johns Hopkins University, and New York University. Diebold's research and teaching are firmly rooted in policy and industry; during 1986-1989 he served as an economist under Paul Volcker and

Alan Greenspan at the Board of Governors of the Federal Reserve System in Washington DC, during 2007-2008 he served as an Executive Director at Morgan Stanley Investment Management, and during 2012-2013 he served as Chairman of the Federal Reserve System's Model Validation Council. All his degrees are from the University of Pennsylvania; he received his B.S. from the Wharton School in 1981 and his Economics Ph.D. in 1986. He is married with three children and lives in suburban Philadelphia.

# About the Cover



The graphic, “Hallucination” by Mario Marcus, is from Wikimedia Commons. I used it both simply because I like it and because it is reminiscent of a crystal ball.

For details see [http://commons.wikimedia.org/wiki/File:Mario\\_Markus--HALLUCIN.jpg](http://commons.wikimedia.org/wiki/File:Mario_Markus--HALLUCIN.jpg). The complete attribution is: Prof. Dr. Mario Markus [GFDL (<http://www.gnu.org/copyleft/fdl.html>), CC-BY-SA-3.0 (<http://creativecommons.org/licenses/by-sa/3.0>) or CC-BY-SA-3.0-de (<http://creativecommons.org/licenses/by-sa/3.0/de/deed.en>)], via Wikimedia Commons.



# Brief Table of Contents

<b>About the Author</b>	<b>ix</b>
<b>About the Cover</b>	<b>xi</b>
<b>Preface</b>	<b>xxix</b>
<b>I Getting Started</b>	<b>1</b>
<b>1 Introduction</b>	<b>3</b>
<b>2 Universal Considerations</b>	<b>21</b>
<b>II Cross Sections: Basics</b>	<b>47</b>
<b>3 Predictive Regression: Review and Interpretation</b>	<b>49</b>
<b>4 Forecast Model Building and Use</b>	<b>77</b>
<b>III Time Series: A Components Perspective</b>	<b>109</b>
<b>5 Trend and Seasonality</b>	<b>111</b>
<b>6 Cycles I: Autoregressions and Wold's Chain Rule</b>	<b>141</b>
<b>7 Cycles II: The Wold Representation and Its Approximation</b>	<b>199</b>
<b>8 Noise: Conditional Variance Dynamics</b>	<b>273</b>

<b>9 Assembling the Components</b>	<b>311</b>
<b>IV Forecast Evaluation and Combination</b>	<b>331</b>
<b>10 Point Forecast Evaluation</b>	<b>333</b>
<b>11 Interval and Density Forecast Evaluation</b>	<b>383</b>
<b>12 Model-Based Forecast Combination</b>	<b>397</b>
<b>13 Market-Based Forecast Combination</b>	<b>421</b>
<b>14 Survey-Based Forecast Combination</b>	<b>437</b>
<b>V More</b>	<b>443</b>
<b>15 Selection, Shrinkage, and Distillation</b>	<b>445</b>
<b>16 Multivariate: Vector Autoregression</b>	<b>461</b>
<b>VI Appendices</b>	<b>507</b>
<b>A Elements of Probability and Statistics</b>	<b>509</b>
<b>B Elements of Nonparametrics</b>	<b>523</b>
<b>C Problems and Complements Data</b>	<b>535</b>
<b>D Some Pop and “Cross-Over” Books and Sites Worth Examining</b>	<b>575</b>
<b>E Construction of the Wage Datasets</b>	<b>577</b>

# Detailed Table of Contents

<b>About the Author</b>	<b>ix</b>
<b>About the Cover</b>	<b>xi</b>
<b>Preface</b>	<b>xxix</b>
<b>I Getting Started</b>	<b>1</b>
<b>1 Introduction</b>	<b>3</b>
1.1 Welcome . . . . .	3
1.2 Who Forecasts, and Why? . . . . .	4
1.3 Useful Materials . . . . .	9
1.3.1 Books . . . . .	9
1.3.2 Online Information and Data . . . . .	10
1.3.3 Software (and a Tiny bit of Hardware) . . . . .	11
1.3.4 Journals and Professional Organizations . . . . .	13
1.4 Final Thoughts . . . . .	13
1.5 Tips on How to use this book . . . . .	14
1.6 Exercises, Problems and Complements . . . . .	17
<b>2 Universal Considerations</b>	<b>21</b>
2.1 The Forecast Object . . . . .	23
2.2 The Information Set . . . . .	25
2.2.1 Univariate vs. Multivariate . . . . .	25
2.2.2 Expert Opinion and Judgment . . . . .	26
2.2.3 Information Sets in Forecast Evaluation . . . . .	26
2.3 Model Uncertainty and Improvement . . . . .	27
2.4 The Forecast Horizon . . . . .	27

2.4.1	<i>h</i> -Step-Ahead Forecasts . . . . .	27
2.4.2	<i>h – Step</i> Ahead Path Forecasts . . . . .	28
2.4.3	Nowcasting and Backcasting . . . . .	28
2.5	Structural Change . . . . .	30
2.6	The Forecast Statement . . . . .	30
2.6.1	Time Series . . . . .	30
2.6.2	Events . . . . .	34
2.6.3	Probability Forecasts as Point and/or Density Forecasts	35
2.7	Forecast Presentation . . . . .	35
2.7.1	Graphics for Forecasts . . . . .	35
2.7.2	Graphics for Forecast Evaluation . . . . .	35
2.8	The Decision Environment and Loss Function . . . . .	35
2.8.1	Loss Functions . . . . .	35
2.8.2	Optimal Forecasts with Respect to a Loss Function . .	37
2.8.3	State-Dependent Loss . . . . .	39
2.9	Model Complexity and the Parsimony Principle . . . . .	39
2.10	Unobserved Components . . . . .	41
2.11	Concluding Remarks . . . . .	41
2.12	Exercises, Problems and Complements . . . . .	42
<b>II</b>	<b>Cross Sections: Basics</b>	<b>47</b>
<b>3</b>	<b>Predictive Regression: Review and Interpretation</b>	<b>49</b>
3.1	Regression as Curve Fitting . . . . .	49
3.1.1	Simple Regression . . . . .	49
3.1.2	Multiple Regression . . . . .	52
3.2	Regression as a Probability Model . . . . .	53
3.2.1	A Population Model and a Sample Estimator . . . . .	53
3.2.2	Notation, Assumptions and Results: The Full Ideal Conditions . . . . .	54
	A Bit of Matrix Notation . . . . .	54
	Assumptions: The Full Ideal Conditions (FIC) . . . . .	55
	Results Under the FIC . . . . .	56
3.3	A Typical Regression Analysis . . . . .	56
3.3.1	Coefficient Estimates, Standard Errors, <i>t</i> Statistics and <i>p</i> -Values . . . . .	57

3.3.2	Residual Plot . . . . .	60
3.3.3	Mean dependent var . . . . .	61
3.3.4	S.D. dependent var . . . . .	61
3.3.5	Sum squared resid . . . . .	62
3.3.6	F-statistic . . . . .	62
3.3.7	Prob(F-statistic) . . . . .	63
3.3.8	S.E. of regression . . . . .	63
3.3.9	R-squared . . . . .	64
3.3.10	Adjusted R-squared . . . . .	65
3.3.11	Durbin-Watson stat . . . . .	65
3.3.12	Akaike info criterion and Schwarz criterion . . . . .	66
3.3.13	Log Likelihood . . . . .	67
3.4	Regression From a Forecasting Perspective . . . . .	67
3.4.1	The Key to Everything (or at Least Many Things) . .	67
3.4.2	Why Take a Probabilistic Approach to Regression, as Opposed to Pure Curve Fitting? . . . . .	69
3.4.3	Regression For Estimating Conditional Means is Re- gression for Forecasting . . . . .	69
3.4.4	LS and Quadratic Loss . . . . .	70
3.4.5	Estimated Coefficient Signs and Sizes . . . . .	70
3.4.6	Standard Errors, <i>t</i> Statistics, <i>p</i> -values, <i>F</i> Statistic, Log Likelihood, etc. . . . .	70
3.4.7	Fitted Values and Residuals . . . . .	70
3.4.8	Mean and Variance of Dependent Variable . . . . .	71
3.4.9	$R^2$ and $\bar{R}^2$ . . . . .	71
3.4.10	<i>SSR</i> (or <i>MSE</i> ), <i>SER</i> (or Residual $s^2$ ), <i>AIC</i> and <i>SIC</i>	72
3.4.11	Durbin-Watson . . . . .	72
3.4.12	Residual Plots . . . . .	73
3.5	Exercises, Problems and Complements . . . . .	73
<b>4</b>	<b>Forecast Model Building and Use</b>	<b>77</b>
4.1	Cross-Section Prediction . . . . .	77
4.1.1	Point Prediction . . . . .	78
4.1.2	Density Prediction for <i>D</i> Gaussian . . . . .	78
4.1.3	Density Prediction for <i>D</i> Parametric Non-Gaussian .	79
4.1.4	Making the Forecasts Feasible . . . . .	79

4.1.5	Density Prediction for $D$ Non-Parametric . . . . .	80
4.1.6	Density Forecasts for $D$ Nonparametric and Acknowledging Parameter Estimation Uncertainty . . . . .	80
4.1.7	Incorporating Heteroskedasticity . . . . .	81
4.2	Wage Prediction Conditional on Education and Experience . . . . .	82
4.2.1	The CPS Dataset . . . . .	82
4.2.2	Regression . . . . .	84
4.2.3	Point Prediction by Exponentiating vs. Simulation . . . . .	85
4.2.4	Density Prediction for $D$ Gaussian . . . . .	86
4.2.5	Density Forecasts for $D$ Gaussian and Acknowledging Parameter Estimation Uncertainty . . . . .	87
4.2.6	Density Forecasts for $D$ Gaussian, Acknowledging Parameter Estimation Uncertainty, and Allowing for Heteroskedasticity . . . . .	88
4.2.7	Density Prediction for $D$ Nonparametric . . . . .	92
4.2.8	Density Forecasts for $D$ Nonparametric and Acknowledging Parameter Estimation Uncertainty . . . . .	93
4.2.9	Modeling Directly in Levels . . . . .	94
4.3	Non-Parametric Estimation of Conditional Mean Functions . . . . .	96
4.3.1	Global Nonparametric Regression: Series . . . . .	96
	The Curse of Dimensionality . . . . .	97
	Bandwidth Selection and the Bias-Variance Tradeoff . . . . .	98
4.3.2	Local Nonparametric Regression: Nearest-Neighbor . . . . .	98
	Unweighted Locally-Constant Regression . . . . .	98
	Weighted Locally-Linear Regression . . . . .	99
	“Robustness Iterations” . . . . .	100
4.3.3	Forecasting Perspectives . . . . .	101
	On Global vs. Local Smoothers for Forecasting . . . . .	101
	Nearest Neighbors as a General Forecasting Method . . . . .	102
4.4	Wage Prediction, Continued . . . . .	102
4.4.1	Point Wage Prediction . . . . .	102
4.4.2	Density Wage Prediction . . . . .	102
4.5	Exercises, Problems and Complements . . . . .	102
4.6	Notes . . . . .	107

<b>III Time Series: A Components Perspective</b>	<b>109</b>
<b>5 Trend and Seasonality</b>	<b>111</b>
5.1 The Forecasting the Right-Hand-Side Variables (FRV) Problem	111
5.2 Deterministic Trend . . . . .	113
5.2.1 Trend Models . . . . .	113
5.2.2 Trend Estimation . . . . .	116
5.2.3 Forecasting Trends . . . . .	118
5.2.4 Forecasting Retail Sales . . . . .	120
5.3 Deterministic Seasonality . . . . .	127
5.3.1 Seasonal Models . . . . .	128
5.3.2 Seasonal Estimation . . . . .	129
5.3.3 Forecasting Seasonals . . . . .	130
5.3.4 Forecasting Housing Starts . . . . .	131
5.4 Exercises, Problems and Complements . . . . .	137
5.5 Notes . . . . .	140
<b>6 Cycles I: Autoregressions and Wold's Chain Rule</b>	<b>141</b>
6.1 Characterizing Cycles . . . . .	142
6.1.1 Covariance Stationary Time Series . . . . .	142
Basic Ideas . . . . .	142
6.2 White Noise . . . . .	150
6.2.1 Basic Ideas . . . . .	150
6.3 Estimation and Inference for the Mean, Autocorrelation and Partial Autocorrelation Functions . . . . .	156
6.3.1 Sample Mean . . . . .	156
6.3.2 Sample Autocorrelations . . . . .	157
6.3.3 Sample Partial Autocorrelations . . . . .	159
6.4 Canadian Employment I: Characterizing Cycles . . . . .	161
6.5 Modeling Cycles With Autoregressions . . . . .	164
6.5.1 Some Preliminary Notation: The Lag Operator . . . . .	164
6.5.2 Autoregressive Processes . . . . .	165
6.5.3 Autoregressive Disturbances and Lagged Dependent Variables . . . . .	166
The $AR(1)$ Process for Observed Series . . . . .	167
6.5.4 The $AR(p)$ Process . . . . .	176
6.6 Canadian Employment II: Modeling Cycles . . . . .	178

6.7	Forecasting Cycles with Autoregressions . . . . .	181
6.7.1	On the FRV Problem . . . . .	181
6.7.2	Information Sets, Conditional Expectations, and Linear Projections . . . . .	181
6.7.3	Point Forecasts for Autoregressions: Wold's Chain Rule	183
6.7.4	Density Forecasts . . . . .	184
6.8	Canadian Employment III: Forecasting . . . . .	186
6.9	Exercises, Problems and Complements . . . . .	190
6.10	Notes . . . . .	198
<b>7</b>	<b>Cycles II: The Wold Representation and Its Approximation</b>	<b>199</b>
7.1	The Wold Representation and the General Linear Process . . .	199
7.1.1	The Wold Representation . . . . .	199
7.1.2	The General Linear Process . . . . .	201
7.2	Approximating the Wold Representation . . . . .	202
7.2.1	Rational Distributed Lags . . . . .	203
7.2.2	Moving Average ( <i>MA</i> ) Models . . . . .	205
	The <i>MA</i> (1) Process . . . . .	205
	The <i>MA</i> ( <i>q</i> ) Process . . . . .	213
7.2.3	Autoregressive ( <i>AR</i> ) Models . . . . .	215
	The <i>AR</i> (1) Process . . . . .	215
	The <i>AR</i> ( <i>p</i> ) Process . . . . .	222
7.2.4	Autoregressive Moving Average ( <i>ARMA</i> ) Models . . .	226
7.3	Forecasting Cycles From a Moving-Average Perspective: Wiener-Kolmogorov . . . . .	228
7.3.1	Optimal Point Forecasts for Finite-Order Moving Averages . . . . .	230
7.3.2	Optimal Point Forecasts for Infinite-Order Moving Averages . . . . .	233
7.3.3	Interval and Density Forecasts . . . . .	235
7.3.4	Making the Forecasts Operational . . . . .	236
7.4	Forecasting Cycles From an Autoregressive Perspective: Wold's Chain Rule . . . . .	238
7.4.1	Point Forecasts of Autoregressive Processes . . . . .	238
7.4.2	Point Forecasts of ARMA processes . . . . .	240
7.4.3	Interval and Density Forecasts . . . . .	242

7.5	Canadian Employment . . . . .	243
7.6	Exercises, Problems and Complements . . . . .	264
7.7	Notes . . . . .	271
<b>8</b>	<b>Noise: Conditional Variance Dynamics</b>	<b>273</b>
8.1	The Basic ARCH Process . . . . .	274
8.2	The GARCH Process . . . . .	280
8.3	Extensions of ARCH and GARCH Models . . . . .	287
8.3.1	Asymmetric Response . . . . .	287
8.3.2	Exogenous Variables in the Volatility Function . . . . .	288
8.3.3	Regression with GARCH disturbances and GARCH-M	289
8.3.4	Component GARCH . . . . .	289
8.3.5	Mixing and Matching . . . . .	290
8.4	Estimating, Forecasting and Diagnosing GARCH Models . . .	290
8.5	Application: Stock Market Volatility . . . . .	293
8.6	Exercises, Problems and Complements . . . . .	303
8.7	Notes . . . . .	309
<b>9</b>	<b>Assembling the Components</b>	<b>311</b>
9.1	Serially Correlated Disturbances . . . . .	312
9.2	Lagged Dependent Variables . . . . .	314
9.2.1	Case Study: Forecasting Liquor Sales with Deterministic Trends and Seasonals . . . . .	314
9.3	Exercises, Problems and Complements . . . . .	328
9.4	Notes . . . . .	329
<b>IV</b>	<b>Forecast Evaluation and Combination</b>	<b>331</b>
<b>10</b>	<b>Point Forecast Evaluation</b>	<b>333</b>
10.1	Absolute Standards for Point Forecasts . . . . .	333
10.1.1	Are errors zero-mean? . . . . .	335
10.1.2	Are 1-step-ahead errors white noise? . . . . .	335
10.1.3	Are $h$ -step-ahead errors are at most $MA(h - 1)$ ? . . .	336
10.1.4	Are $h$ -step-ahead error variances non-decreasing in $h$ ? .	336
10.1.5	Are errors orthogonal to available information? . . . . .	336
10.2	Relative Standards for Point Forecasts . . . . .	338

10.2.1 Accuracy Rankings via Expected Loss . . . . .	338
10.2.2 On MSE vs. MAE . . . . .	341
10.2.3 Benchmark Comparisons . . . . .	343
Predictive $R^2$ . . . . .	343
Theil's U-Statistic . . . . .	343
10.2.4 Measures of Forecastability . . . . .	344
Population Measures . . . . .	345
Sample Measures . . . . .	348
10.2.5 Statistical Assessment of Accuracy Rankings . . . . .	349
A Motivational Example . . . . .	349
The Diebold-Mariano Perspective . . . . .	350
Thoughts on Assumption <i>DM</i> . . . . .	352
10.3 OverSea Shipping . . . . .	353
10.4 Exercises, Problems and Complements . . . . .	377
10.5 Notes . . . . .	382
<b>11 Interval and Density Forecast Evaluation</b>	<b>383</b>
11.1 Interval Forecast Evaluation . . . . .	383
11.1.1 Absolute Standards . . . . .	383
On Correct Unconditional vs. Conditional Coverage .	383
Christoffersen's Absolute Interval Forecast Evaluation .	384
On Testing <i>iid</i> in Forecast Evaluation . . . . .	385
11.1.2 Relative Standards . . . . .	386
11.2 Density Forecast Evaluation . . . . .	386
11.2.1 Absolute Standards . . . . .	386
Theory . . . . .	386
Practical Application . . . . .	388
11.2.2 Additional Discussion . . . . .	389
Parameter Estimation Uncertainty . . . . .	389
Improving Mis-Calibrated Density Forecasts . . . . .	390
Multi-Step Density Forecasts . . . . .	390
11.2.3 Relative Standards . . . . .	391
11.3 Stock Return Density Forecasting . . . . .	391
11.3.1 A Preliminary GARCH Simulation . . . . .	391
11.3.2 Daily S&P 500 Returns . . . . .	394
11.4 Exercises, Problems and Complements . . . . .	395

11.5 Notes . . . . .	395
<b>12 Model-Based Forecast Combination</b>	<b>397</b>
12.1 Forecast Encompassing . . . . .	397
12.2 Variance-Covariance Forecast Combination . . . . .	399
12.2.1 Bivariate Case . . . . .	399
12.2.2 General Case . . . . .	402
12.3 Regression-Based Forecast Combination . . . . .	402
12.3.1 Time-Varying Combining Weights . . . . .	403
12.3.2 Serial Correlation . . . . .	404
12.3.3 Shrinkage of Combining Weights Toward Equality . .	405
12.3.4 Nonlinear Combining Regressions . . . . .	406
12.3.5 Regularized Regression for Combining Large Numbers of Forecasts . . . . .	406
12.4 Application: OverSea Shipping Volume Revisited . . . . .	406
12.5 On the Optimality of Equal Weights . . . . .	408
12.5.1 Under Quadratic Loss . . . . .	408
12.5.2 Under Minimax Loss . . . . .	411
12.6 Interval Forecast Combination . . . . .	415
12.7 Density Forecast Combination . . . . .	415
12.7.1 Choosing Weights to Optimize a Predictive Likelihood	415
12.7.2 Choosing Weights Optimize Conditional Calibration .	415
12.8 Exercises, Problems and Complements . . . . .	415
12.9 Notes . . . . .	418
<b>13 Market-Based Forecast Combination</b>	<b>421</b>
13.1 Financial Markets . . . . .	421
13.1.1 General Principles . . . . .	422
Point Forecasts From Forward Markets . . . . .	422
Point Forecasts From Futures Markets . . . . .	422
Density Forecasts From Options Markets (Using Sets of Options) . . . . .	423
Event Probability Forecasts From Digital Options Mar- kets . . . . .	423
Density Forecasts From Digital Options Markets (Us- ing Sets of Digital Options) . . . . .	423
13.1.2 More . . . . .	423

Volatility Forecasts From Options Markets . . . . .	423
Correlation Forecasts From Trios of Implied Volatilities	424
Skewness Forecasts From Risk Reversals . . . . .	425
Inflation Forecasts From Indexed Bonds . . . . .	425
Inflation Forecasts from Bond Yields . . . . .	425
Bond Yield Forecasts From the Term Premium . . . .	425
Real Activity Forecasts From the Term Premium . . .	426
Real Activity Forecasts From the Default Premium . .	426
Long-Run Equity Return Forecasts from the Dividend Yield . . . . .	427
13.2 “Prediction Markets” . . . . .	427
13.2.1 Arrow-Debreu Contingent Claims . . . . .	427
13.2.2 Parimutual Betting Markets . . . . .	427
13.3 Issues with Market-Based Forecasts . . . . .	427
13.3.1 Market Inefficiencies and No-Arbitrage Conditions . . .	428
13.3.2 Moral Hazard and Market Manipulation . . . . .	428
13.3.3 True Moral Issues . . . . .	428
13.3.4 Risk Neutrality . . . . .	428
13.3.5 Beyond Risk Neutrality . . . . .	429
13.3.6 A Bit More on Market Efficiency . . . . .	429
13.4 Exercises, Problems and Complements . . . . .	430
13.5 Notes . . . . .	432
<b>14 Survey-Based Forecast Combination</b>	<b>437</b>
14.1 Survey-Based Point Forecast Combination . . . . .	437
14.1.1 Surveys and the Wisdom of Crowds . . . . .	438
14.1.2 Delphi, Focus Groups, and Related Methods . . . .	438
14.1.3 Cross-Sectional Forecast Dispersion vs. True Uncertainty	438
14.2 Survey-Based Density Forecast Combination . . . . .	439
14.3 Exercises, Problems and Complements . . . . .	439
14.4 Notes . . . . .	441
<b>V More</b>	<b>443</b>
<b>15 Selection, Shrinkage, and Distillation</b>	<b>445</b>
15.1 All-Subsets Model Selection I: Information Criteria . . . . .	445

15.2 All-Subsets Model Selection II: Cross Validation . . . . .	452
15.3 Stepwise Selection . . . . .	453
15.3.1 Forward . . . . .	453
15.3.2 Backward . . . . .	454
15.4 One-Shot Estimation: Bayesian Shrinkage . . . . .	454
15.5 One-Shot Estimation: Selection <i>and</i> Shrinkage . . . . .	455
15.5.1 Penalized Estimation . . . . .	455
15.5.2 The Lasso . . . . .	455
Elastic Net . . . . .	456
Adaptive Lasso . . . . .	458
Adaptive Elastic Net . . . . .	458
15.6 Distillation: Principal Components . . . . .	458
15.6.1 Distilling “X Variables” into Principal Components . .	458
15.6.2 Principal Components Regression . . . . .	459
15.7 Exercises, Problems and Complements . . . . .	459
15.8 Notes . . . . .	460
<b>16 Multivariate: Vector Autoregression</b>	<b>461</b>
16.1 Distributed Lag Models . . . . .	462
16.2 Regressions with Lagged Dependent Variables, and Regres- sions with <i>ARMA</i> Disturbances . . . . .	464
16.3 Vector Autoregressions . . . . .	467
16.4 Predictive Causality . . . . .	469
16.5 Impulse-Response Functions . . . . .	472
16.6 Variance Decompositions . . . . .	476
16.7 Application: Housing Starts and Completions . . . . .	477
16.8 Exercises, Problems and Complements . . . . .	493
16.9 Notes . . . . .	501
<b>VI Appendices</b>	<b>507</b>
<b>A Elements of Probability and Statistics</b>	<b>509</b>
A.1 Populations: Random Variables, Distributions and Moments .	509
A.1.1 Univariate . . . . .	509
A.1.2 Multivariate . . . . .	512
A.2 Samples: Sample Moments . . . . .	513

A.2.1	Univariate . . . . .	513
A.2.2	Multivariate . . . . .	516
A.3	Finite-Sample and Asymptotic Sampling Distributions of the Sample Mean . . . . .	516
A.3.1	Exact Finite-Sample Results . . . . .	517
A.3.2	Approximate Asymptotic Results (Under Weaker Assumptions) . . . . .	518
A.4	Exercises, Problems and Complements . . . . .	519
A.5	Notes . . . . .	521
<b>B</b>	<b>Elements of Nonparametrics</b>	<b>523</b>
B.1	Density Estimation . . . . .	523
B.1.1	The Basic Problem . . . . .	523
B.1.2	Kernel Density Estimation . . . . .	524
B.1.3	Bias-Variance Tradeoffs . . . . .	525
	Inescapable Bias-Variance Tradeoff (in Practice, Fixed $N$ ) . . . . .	525
	Escapable Bias-Variance Tradeoff (in Theory, $N \rightarrow \infty$ ) . . . . .	525
	Convergence Rate . . . . .	525
B.1.4	Optimal Bandwidth Choice . . . . .	526
B.2	Multivariate . . . . .	527
B.3	Functional Estimation . . . . .	529
B.4	Local Nonparametric Regression . . . . .	530
B.4.1	Kernel Regression . . . . .	530
B.4.2	Nearest-Neighbor Regression . . . . .	530
	Basic Nearest-Neighbor Regression . . . . .	530
	Locally-Weighted Nearest-Neighbor Regression (Locally Polynomial, Non-Uniform Weighting) . . . . .	531
B.5	Global Nonparametric Regression . . . . .	531
B.5.1	Series (Sieve, Projection, ...) . . . . .	531
B.5.2	Neural Networks . . . . .	532
B.5.3	More . . . . .	532
B.6	Time Series Aspects . . . . .	532
B.7	Exercises, Problems and Complements . . . . .	533
B.8	Notes . . . . .	533

<b>C Problems and Complements Data</b>	<b>535</b>
C.1 Liquor Sales . . . . .	535
C.2 Housing Starts and Completions . . . . .	536
C.3 Shipping Volume . . . . .	548
C.4 Hungarian Exchange Rate . . . . .	565
C.5 Eurostar . . . . .	568
C.6 BankWire Transfers . . . . .	568
C.7 Nile.com Hits . . . . .	569
C.8 Thompson Energy Investors . . . . .	570
<b>D Some Pop and “Cross-Over” Books and Sites Worth Examining</b>	<b>575</b>
<b>E Construction of the Wage Datasets</b>	<b>577</b>



# Preface

Most good texts arise from the desire to leave one's stamp on a discipline by training future generations of students, coupled with the recognition that existing texts are inadequate in various respects. That was certainly the motivation behind my earlier *Elements of Forecasting* ("Elements"), and *Elements* helped train so many students, going through four successful editions during fifteen years.

But I have refused to do a fifth edition; instead, I feel that it's time to begin afresh. Two key reasons motivate the new start. The first is intellectual. Forecasting has changed tremendously in recent decades, and continually patching an old book only works for so long. This new book ("Forecasting") contains a wealth of new material and new visions, newly synthesized.

The second reason is technological. I want a book alive with color photos and graphics, extensively hyperlinked, with audio and video. I want to be able to update it continuously and distribute it instantly. And I want it to be widely affordable, \$29 (say), not \$290, or better yet, free. In short, I want my readers to escape the shackles of Middle Ages printing-press technology, benefiting instead from the pedagogical wonders of modern e-technology.

Beyond new structure, new and more advanced material, and e-awareness, a number of features distinguish *Forecasting*, many of which were shared by the earlier *Elements*. First, *Forecasting* does not attempt to be exhaustive in coverage. In fact, the coverage is intentionally selective, focusing on the core techniques with the widest applicability. It is designed so that its earlier chapters can be realistically covered in a one-semester course, with the remaining chapters of use for more advanced courses and for independent study. Core material appears in the main text of the various chapters, and additional material that expands on the depth and breadth of coverage is provided in the Exercises, Problems and Complements (EPC) at the end of

each chapter.

Second, *Forecasting* is applications-oriented. It illustrates all methods with detailed real-world applications that reflect typical forecasting situations. In many chapters, the application is the centerpiece of the presentation. In various places, it uses applications not simply to illustrate the methods but also to drive home an important lesson via truly realistic examples: not everything works perfectly in the real world!

Third, *Forecasting* is in touch with modern modeling and forecasting software. I supply some code in EViews, R and Python. I like all of them, but at the same time, nothing is wed to any particular software. Students and instructors can use whatever computing environment they like best.

Drafts of *Forecasting*, like the many editions of the earlier *Elements*, have found wide use among students in many fields, including economics, business, finance, public policy, statistics, and even engineering. It is directly accessible at the undergraduate and master's levels; the only prerequisite is an introductory statistics course that includes linear regression. Simultaneously *Forecasting* will also be of interest to those with more advanced preparation, because of the hard-to-find direct focus on forecasting – as opposed, for example, to general statistics, econometrics, or time series analysis. I have used the material successfully for many years as a background for various other undergraduate and graduate courses (including Ph.D.), and as the primary material for master's-level Executive Education courses given to professionals in business, finance, economics and government.

Many coauthors, colleagues and students contributed to the development some explicitly, some implicitly. The National Science Foundation, the Wharton Financial Institutions Center, and the Guggenheim Foundation provided financial support for much of the underlying research. The University of Pennsylvania provided an unparalleled 25-year intellectual home, the perfect incubator for the ideas that have congealed here.

My hope is that if you liked *Elements*, you'll love *Forecasting*, sharing with me the excitement of the rapidly-evolving field. That rapid evolution is related to the many errors of commission and omission that surely remain, despite my ongoing efforts to eliminate them, for which I apologize in advance.

Francis X. Diebold  
Philadelphia

Tuesday 1<sup>st</sup> August, 2017

Forecasting

# Part I

## Getting Started



# Chapter 1

## Introduction



Figure 1.1: This is Not What This Book is About

### 1.1 Welcome

**Forecasting** is important — forecasts are constantly made in business, finance, economics, government, and many other fields, and they guide many important decisions. As with anything else, there are good and bad ways to forecast. This book is about the good ways: modern, rigorous, replicable, largely-quantitative statistical/econometric methods – their strengths *and* their limitations. That’s why I dislike the above picture of the crystal ball – it bows to the common misconception among the uninitiated that forecasting

is some sort of dubious mystical activity, like fortune telling or astrology. But how could a forecasting book *not* begin with a picture like that? So I decided to lighten up, if only for a moment.

## 1.2 Who Forecasts, and Why?

Forecasts are made and used in numerous fields. To develop a feel for the tremendous diversity of forecasting applications, let's sketch some of the areas where forecasts feature prominently, and the corresponding diversity of decisions that they support.

One key field is economics, broadly defined. Governments, businesses, policy organizations, central banks, financial services firms, and economic consulting firms around the world routinely forecast major economic variables, such as gross domestic product (GDP), unemployment, consumption, investment, the price level, and interest rates. Governments use such forecasts to guide monetary and fiscal policy, and private firms use them for strategic planning, because economy-wide economic fluctuations typically have industry-level and firm-level effects. In addition to forecasting “standard” variables such as GDP, economists sometimes make more exotic forecasts, such as the stage of the business cycle that we'll be in six months from now (expansion or contraction), the state of future stock market activity (bull or bear), or the state of future foreign exchange market activity (appreciation or depreciation). Again, such forecasts are of obvious use to both governments and firms – if they're accurate!

Another key area is business and all its subfields. These include management strategy of all types including operations management and control (hiring, production, inventory, investment, ...), marketing (pricing distributing, advertising, ...), and accounting (budgeting using revenue and expenditure forecasts), etc. Sales forecasting is a good example. Firms routinely forecast

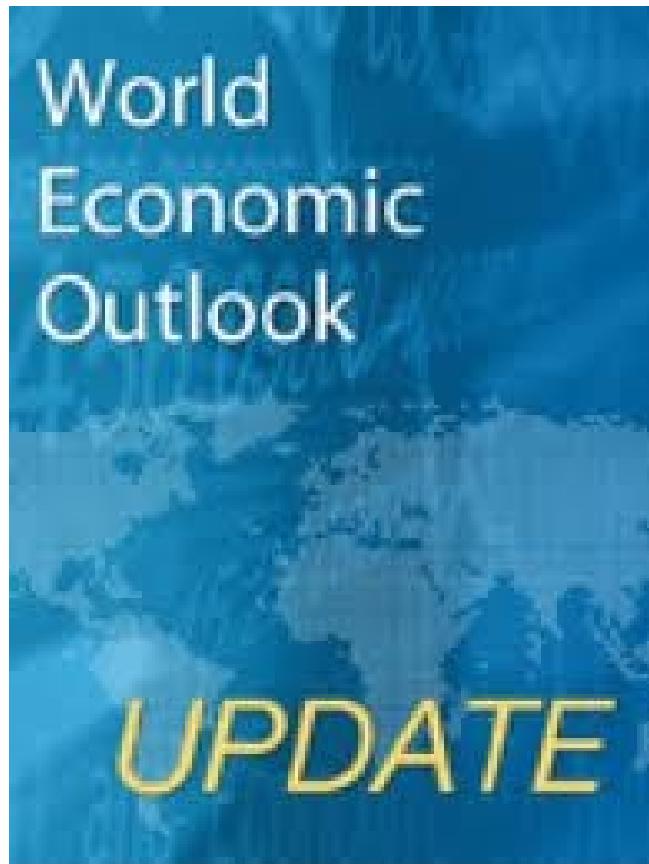


Figure 1.2: Economics: World Economic Outlook

sales to help guide management decisions in inventory management, sales force management, and production planning, as well as strategic planning regarding product lines, new market entry, and so on.

More generally, firms use forecasts to decide what to produce (What product or mix of products should be produced?), when to produce (Should we build up inventories now in anticipation of high future demand? How many shifts should be run?), how much to produce and how much capacity to build (What are the trends in market size and market share? Are there cyclical or seasonal effects? How quickly and with what pattern will a newly-built plant or a newly-installed technology depreciate?), and where to produce (Should we have one plant or many? If many, where should we locate them?). Firms

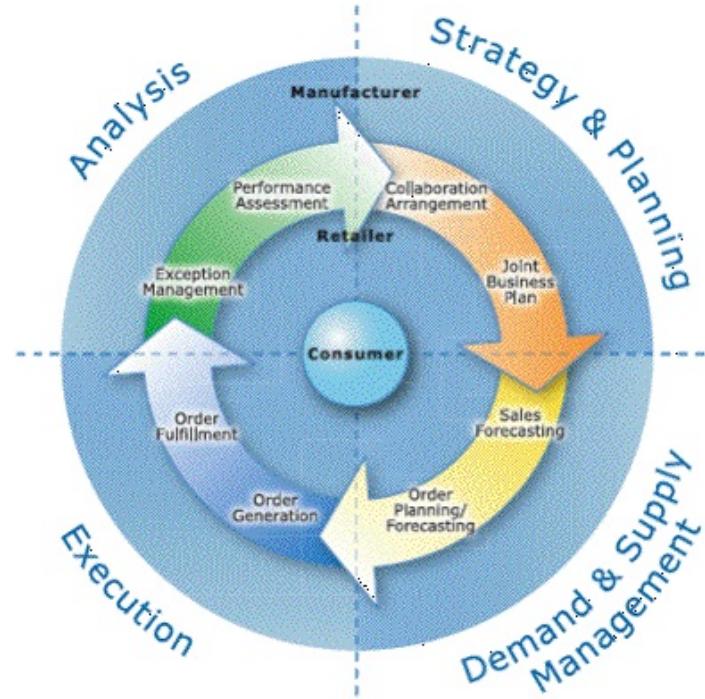


Figure 1.3: Business: Sales Forecasting

also use forecasts of future prices and availability of inputs to guide production decisions.

Forecasting is also crucial in financial services, including asset management, asset pricing, mergers and acquisitions, investment banking, and insurance. Portfolio managers, for example, have keen interest in forecasting asset returns (stock returns, interest rates, exchange rates, and commodity prices) and such forecasts are made routinely. There is endless debate about the success of forecasts of asset returns. On the one hand, asset returns should be very hard to forecast; if they were easy to forecast, you could make a fortune easily, and any such “get rich quick” opportunities should already have been exploited. On the other hand, those who exploited them along the way may well have gotten rich! Thus, we expect that simple, widely-available methods for forecasting should have little success in financial markets, but



Figure 1.4: Finance: A Trading Room

there may well be profits to be made from using new and sophisticated techniques to uncover and exploit previously-unnoticed patterns in financial data (at least for a short time, until other market participants catch on or your own trading moves the market).

Forecasting is similarly central to financial risk management. The forecasting of asset return volatility is related to the forecasting of asset returns. In recent decades, practical methods for volatility forecasting have been developed and widely applied. Volatility forecasts are crucial for evaluating and insuring risks associated with asset portfolios. Volatility forecasts are also crucial for firms and investors who need to price assets such as options and other derivatives.

Finally, forecasting is central to a variety of consulting firms, many of which support the business functions already mentioned. Litigation support is a particularly active area. Forecasting is central to damage assessment (e.g., lost earnings), “but for” analyses and event studies, etc.

The above examples are just the tip of the iceberg. To take another example, demographers routinely forecast the populations of countries and



Figure 1.5: Consulting: Litigation Support

regions all over the world, often in disaggregated form, such as by age, sex, and race. Population forecasts are crucial for planning government expenditure on health care, infrastructure, social insurance, anti-poverty programs, and so forth. Many private-sector decisions, such as strategic product line decisions by businesses, are guided by demographic forecasts of particular targeted population subgroups. Population in turn depends on births, deaths, immigration and emigration, which also are forecasted routinely.

To take just one more example, many events corresponding to crises of various sorts are frequently forecasted. Such forecasts are routinely issued as probabilities. For example, in both consumer and commercial lending, banks generate default probability forecasts and refuse loans if the probability is deemed too high. Similarly, international investors of various sorts are concerned with probabilities of default, currency devaluations, military coups, etc., and use forecasts of such events to inform their portfolio allocation decisions.

The variety of forecasting tasks that we've just sketched was selected to help you begin to get a feel for the depth and breadth of the field. Surely you can think of many more situations in which forecasts are made and used to guide decisions. With so many different forecasting applications, you might fear that a huge variety of forecasting techniques exists, and that you'll have to master all of them. Fortunately, that's not the case. Instead, a rela-

tively small number of tools form the common core of almost all forecasting methods. Needless to say, the details differ if one is forecasting Intel’s stock price one day and the population of Scotland the next, but the underlying forecasting principles are identical. We will focus on those underlying core principles.

## 1.3 Useful Materials

As you begin your study of forecasting, it’s important that you begin to develop an awareness of a variety of useful and well-known forecasting textbooks, forecasting journals where original research is published, forecasting software, data sources, professional organizations, etc.

### 1.3.1 Books

A number of good books exist that complement this one; some are broader, some are more advanced, and some are more specialized. Here we’ll discuss a few that are more broad or more advanced. We’ll mention more specialized books in subsequent chapters when appropriate.

Wonnacott and Wonnacott (1990) remains a time-honored classic statistics text, which you may wish to consult to refresh your memory on statistical distributions, estimation and hypothesis testing. Anderson et al. (2008) is a well-written and more-recent statistics text, containing a very accessible discussion of linear regression, which we use extensively throughout this book. Pindyck and Rubinfeld (1997) remains one of the all-time great introductory econometrics texts, and it has unusually-strong treatment of time-series and forecasting. It’s a useful refresher for basic statistical topics, as well as a good introduction to more advanced **econometric models**.

As a student of forecasting, you’ll also want to familiarize yourself with the broader time series analysis literature. Most forecasting methods are

concerned with forecasting time series – data recorded over time. The modeling and forecasting of time series are so important that an entire field called “**time series analysis**” has arisen. Forecasting is intimately related to time series analysis, because quantitative time series forecasting techniques are based on quantitative time series models. Thus, forecasting requires knowledge of time series modeling techniques, and we therefore devote a substantial portion of this book to time series modeling. Chatfield (2006) is a good introductory time series book, which you’ll find useful as a background reference. More advanced books, which you may want to consult later, include Granger and Newbold (1986), a classic packed with insight and explicitly oriented toward those areas of time series analysis relevant for forecasting. Finally, Hamilton (1994) and Shumway and Stoffer (2011) are fine advanced texts suitable for Ph.D.-level study.

### 1.3.2 Online Information and Data

A variety of information of interest to forecasters is available on the web. The best way to learn about what’s out there is to spend a few hours searching the web for whatever interests you. Here we mention just a few key “must-know” sites. [Resources for Economists](#), maintained by the American Economic Association, is a fine portal to almost anything of interest to economists. It contains hundreds of links to data sources, journals, professional organizations, and so on. [FRED \(Federal Reserve Economic Data\)](#) at the Federal Reserve Bank of St. Louis is a tremendously convenient source for economic data, as is [Quandl](#). [Forecasting Principles](#) has a wealth of data well beyond economics, as well as extensive additional information of interest to forecasters. The [National Bureau of Economic Research](#) site has data on U.S. business cycles, and the [Real-Time Data Research Center](#) at the Federal Reserve Bank of Philadelphia has real-time vintage macroeconomic data.

**RFE: Resources for Economists on the Internet**

**ISSN 1081-4248**  
**Vol. 14, No. 8**  
**October 25, 2012**

Editor: [Bill Goffe](#)  
 Dept. of Economics, Penn State University  
 Editorial Assistant: Adrienne Mullins

- [Introduction](#)
- [Data](#)
- [Dictionaries, Glossaries, & Encyclopedias](#)
- [Economists, Departments, & Universities](#)
- [Forecasting & Consulting](#)
- [Jobs, Grants, Grad School, & Advice](#)
- [Mailing Lists & Forums](#)
- [Meetings & Conferences](#)
- [News Media](#)
- [Organizations & Associations](#)
- [Other Internet Guides](#)
- [Scholarly Communication](#)
- [Software](#)
- [Teaching Resources](#)

Figure 1.6: Resources for Economists Web Page

### 1.3.3 Software (and a Tiny bit of Hardware)

Just as some journals specialize exclusively in forecasting, so too do some software packages. But just as the most important forecasting articles often appear in journals much broader than the specialized forecasting journals, so too are forecasting tools scattered throughout econometric / statistical software packages with capabilities much broader than forecasting alone. One of the best such packages is **Eviews**, a modern object-oriented environment with extensive time series, modeling and forecasting capabilities. It implements almost all of the methods described in this book, and many more. Eviews reflects a balance of generality and specialization that makes it ideal for the sorts of tasks that will concern us, and most of the examples in this book are

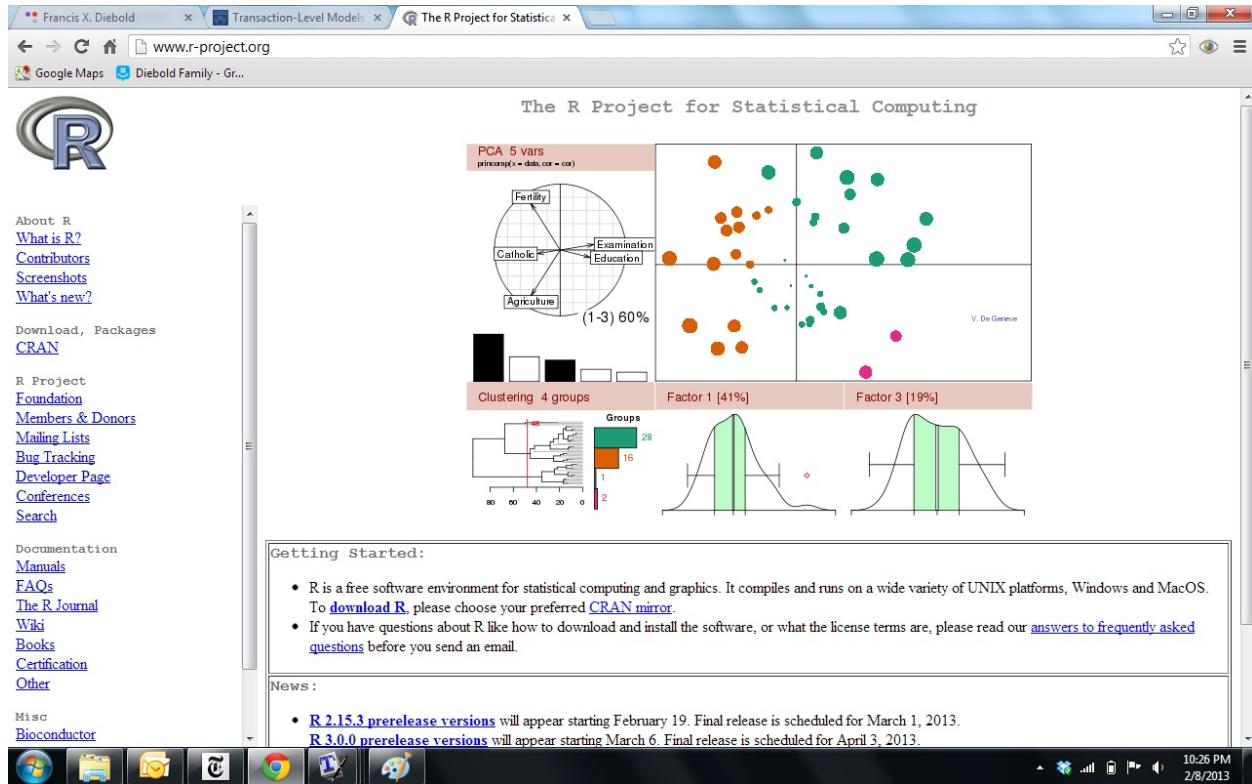


Figure 1.7: The R Homepage

done using it. If you feel more comfortable with another package, however, that's fine – none of our discussion is wed to Eviews in any way, and most of our techniques can be implemented in a variety of packages.

Eviews is an example of a very high-level modeling environment. If you go on to more advanced modeling and forecasting, you'll probably want also to have available slightly lower-level ("mid-level") environments in which you can quickly program, evaluate and apply new tools and techniques. **R** is one very powerful and popular such environment, with special strengths in modern statistical methods and graphical data analysis. It is available for free as part of a major open-source project. In this author's humble opinion, R is the key mid-level environment for the foreseeable future.<sup>1</sup>

If you need real speed, such as for large simulations, you will likely need

<sup>1</sup>Python and Julia are also powerful mid-level environments.

a low-level environment like Fortran or C++. And in the limit (and on the hardware side), if you need blazing-fast parallel computing for massive simulations etc., graphics cards (graphical processing units, or GPU's) provide stunning gains, as documented for example in Aldrich et al. (2011).

### 1.3.4 Journals and Professional Organizations

Forecasting cuts across many literatures, including statistics, econometrics, machine learning, and many others.

A number of journals cater to the forecasting community. *International Journal of Forecasting*, for example, is a leading academic forecasting journal, which contains a mixture of newly-proposed methods, evaluation of existing methods, practical applications, and book and software reviews. It is an official journal of the International Institute of Forecasters, which also publishes *Foresight* (a super-applied journal for industry professionals) and *The Oracle* (an online newsletter), and sponsors the *Forecasting Principles* site. Other organizations with a strong focus on forecasting methods include the Econometric Society and the Society for Financial Econometrics (SoFiE).

Although there are a number of journals devoted to forecasting, its interdisciplinary nature results in a rather ironic outcome: A substantial fraction of the best forecasting research is published not in the forecasting journals, but rather in the broader applied econometrics and statistics journals, such as *Journal of Econometrics*, *Journal of Business and Economic Statistics*, and *Journal of Applied Econometrics*, among many others.

## 1.4 Final Thoughts

Forecasts guide decisions, and good forecasts help to produce good decisions. In the remainder of this book, we'll motivate, describe, and compare modern forecasting methods. You'll learn how to build and evaluate forecasts and

forecasting models, and you'll be able to use them to improve your decisions.

Forecasting is inextricably linked to the building of **statistical models**. Before we can forecast a variable of interest, we typically build a model for it and estimate the model's parameters using observed historical data. Typically, the estimated model summarizes dynamic patterns in the data; that is, the estimated model provides a statistical characterization of the links between the present and the past. More formally, an estimated **forecasting model** provides a characterization of what we expect in the present, conditional upon the past, from which we infer what to expect in the future, conditional upon the present and past. Quite simply, we use the estimated forecasting model to extrapolate the observed historical data.

In this book we focus on core modeling and forecasting methods that are very widely applicable. We begin by introducing several fundamental issues relevant to any forecasting exercise, and then we treat the construction, use, and evaluation of modern forecasting models. We give special attention to basic methods for forecasting trend, seasonality and cycles, as well as methods for evaluating and combining forecasts. Most chapters contain a detailed application; examples include forecasting retail sales, housing starts, employment, liquor sales, exchange rates, shipping volume, and stock market volatility.

## 1.5 Tips on How to use this book

As you navigate through the book, keep the following in mind.

- Hyperlinks to internal items (table of contents, index, footnotes, etc.) appear in red.
- Hyperlinks to bibliographical references appear in green.
- Hyperlinks to external items (web pages, video, etc.) appear in cyan.<sup>2</sup>

---

<sup>2</sup>Obviously web links sometimes go dead. I make every effort to keep them updated in the latest edition

- Hyperlinks to external files appear in blue.
- Many graphics are clickable to reach related material, as are, for example, all pictures in this chapter.
- Key concepts appear in bold. They also appear in the (hyperlinked) index and so can be referenced instantly.
- Additional course-related materials (slides, code, data) appear on the book’s website at <http://www.ssc.upenn.edu/~fdiebold/Textbooks.html>.
- Datasets appear in Appendix C, from which they may be copied and pasted directly.
- The examples that appear throughout should not be taken as definitive or complete treatments – there is no such thing! A good idea is to think of the implicit “Problem 0” in each chapter’s Exercises, Problems and Complements (EPC) section as “Critique the modeling and forecasting in this chapter’s empirical example, obtain the relevant data, and produce a superior modeling and forecasting analysis.”
- All data used in examples are fictitious. Sometimes they are based on real data for various real countries, firms, etc., and sometimes they are artificially constructed. Ultimately, however, any resemblance to particular countries, firms, etc. should be viewed as coincidental and irrelevant.
- The end-of-chapter EPC’s are of central importance and should be studied carefully. Exercises are generally straightforward checks of your understanding. Problems, in contrast, are generally significantly more in-

---

(but no guarantees of course!). If you’re encountering an unusual number of dead links, you’re probably using an outdated edition.

volved, whether analytically or computationally. Complements generally introduce important auxiliary material not covered in the main text.

## 1.6 Exercises, Problems and Complements

### 1. The basic forecasting framework.

True or false:

- a. The underlying principles of time-series forecasting differ radically depending on the time series being forecast.
- b. Ongoing improvements in forecasting methods will eventually enable perfect prediction.
- c. There is no way to learn from a forecast's historical performance whether and how it could be improved.

### 2. Data and forecast timing conventions.

Suppose that, in a particular monthly data set, time  $t = 10$  corresponds to September 1960.

- a. Name the month and year of each of the following times:  $t + 5$ ,  $t + 10$ ,  $t + 12$ ,  $t + 60$ .
- b. Suppose that a series of interest follows the simple process  $y_t = y_{t-1} + 1$ , for  $t = 1, 2, 3, \dots$ , meaning that each successive month's value is one higher than the previous month's. Suppose that  $y_0 = 0$ , and suppose that at present  $t = 10$ . Calculate the forecasts  $y_{t+5,t}$ ,  $y_{t+10,t}$ ,  $y_{t+12,t}$ ,  $y_{t+60,t}$ , where, for example,  $y_{t+5,t}$  denotes a forecast made at time  $t$  for future time  $t + 5$ , assuming that  $t = 10$  at present.

### 3. Degrees of forecastability.

Which of the following can be forecast perfectly? Which can not be forecast at all? Which are somewhere in between? Explain your answers, and be careful!

- a. The direction of change tomorrow in a country's stock market;

- b. The eventual lifetime sales of a newly-introduced automobile model;
  - c. The outcome of a coin flip;
  - d. The date of the next full moon;
  - e. The outcome of a (fair) lottery.
4. Forecasting in daily life.

We all forecast, all the time, implicitly if not explicitly.

- a. Sketch in detail three forecasts that you make routinely, and probably informally, in your daily life. What makes you believe that the things your forecast are in fact forecastable? What does that even *mean*? What factors might introduce error into your forecasts?
  - b. What decisions are aided by your three forecasts? How might the degree of predictability of the forecast object affect your decisions?
  - c. For each of your forecasts, what is the value to you of a “good” as opposed to a “bad” forecast?
  - d. How might you measure the “goodness” of your three forecasts?
5. Forecasting in business, finance, economics, and government.

What sorts of forecasts would be useful in the following decision-making situations? Why? What sorts of data might you need to produce such forecasts?

- a. Shop-All-The-Time Network (SATTN) needs to schedule operators to receive incoming calls. The volume of calls varies depending on the time of day, the quality of the TV advertisement, and the price of the good being sold. SATTN must schedule staff to minimize the loss of sales (too few operators leads to long hold times, and people hang up if put on hold) while also considering the loss associated with hiring excess employees.

- b. You're a U.S. investor holding a portfolio of Japanese, British, French and German stocks and government bonds. You're considering broadening your portfolio to include corporate stocks of Tambia, a developing economy with a risky emerging stock market. You're only willing to do so if the Tambian stocks produce higher portfolio returns sufficient to compensate you for the higher risk. There are rumors of an impending military coup, in which case your Tambian stocks would likely become worthless. There is also a chance of a major Tambian currency depreciation, in which case the dollar value of your Tambian stock returns would be greatly reduced.
- c. You are an executive with Grainworld, a huge corporate farming conglomerate with grain sales both domestically and abroad. You have no control over the price of your grain, which is determined in the competitive market, but you must decide what to plant and how much, over the next two years. You are paid in foreign currency for all grain sold abroad, which you subsequently convert to dollars. Until now the government has bought all unsold grain to keep the price you receive stable, but the agricultural lobby is weakening, and you are concerned that the government subsidy may be reduced or eliminated in the next decade. Meanwhile, the price of fertilizer has risen because the government has restricted production of ammonium nitrate, a key ingredient in both fertilizer and terrorist bombs.
- d. You run BUCO, a British utility supplying electricity to the London metropolitan area. You need to decide how much capacity to have on line, and two conflicting goals must be resolved in order to make an appropriate decision. You obviously want to have enough capacity to meet average demand, but that's not enough, because demand is uneven throughout the year. In particular, demand skyrockets during summer heat waves – which occur randomly – as more and more peo-

ple run their air conditioners constantly. If you don't have sufficient capacity to meet peak demand, you get bad press. On the other hand, if you have a large amount of excess capacity over most of the year, you also get bad press.

## 6. Finding and using data on the web.

Search the web for information on U.S. retail sales, U.K. stock prices, German GDP, and Japanese federal government expenditures. Using graphical methods, compare and contrast the movements of each series.

## 7. Software differences and bugs: caveat emptor.

Be warned: no software is perfect. In fact, all software is highly imperfect! The results obtained when modeling or forecasting in different software environments may differ – sometimes a little and sometimes a lot – for a variety of reasons. The details of implementation may differ across packages, for example, and small differences in details can sometimes produce large differences in results. Hence, it is important that you understand *precisely* what your software is doing (insofar as possible, as some software documentation is more complete than others). And of course, quite apart from correctly-implemented differences in details, deficient implementations can and do occur: there is no such thing as bug-free software.

## 8. Forecasting vs. prediction.

We will use the terms *prediction* and *forecasting* interchangeably, using either term in all environments (time-series environments), cross-section environments, etc.)

# Chapter 2

## Universal Considerations

In Chapter 1 we sketched a variety of areas where forecasts are used routinely. Here we begin by highlighting, in no particular order, a number of considerations relevant for *any* forecasting task. We introduce the those considerations as *questions*.

1. **(Forecast Object)** What is the object that we want to forecast? Is it a time series, such as sales of a firm recorded over time, or an event, such as devaluation of a currency, or something else? Appropriate forecasting strategies depend on the nature of the object being forecast.
2. **(Information Set)** On what information will the forecast be based? In a time series environment, for example, are we forecasting one series, several, or thousands? And what is the quantity and quality of the data? Appropriate forecasting strategies depend on the information set, broadly interpreted to not only quantitative data but also expert opinion, judgment, and accumulated wisdom.
3. **(Model Uncertainty and Improvement)** Does our forecasting model match the true DGP? Of course not. One must never, ever, be so foolish as to be lulled into such a naive belief. All models are false: they are *intentional* abstractions of a much more complex reality. A model might be useful for certain purposes and poor for others. Models that once

worked well may stop working well. One must continually diagnose and assess both empirical performance and consistency with theory. The key is to work continuously toward model improvement.

4. (**Forecast Horizon**) What is the forecast horizon of interest, and what determines it? Are we interested, for example, in forecasting one month ahead, one year ahead, or ten years ahead (called ***h*-step-ahead forecasts**, in this case for  $h = 1$ ,  $h = 12$  and  $h = 120$  months)? Appropriate forecasting strategies likely vary with the horizon.

#### 5. (**Structural Change**)

Are the approximations to reality that we use for forecasting (i.e., our models) stable over time? Generally not. Things can change for a variety of reasons, gradually or abruptly, with obviously important implications for forecasting. Hence we need methods of detecting and adapting to structural change.

6. (**Forecast Statement**) How will our forecasts be stated? If, for example, the object to be forecast is a time series, are we interested in a single “best guess” forecast, a “reasonable range” of possible future values that reflects the underlying uncertainty associated with the forecasting problem, or a full probability distribution of possible future values? What are the associated costs and benefits?

#### 7. (**Forecast Presentation**)

How best to *present* forecasts? Except in the simplest cases, like a single *h*-step-ahead point forecast, graphical methods are valuable, not only for forecast presentation but also for forecast construction and evaluation.

8. (**Decision Environment and Loss Function**) What is the **decision environment** in which the forecast will be used? In particular, what decision will the forecast guide? How do we quantify what we mean

by a “good” forecast, and in particular, the cost or loss associated with forecast errors of various signs and sizes?

9. (**Model Complexity and the Parsimony Principle**) What sorts of models, in terms of complexity, tend to do best for forecasting in business, finance, economics, and government? The phenomena that we model and forecast are often tremendously complex, but it does not necessarily follow that our forecasting models should be complex. Bigger forecasting models are not necessarily better, and indeed, all else equal, smaller models are generally preferable (the “parsimony principle”).
10. (**Unobserved Components**) In the leading time case of time series, have we successfully modeled trend? Seasonality? Cycles? Some series have all such components, and some not. They are driven by very different factors, and each should be given serious attention.

## 2.1 The Forecast Object

There are many objects that we might want to forecast. In business and economics, the forecast object is typically one of three types: **event outcome**, **event timing**, or **time series**.

Event outcome forecasts are relevant to situations in which an event is certain to take place at a given time but the outcome is uncertain. For example, many people are interested in whether the current chairman of the Board of Governors of the U.S. Federal Reserve System will eventually be reappointed. The “event” is the reappointment decision; the decision will occur at the end of the term. The outcome of this decision is confirmation or denial of the reappointment.

Event timing forecasts are relevant when an event is certain to take place and the outcome is known, but the timing is uncertain. A classic example of an event timing forecast concerns business cycle turning points. There are

two types of turning points: peaks and troughs. A peak occurs when the economy moves from expansion into recession, and a trough occurs when the economy moves from recession into expansion. If, for example, the economy is currently in an expansion, then there is no doubt that the next turning point will be a peak, but there is substantial uncertainty as to its *timing*. Will the peak occur this quarter, this year, or ten years from now?

Time series forecasts are relevant when the future value of a time series is of interest and must be projected. As we'll see, there are many ways to make such forecasts, but the basic forecasting setup doesn't change much. Based upon the history of the time series (and possibly a variety of other types of information as well, such as the histories of related time series, or subjective considerations), we want to project future values of the series. For example, we may have data on the number of Apple computers sold in Germany in each of the last 60 months, and we may want to use that data to forecast the number of Apple computers to be sold in Germany in each month of the next year.

There are at least two reasons why time series forecasts are by far the most frequently encountered in practice. First, most business, economic and financial data are time series; thus, the general scenario of projecting the future of a series for which we have historical data arises constantly. Second, the technology for making and evaluating time-series forecasts is well-developed and the typical time series forecasting scenario is precise, so time series forecasts can be made and evaluated routinely. In contrast, the situations associated with event outcome and event timing forecasts arise less frequently and are often less amenable to quantitative treatment.

## 2.2 The Information Set

The quality of our forecasts is limited by the quality and quantity of information available when forecasts are made. Any forecast we produce is conditional upon the information used to produce it, whether explicitly or implicitly.

### 2.2.1 Univariate vs. Multivariate

The idea of an information set is fundamental to constructing good forecasts. In forecasting a series,  $y$ , using historical data from time 1 to time  $T$ , sometimes we use the **univariate information set**, which is the set of historical values of  $y$  up to and including the present,

$$\Omega_T = \{y_T, y_{T-1}, \dots, y_1\}.$$

In a univariate environment, then, a single variable is modeled and forecast solely on the basis of its own past. Univariate approaches to forecasting may seem simplistic, and in some situations they are, but they are tremendously important and worth studying for at least two reasons. First, although they are simple, they are not necessarily simplistic, and a large amount of accumulated experience suggests that they often perform admirably. Second, it's necessary to understand univariate forecasting models before tackling more complicated multivariate models.

Alternatively, sometimes we use the **multivariate information set**

$$\Omega_T = \{y_T, x_T, y_{T-1}, x_{T-1}, \dots, y_1, x_1\},$$

where the  $x$ 's are a set of additional variables potentially related to  $y$ . In a multivariate environment, a variable (or each member of a set of variables) is modeled on the basis of its own past, as well as the past of other variables, thereby accounting for and exploiting cross-variable interactions. Multivari-

ate models have the *potential* to produce forecast improvements relative to univariate models, because they exploit more information to produce forecasts.

### 2.2.2 Expert Opinion and Judgment

Regardless of whether the information set is univariate or multivariate, it's always important to think hard about what information is available, what additional information could be collected or made available, the form of the information (e.g., quantitative or qualitative), and so on. A holistic view of an information involves far more than just the past history of one or a few quantitative variables; instead, it involves theoretical perspectives, expert judgment, contextual knowledge, and so on.

So you should take a broad view of what's meant by a "model." Try to incorporate views of experts and even non-experts. (Sometimes the alleged experts are not so expert, and the alleged non-experts are quite insightful.) Surveys, Bayesian priors and shrinkage, forecast combination, and prediction markets, which we'll discuss in due course, all attempt to do that.

### 2.2.3 Information Sets in Forecast Evaluation

The idea of an information set is also fundamental for evaluating forecasts: the basic principle of forecast evaluation is that a "good" forecast has corresponding errors that are unforecastable using information available when the forecast is made. When evaluating a forecast, we're sometimes interested in whether the forecast could be improved by using a given set of information more efficiently, and we're sometimes interested in whether the forecast could be improved by using more information. Either way, the ideas of information and information sets play crucial roles in forecast evaluation.

## 2.3 Model Uncertainty and Improvement

One must never, ever, be so foolish as to be lulled into believing that one's model coincides with the true DGP. Indeed all models are false. Does that mean that models and modeling are somehow discredited, or worthless? Not at all, and their use continues to expand. The uninitiated are sometimes suspicious, because their lack of understanding of models and modeling leads them to have unreasonable expectations of models, but there's no other way forward. As George Box (1979) famously and correctly noted, "All models are false, but some are useful."

Related, a model might be useful for certain purposes and poor for others. Models that once worked well may stop working well. One must continually diagnose and assess both empirical performance and consistency with theory. That is, the key is to think continually about how to *improve* models. And always remember: *It takes a model to beat a model.*

## 2.4 The Forecast Horizon

### 2.4.1 $h$ -Step-Ahead Forecasts

The forecast horizon is defined as the number of periods between today and the date of the forecast we make. For example, if we have annual data, and it's now year  $T$ , then a forecast of GDP for year  $T + 2$  has a forecast horizon of 2 steps. The meaning of a step depends on the frequency of observation of the data. For monthly data a step is one month, for quarterly data a step is one quarter (three months), and so forth. In general, we speak of an  **$h$ -step ahead forecast**, where the horizon  $h$  is at the discretion of the user.<sup>1</sup>

The horizon is important for at least two reasons. First, of course, the forecast changes with the forecast horizon. Second, the best forecasting model

---

<sup>1</sup>The choice of  $h$  depends on the decision that the forecast will guide. The nature of the decision environment typically dictates whether "short-term", "medium-term", or "long-term" forecasts are needed.

will often change with the forecasting horizon as well. All of our forecasting models are approximations to the underlying dynamic patterns in the series we forecast; there's no reason why the best approximation for one purpose (e.g., short-term forecasting) should be the same as the best approximation for another purpose (e.g., long-term forecasting).

### 2.4.2 $h$ – Step Ahead Path Forecasts

Let's distinguish between what we've called  $h$ -step-ahead forecasts and what we'll call  **$h$ -step-ahead path forecasts**, sometimes also called  **$h$ -step-ahead extrapolation forecasts**. In  $h$ -step-ahead forecasts, the horizon is always fixed at the same value,  $h$ . For example, every month we might make a 4-month-ahead forecast. Alternatively, in path forecasts, the horizon includes all steps from 1-step-ahead to  $h$ -steps-ahead. There's nothing particularly deep or difficult about the distinction, but it's useful to make it, and we'll use it subsequently.

Suppose, for example, that you observe a series from some initial time 1 to some final time  $T$ , and you plan to forecast the series.<sup>2</sup> We illustrate the difference between  $h$ -step-ahead and  $h$ -step-ahead path forecasts in Figures 2.1 and 2.2. In Figure 2.1 we show a 4-step-ahead point forecast, and in Figure 2.2 we show a 4-step-ahead path point forecast. The path forecast is nothing more than a set consisting of 1-, 2-, 3-, and 4-step-ahead forecasts.

### 2.4.3 Nowcasting and Backcasting

Quite apart from making informative and useful guesses about the future (forecasting), often we're interested in the present ("nowcasting") – which is also subject to lots of uncertainty – or even the past ("backcasting"). Many

---

<sup>2</sup>For a sample of data on a series  $y$ , we'll typically write  $\{y_t\}_{t=1}^T$ . This notation means "we observe the series  $y$  from some beginning time  $t = 1$  to some ending time  $t = T$ ".

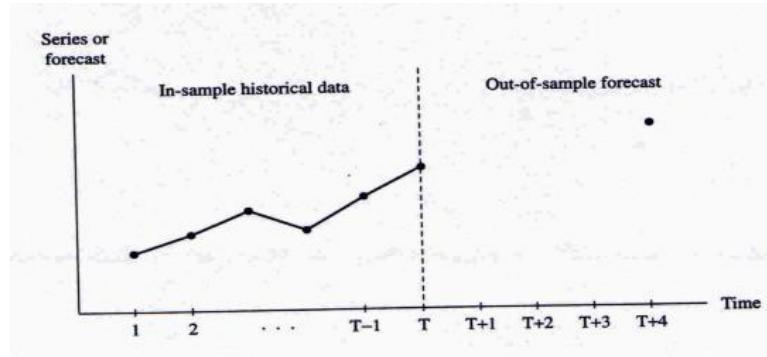


Figure 2.1: 4-Step-Ahead Point Forecast

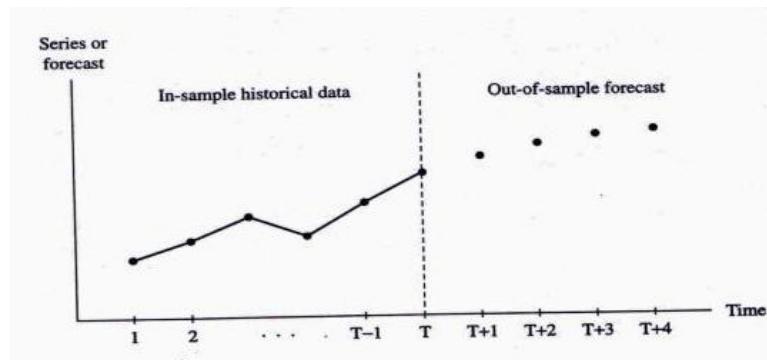


Figure 2.2: 4-Step-Ahead Path Forecast

of our models and methods will be relevant there as well.<sup>3</sup>

## 2.5 Structural Change

In time series, we rely on the future being like the present/past in terms of dynamic relationships (e.g., the next twenty years vs. the last twenty years). But that's not always true. Structural change can be gradual or abrupt.

In cross sections, we rely on fitted relationships being relevant for new cases from the original population, and often even for new populations. But again, that's not always true. For example, the effect of class size on test scores may differ for 10-year olds in California vs. 6-year olds in Maine.

Structural change can affect any or all parameters of a model, and the breaks can be large or small.

Structural change is a type of non-linearity; indeed abrupt structural change is often handled with dummy variable models, and gradual structural change is often handled with smoothly-time-varying parameter models.

## 2.6 The Forecast Statement

### 2.6.1 Time Series

When we make a forecast, we must decide if the forecast will be (1) a single number (a “best guess”), (2) a range of numbers, into which the future value can be expected to fall a certain percentage of the time, or (3) an entire probability distribution for the future value. In short, we need to decide upon the forecast type.

More precisely, we must decide if the forecast will be (1) a **point forecast**, (2) an **interval forecast**, or (3) a **density forecast**. A point forecast is a single number. For example, one possible point forecast of the growth rate of

---

<sup>3</sup>For an example of nowcasting see the [ADS Index at FRB Philadelphia](#), and for an example of backcasting see [GDPplus](#), also at FRB Philadelphia.

the total number of web pages over the next year might be +23.3%; likewise, a point forecast of the growth rate of U.S. real GDP over the next year might be +1.3%. Point forecasts are made routinely in numerous applications, and the methods used to construct them vary in difficulty from simple to sophisticated. The defining characteristic of a point forecast is simply that it is a single number.

A good point forecast provides a simple and easily-digested guide to the future of a time series. However, random and unpredictable “shocks” affect all of the series that we forecast. As a result of such shocks, we expect nonzero forecast errors, even from very good forecasts. Thus, we may want to know the degree of confidence we have in a particular point forecast. Stated differently, we may want to know how much uncertainty is associated with a particular point forecast. The uncertainty surrounding point forecasts suggests the usefulness of an interval forecast.

An interval forecast is not a single number; rather, it is a range of values in which we expect the realized value of the series to fall with some (pre-specified) probability.<sup>4</sup> Continuing with our examples, a 90% interval forecast for the growth rate of web pages might be the interval [11.3%, 35.3%] ( $23.3\% \pm 12\%$ ). That is, the forecast states that with probability 90% the future growth rate of web pages will be in the interval [11.3%, 35.3%]. Similarly, a 90% interval forecast for the growth rate of U.S. real GDP might be [-2.3%, 4.3%] ( $1.3\% \pm 3\%$ ); that is, the forecast states that with probability 90% the future growth rate of U.S. real GDP will be in the interval [-2.3%, 4.3%].

A number of remarks are in order regarding interval forecasts. First, the length (size) of the intervals conveys information regarding forecast uncertainty. The GDP growth rate interval is much shorter than the web page

---

<sup>4</sup>An interval forecast is very similar to the more general idea of a *confidence interval* that you studied in statistics. An interval forecast is simply a confidence interval for the true (but unknown) future value of a series, computed using a sample of historical data. We'll say that  $[a, b]$  is a  $100(1 - \alpha)\%$  interval forecast if the probability of the future value being less than  $a$  is  $\alpha/2$  and the probability of the future value being greater than  $b$  is also  $\alpha/2$ .

growth rate interval; this reflects the fact that there is less uncertainty associated with the real GDP growth rate forecast than the web page growth rate forecast. Second, interval forecasts convey more information than point forecasts: given an interval forecast, you can construct a point forecast by using the midpoint of the interval.<sup>5</sup> Conversely, given only a point forecast, there is no way to infer an interval forecast.

Finally, we consider density forecasts. A density forecast gives the entire density (or probability distribution) of the future value of the series of interest. For example, the density forecast of future web page growth might be normally distributed with a mean of 23.3% and a standard deviation of 7.32%. Likewise, the density forecast of future real GDP growth might be normally distributed with a mean of 1.3% and a standard deviation of 1.83%. As with interval forecasts, density forecasts convey more information than point forecasts. Density forecasts also convey more information than interval forecasts, because given a density, interval forecasts at any desired confidence level are readily constructed. For example, if the future value of a series  $x$  is distributed as  $N(\mu, \sigma^2)$ , then a 95% interval forecast of  $x$  is  $\mu \pm 1.96\sigma$ , a 90% interval forecast of  $x$  is  $\mu \pm 1.64\sigma$ , and so forth. Continuing with our example, the relationships between density, interval, and point forecasts are made clear in Figures 2.3 (web page growth) and 2.4 (GDP growth).

To recap, there are three time series forecast types: point, interval, and density. Density forecasts convey more information than interval forecasts, which in turn convey more information than point forecasts. This may seem to suggest that density forecasts are always the preferred forecast, that density forecasts are the most commonly used forecasts in practice, and that we should focus most of our attention in this book on density forecasts.

In fact, the opposite is true. Point forecasts are the most commonly used

---

<sup>5</sup>An interval forecast doesn't *have* to be symmetric around the point forecast, so that we wouldn't necessarily infer a point forecast as the midpoint of the interval forecast, but in many cases such a procedure is appropriate.

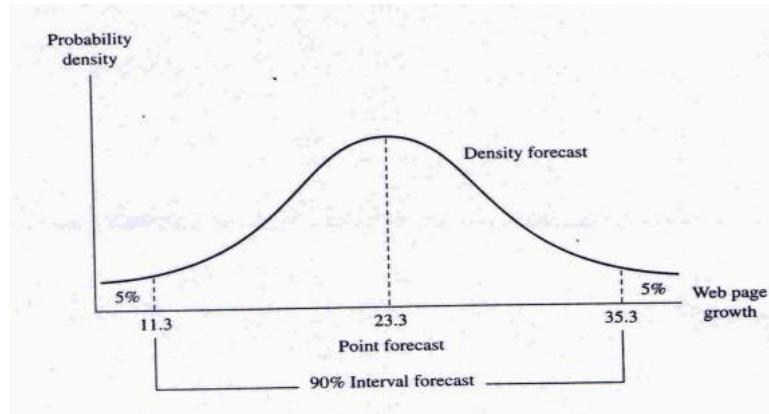


Figure 2.3: Web Page Growth Forecasts

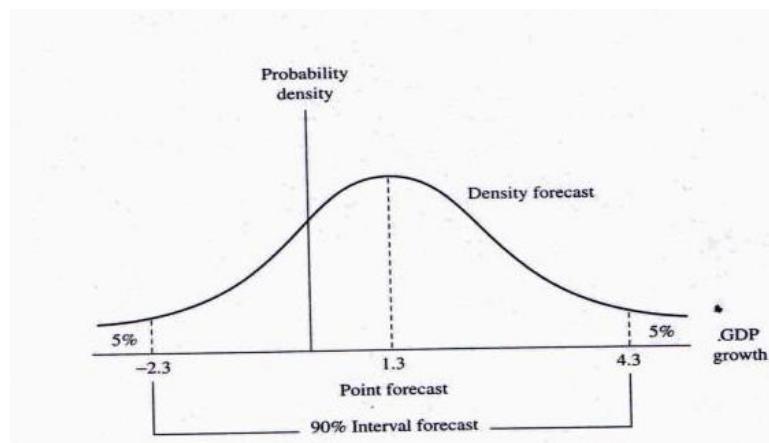


Figure 2.4: GDP Growth Forecasts

forecasts in practice, interval forecasts are a rather distant second, and density forecasts are rarely made. There are at least two reasons. First, the construction of interval and density forecasts requires either (a) additional and possibly incorrect assumptions relative to those required for construction of point forecasts, or (b) advanced and computer-intensive methods involving, for example, extensive simulation. Second, point forecasts are often easier to understand and act upon than interval or density forecasts. That is, the extra information provided by interval and density forecasts is not necessarily an advantage when information processing is costly.

### 2.6.2 Events

Thus far we have focused exclusively on types of time series forecasts, because time series are so prevalent and important in numerous fields. It is worth mentioning another forecast type of particular relevance to event outcome and event timing forecasting, the **probability forecast**. To understand the idea of a probability forecast, consider forecasting which of two politicians, X or Y, will win an election. (This is an event-outcome forecasting situation.) If our calculations tell us that the odds favor X, we might issue the forecast simply as “X will win.” This is roughly analogous to the time series point forecasts discussed earlier, in the sense that we’re not reporting any measure of the uncertainty associated with our forecast. Alternatively, we could report the probabilities associated with each of the possible outcomes; for example, “X will win with probability .6, and Y will win with probability .4.” This is roughly analogous to the time series interval or density forecasts discussed earlier, in the sense that it explicitly quantifies the uncertainty associated with the future event with a probability distribution.

Event outcome and timing forecasts, although not as common as time series forecasts, do nevertheless arise in certain important situations and are often stated as probabilities. For example, when a bank assesses the proba-

bility of default on a new loan or a macroeconomist assesses the probability that a business cycle turning point will occur in the next six months, the banker or macroeconomist will often use a probability forecast.

### 2.6.3 Probability Forecasts as Point and/or Density Forecasts

## 2.7 Forecast Presentation

### 2.7.1 Graphics for Forecasts

### 2.7.2 Graphics for Forecast Evaluation

## 2.8 The Decision Environment and Loss Function

Forecasts are not made in a vacuum. The key to generating good and useful forecasts, which we will stress now and throughout, is recognizing that forecasts are made to guide decisions. The link between forecasts and decisions sounds obvious – and it is – but it’s worth thinking about in some depth. Forecasts are made in a wide variety of situations, but in every case forecasts are of value because they aid in decision making. Quite simply, good forecasts help to produce good decisions. Recognition and awareness of the decision making environment is the key to effective design, use and evaluation of forecasting models.

### 2.8.1 Loss Functions

Let  $y$  denote a series and  $\hat{y}$  its forecast. The corresponding **forecast error**,  $e$ , is the difference between the realization and the previously-made forecast,

$$e = y - \hat{y}.$$

We consider loss functions of the form  $L(e)$ . This means that the loss associated with a forecast depends only on the size of the forecast error. We might

require the loss function  $L(e)$  to satisfy three conditions:

1.  $L(0) = 0$ . That is, no loss is incurred when the forecast error is zero. (A zero forecast error, after all, corresponds to a perfect forecast!)
2.  $L(e)$  is continuous. That is, nearly-identical forecast errors should produce nearly-identical losses.
3.  $L(e)$  is increasing on each side of the origin. That is, the bigger the absolute value of the error, the bigger the loss.

Apart from these three requirements, we impose no restrictions on the form of the loss function.

The **quadratic loss** function is tremendously important in practice. First, it's often an arguably-reasonable approximation to realistic loss structures. Second, it's mathematically convenient: It is usually easy to compute, because quadratic objectives have linear first-order conditions.<sup>6</sup>

Quadratic loss is given by

$$L(e) = e^2,$$

and we graph it as a function of the forecast error in Figure 2.5. Because of the squaring associated with the quadratic loss function, it is symmetric around the origin, and in addition, it increases at an increasing rate on each side of the origin, so that large errors are penalized much more severely than small ones.

Another important symmetric loss function is **absolute loss**, or **absolute error loss**, given by

$$L(e) = |e|.$$

Like quadratic loss, absolute loss is increasing on each side of the origin,

---

<sup>6</sup>In contrast, optimal forecasting under **asymmetric loss** is rather involved, and the tools for doing so are still under development. See, for example, Christoffersen and Diebold, 1997.

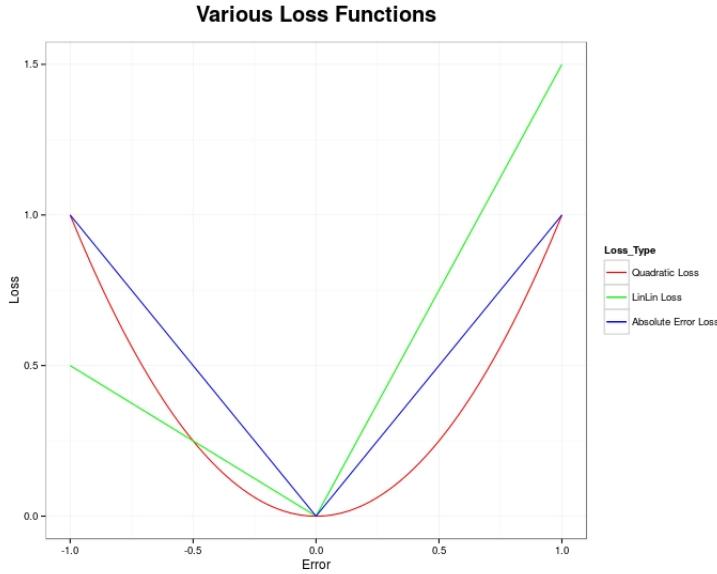


Figure 2.5: Quadratic, Absolute, and Linlin Loss Functions

but loss increases at a constant (linear) rate with the size of the error. We illustrate absolute loss in Figure 2.5.

In certain contexts, symmetric loss functions may not be an adequate distillation of the forecast / decision environment, as would be the case, for example, if negative forecast errors were for some reason generally less costly than positive errors. An important asymmetric loss function is “**linlin loss**” (linear on each side of the origin, with generally different slopes), given by

$$L(e) = \begin{cases} a|e|, & \text{if } e > 0 \\ b|e|, & \text{if } e \leq 0. \end{cases}$$

We show asymmetric linlin loss in Figure 2.5.

### 2.8.2 Optimal Forecasts with Respect to a Loss Function

Much of this book is about how to produce **optimal forecasts**. What precisely do we mean by an optimal forecast? That’s where the loss function

comes in: the optimal forecast is the forecast with smallest conditionally expected loss.

$$\hat{y}(x)^* = \operatorname{argmin}_{\hat{y}(x)} \int \int L(y - \hat{y}(x)) f(y, x) dy dx.$$

Here are some key results:

- Under quadratic loss, the optimal forecast is the conditional mean. That is,

$$\hat{y}(x)^* = E(y|x).$$

Note that  $x$  could be lagged  $y$ .

- Under absolute loss, the optimal forecast is the conditional median,

$$\hat{y}(x)^* = Q_{d \cdot 100\%}(y|x),$$

where  $Q_{d \cdot 100\%}(\cdot)$  denotes the  $d$ -percent conditional quantile function.

- Under lin-lin loss, the optimal forecast is the conditional  $d \cdot 100\%$  quantile, where

$$d = \frac{b}{a+b} = \frac{1}{1+a/b}.$$

That is,

$$\hat{y}(x)^* = Q_{d \cdot 100\%}(y|x).$$

Quite generally under asymmetric  $L(e)$  loss (e.g., linlin), optimal forecasts are biased, whereas the conditional mean forecast is unbiased.<sup>7</sup> Bias is optimal under asymmetric loss because we can gain on average by pushing the forecasts in the direction such that we make relatively few errors of the more costly sign.

---

<sup>7</sup>A forecast is unbiased if its error has zero mean.

### 2.8.3 State-Dependent Loss

In some situations, the  $L(e)$  form of the loss function is too restrictive. Although loss will always be of the form  $L(y, \hat{y})$ , there's no reason why  $y$  and  $\hat{y}$  should *necessarily* enter as  $y - \hat{y}$ . In predicting financial asset returns, for example, interest sometimes focuses on direction of change. A **direction-of-change forecast** takes one of two values – up or down. The loss function associated with a direction of change forecast might be:<sup>8</sup>

$$L(y, \hat{y}) = \begin{cases} 0, & \text{if } \text{sign}(\Delta y) = \text{sign}(\Delta \hat{y}) \\ 1, & \text{if } \text{sign}(\Delta y) \neq \text{sign}(\Delta \hat{y}). \end{cases}$$

With this loss function, if you predict the direction of change correctly, you incur no loss; but if your prediction is wrong, you're penalized.

This is one example of a **state-dependent loss function**, meaning that loss actually depends on the state of the world ( $y$ ), as opposed to just depending on  $e$ . This may sometimes make sense; the cost of a given error may be higher or lower, for example, in different states of the world as indexed by  $y$ .

Under direction-of-change loss, the optimal forecast is the conditional mode. That is,

$$\hat{y}(x)^* = \text{Mode}(y|x).$$

## 2.9 Model Complexity and the Parsimony Principle

It's crucial to tailor forecasting tools to forecasting tasks, and doing so is partly a matter of judgment. Typically the specifics of the situation (e.g., decision environment, forecast object, forecast statement, forecast horizon, information set, etc.) will indicate the desirability of a specific method or

---

<sup>8</sup>The operator “ $\Delta$ ” means “change.” Thus  $\Delta y_t$  is the change in  $y$  from period  $t - 1$  to period  $t$ , or  $y_t - y_{t-1}$ .

modeling strategy. Moreover, as we'll see, formal statistical criteria exist to guide model selection within certain classes of models.

We've stressed that a variety of forecasting applications use a small set of common tools and models. You might guess that those models are tremendously complex, because of the obvious complexity of the real-world phenomena that we seek to forecast. Fortunately, such is not the case. In fact, decades of professional experience suggest just the opposite – simple, parsimonious models tend to be best for out-of-sample forecasting in business, finance, and economics. Hence, the **parsimony principle**: other things the same, simple models are usually preferable to complex models.

There are a number of reasons why smaller, simpler models are often more attractive than larger, more complicated ones. First, by virtue of their parsimony, we can estimate the parameters of simpler models more precisely. Second, because simpler models are more easily interpreted, understood and scrutinized, anomalous behavior is more easily spotted. Third, it's easier to communicate an intuitive feel for the behavior of simple models, which makes them more useful in the decision-making process. Finally, enforcing simplicity lessens the scope for “data mining” – tailoring a model to maximize its fit to historical data. Data mining often results in models that fit historical data beautifully (by construction) but perform miserably in out-of-sample forecasting, because it tailors models in part to the *idiosyncrasies* of historical data, which have no relationship to unrealized future data.

Finally, note that simple models should not be confused with naive models. All of this is well-formalized in the **KISS principle** (appropriately modified for forecasting): “Keep it Sophisticatedly Simple.” We'll attempt to do so throughout.

## 2.10 Unobserved Components

Trend, seasonal, cycle, noise. Deterministic vs. stochastic trend and seasonality.

$$y_t = T_t + S_t + C_t + \varepsilon_t.$$

Or maybe

$$y_t = T_t \times S_t \times C_t \times \varepsilon_t,$$

but of course that's just

$$\ln y_t = \ln T_t + \ln S_t + \ln C_t + \ln \varepsilon_t.$$

## 2.11 Concluding Remarks

This chapter obviously deals with broad issues of general relevance. For the most part, it avoids detailed discussion of specific modeling or forecasting techniques. The rest of the book drills down more deeply.

## 2.12 Exercises, Problems and Complements

1. Properties of loss functions.

State whether the following potential loss functions meet the criteria introduced in the text, and if so, whether they are symmetric or asymmetric:

a.  $L(e) = e^2 + e$

b.  $L(e) = e^4 + 2e^2$

c.  $L(e) = 3e^2 + 1$

d.  $L(e) = \begin{cases} \sqrt{e} & \text{if } e > 0 \\ . & \\ |e| & \text{if } e \leq 0. \end{cases}$

2. Relationships among point, interval and density forecasts.

For each of the following density forecasts, how might you infer “good” point and ninety percent interval forecasts? Conversely, if you started with your point and interval forecasts, could you infer “good” density forecasts?

Be sure to defend your definition of “good.”

a. Future  $y$  is distributed as  $N(10, 2)$ .

b.  $P(y) = \begin{cases} \frac{y-5}{25} & \text{if } 5 < y < 10 \\ -\frac{y-15}{25} & \text{if } 10 < y < 15 \\ 0 & \text{otherwise.} \end{cases}$

3. Forecasting at short through long horizons.

Consider the claim, “The distant future is harder to forecast than the near future.” Is it sometimes true? Usually true? Always true? Why or why not? Discuss in detail. Be sure to define “harder.”

4. “Real” forecasts vs. “goal” or “advocacy” forecasts.

Many things that seem like forecasts are not at all real forecasts. Every politician forecasts that she will win the election. Should you take such forecasts seriously? Every lawyer forecasts that his client will win. Should you take such forecasts seriously? Simultaneously, hidden away from the public, serious, scientifically disinterested forecasts are routinely made and used successfully in numerous endeavors. The problem is that the public routinely sees the former (e.g., from television pundits) and rarely sees the latter.

5. Univariate and multivariate information sets.

- a. Which of the following modeling situations involve univariate information sets? Multivariate?
  - i. Using a stock’s price history to forecast its price over the next week;
  - ii. Using a stock’s price history and volatility history to forecast its price over the next week;
  - iii. Using a stock’s price history and volatility history to forecast its price and volatility over the next week.
- b. Keeping in mind the distinction between univariate and multivariate information sets, consider a wine merchant seeking to forecast the price per case at which a fine vintage of Chateau Latour, one of the greatest Bordeaux wines, will sell when it is thirty years old, at which time it will be fully mature.

- i. What sorts of univariate forecasting approaches can you imagine that might be relevant?
  - ii. What sorts of multivariate forecasting approaches can you imagine that might be relevant? What other variables might be used to predict the Latour price?
  - iii. What are the comparative costs and benefits of the univariate and multivariate approaches to forecasting the Latour price?
  - iv. Would you adopt a univariate or multivariate approach to forecasting the Latour price? Why?
6. Assessing forecasting situations.

For each of the following scenarios, discuss the decision environment, the nature of the object to be forecast, the forecast type, the forecast horizon, the loss function, the information set, and what sorts of simple or complex forecasting approaches you might entertain.

- a. You work for Airborne Analytics, a highly specialized mutual fund investing exclusively in airline stocks. The stocks held by the fund are chosen based on your recommendations. You learn that a newly rich oil-producing country has requested bids on a huge contract to deliver thirty state-of-the-art fighter planes, but that only two companies submitted bids. The stock of the successful bidder is likely to rise.
- b. You work for the Office of Management and Budget in Washington DC and must forecast tax revenues for the upcoming fiscal year. You work for a president who wants to maintain funding for his pilot social programs, and high revenue forecasts ensure that the programs keep their funding. However, if the forecast is too high, and the president runs a large deficit at the end of the year, he will be seen as fiscally irresponsible, which will lessen his probability of reelection.

Furthermore, your forecast will be scrutinized by the more conservative members of Congress; if they find fault with your procedures, they might have fiscal grounds to undermine the President's planned budget.

- c. You work for D&D, a major Los Angeles advertising firm, and you must create an ad for a client's product. The ad must be targeted toward teenagers, because they constitute the primary market for the product. You must (somehow) find out what kids currently think is "cool," incorporate that information into your ad, and make your client's product attractive to the new generation. If your hunch is right, your firm basks in glory, and you can expect multiple future clients from this one advertisement. If you miss, however, and the kids don't respond to the ad, then your client's sales fall and the client may reduce or even close its account with you.

## 7. Box vs. Wiener on Models and Modeling.

We earlier mentioned George Box's memorable view that "All models are false, but some are useful." Norbert Wiener, an equally important applied mathematician on whose work much of this book builds, had a different and also-memorable view, asserting that "The best material model of a cat is another, or preferably the same, cat."<sup>9</sup> What did Wiener mean? What is your view?

## 8. Forecasting as an ongoing process in organizations.

We could add another very important item to this chapter's list of considerations basic to successful forecasting – forecasting in organizations is an ongoing process of building, using, evaluating, and improving forecasting models. Provide a concrete example of a forecasting model used in business, finance, economics or government, and discuss ways in which

---

<sup>9</sup>Attributed by Wikiquote to Wiener and Rosenblueth's *Philosophy of Science*, 1945.

each of the following questions might be resolved prior to, during, or after its construction.

- a. Are the data “dirty”? For example, are there “**ragged edges**” (different starting and ending dates of different series)? Are there **missing observations**? Are there aberrant observations, called **outliers**, perhaps due to **measurement error**?  
Are the data stored in a format that inhibits computerized analysis?
- b. Has software been written for importing the data in an ongoing forecasting operation?
- c. Who will build and maintain the model?
- d. Are sufficient resources available (time, money, staff) to facilitate model building, use, evaluation, and improvement on a routine and ongoing basis?
- e. How much time remains before the first forecast must be produced?
- f. How many series must be forecast, and how often must ongoing forecasts be produced?
- g. What level of data **aggregation** or **disaggregation** is desirable?
- h. To whom does the forecaster or forecasting group report, and how will the forecasts be communicated?
- i. How might you conduct a “forecasting audit”?

# **Part II**

## **Cross Sections: Basics**



# Chapter 3

## Predictive Regression: Review and Interpretation

Ideas that fall under the general heading of “**regression analysis**” are crucial for building forecasting models, using them to produce forecasts, and evaluating those forecasts. Here we provide a linear regression refresher. Again, be warned: this chapter is no substitute for a full-introduction to regression, which you should have had already.

### 3.1 Regression as Curve Fitting

#### 3.1.1 Simple Regression

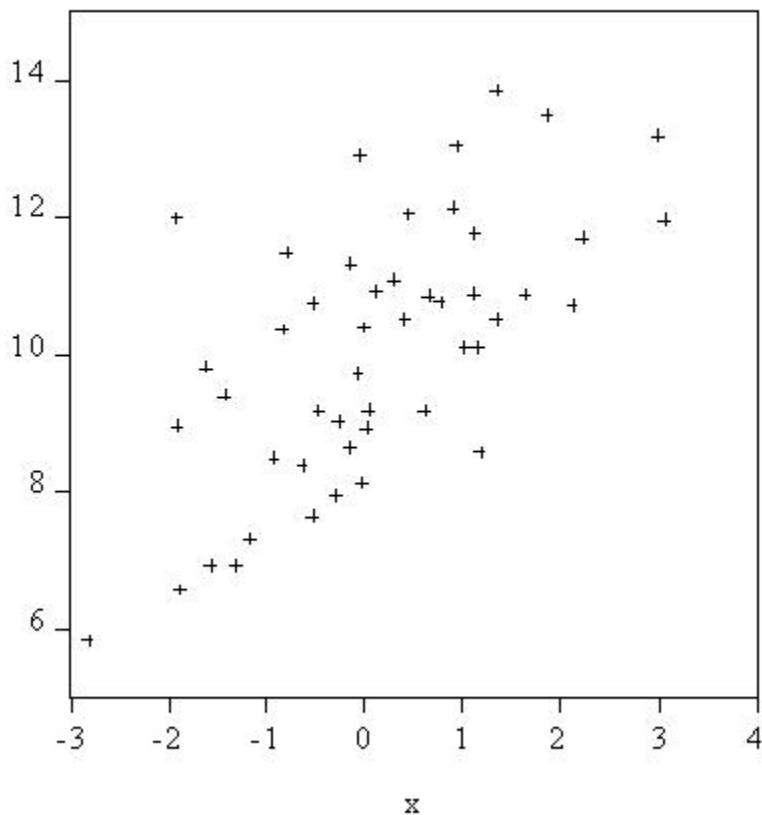
Suppose that we have data on two variables (“simple,” or “bivariate,” regression),  $y$  and  $x$ , as in Figure 1, and suppose that we want to find the linear function of  $x$  that best fits the data points, in the sense that the sum of squared vertical distances of the data points from the fitted line is minimized. When we “run a regression,” or “fit a regression line,” that’s what we do. The estimation strategy is called **least squares**.

In Figure 2, we illustrate graphically the results of regressing  $y$  on  $x$ , which we sometimes denote by  $y \rightarrow c, x$ .<sup>1</sup> The best-fitting line slopes upward,

---

<sup>1</sup>The “c” denotes inclusion of a constant, or intercept, term.

**Figure 1**  
Scatterplot of  $y$  versus  $x$



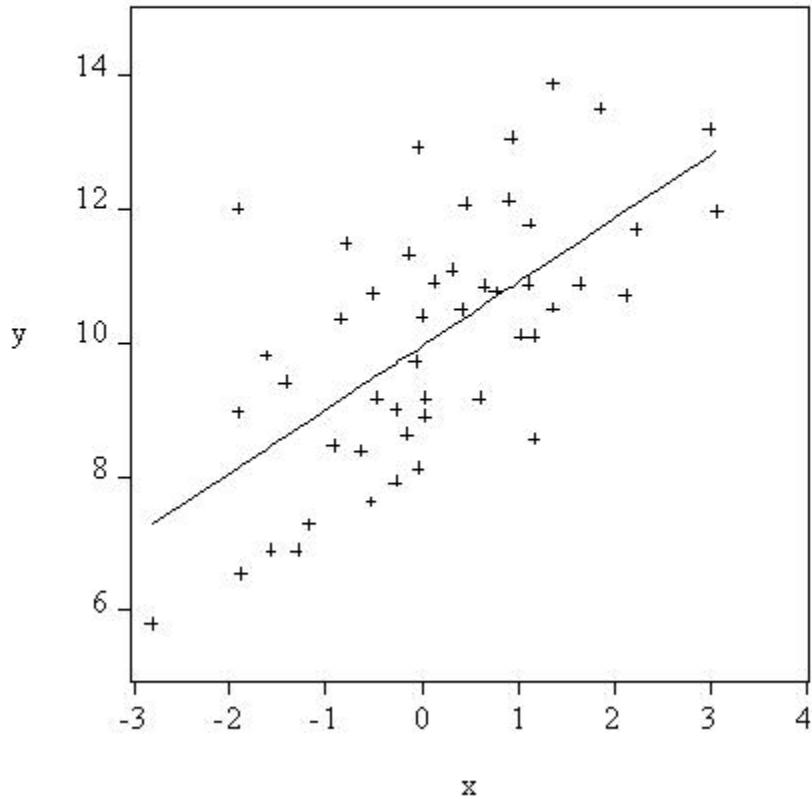
reflecting the positive correlation between  $y$  and  $x$ . Note that the data points don't satisfy the fitted linear relationship exactly; rather, they satisfy it on average.

Let us elaborate on the fitting of regression lines, and the reason for the name “least squares.” When we run the regression, we use a computer to fit the line by solving the problem

$$\min_{\beta} \sum_{t=1}^T (y_t - \beta_1 - \beta_2 x_t)^2,$$

where  $\beta$  is shorthand notation for the set of two parameters,  $\beta_1$  and  $\beta_2$ . We denote the set of estimated, or fitted, parameters by  $\hat{\beta}$ , and its elements by

**Figure 2**  
 Scatterplot of  $y$  versus  $x$   
 Regression Line Superimposed



$\hat{\beta}_1$  and  $\hat{\beta}_2$ .

The regression **fitted values** are

$$\hat{y}_t = \hat{\beta}_1 + \hat{\beta}_2 x_t,$$

$t = 1, \dots, T$ . The regression **residuals** are simply the difference between actual and fitted values. We write

$$e_t = y_t - \hat{y}_t,$$

$t = 1, \dots, T$ .

In all linear regressions (even with multiple RHS variables, to which we turn shortly), the least-squares estimator has a simple formula. We use a

computer to evaluate the formula, simply, stably, and instantaneously.

### 3.1.2 Multiple Regression

Extension to the general multiple linear regression model, with an arbitrary number of right-hand-side variables ( $K$ , including the constant), is immediate. We simply run  $y \rightarrow c, x_2, \dots, x_K$ , again picking the parameters to minimize the sum of squared residuals, and everything goes through as in the case of simple regression.

The least squares estimator is

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y, \quad (3.1)$$

where  $X$  is a  $T \times K$  matrix,

$$X = \begin{pmatrix} 1 & x_{21} & x_{31} & \dots & x_{K1} \\ 1 & x_{22} & x_{32} & \dots & x_{K2} \\ \vdots & & & & \\ 1 & x_{2T} & x_{3T} & \dots & x_{KT} \end{pmatrix}$$

and  $y$  is a  $T \times 1$  vector,  $y' = (y_1, y_2, \dots, y_T)$ . The time- $t$  fitted value is

$$\hat{y}_t = x_t' \hat{\beta},$$

where  $x_t' = (x_{1t}, \dots, x_{Kt})$  is the time- $t$  vector of  $x$ 's, and the time- $t$  residual is

$$e_t = y_t - \hat{y}_t.$$

The vector of fitted values is

$$\hat{y} = X \hat{\beta},$$

and the vector of residuals is

$$e = y - \hat{y}.$$

## 3.2 Regression as a Probability Model

We work with the full multiple regression model; simple regression is of course a special case.

### 3.2.1 A Population Model and a Sample Estimator

Thus far we have *not* postulated a probabilistic model that relates  $y_t$  and  $x_t$ ; instead, we simply ran a mechanical regression of  $y_t$  on  $x_t$  to find the best fit to  $y_t$  formed as a linear function of  $x_t$ . It's easy, however, to construct a probabilistic framework that lets us make statistical assessments about the properties of the fitted line. Assume, for example, that  $y_t$  is linearly related to an exogenously-determined  $x_t$ , with an independent and identically distributed zero-mean (iid) Gaussian **disturbance**:

$$y_t = \beta_1 + \beta_2 x_{2t} + \dots + \beta_K x_{Kt} + \varepsilon_t = x_t' \beta + \varepsilon_t$$

$$\varepsilon_t \sim iidN(0, \sigma^2),$$

$t = 1, \dots, T$ . The intercept of the line is  $\beta_1$ , the slope parameters are the  $\beta_i$ 's, and the variance of the disturbance is  $\sigma^2$ .<sup>2</sup> Collectively, we call the  $\beta$ 's the model's **parameters**. The index  $t$  keeps track of time; the data sample begins at some time we've called "1" and ends at some time we've called " $T$ ", so we write  $t = 1, \dots, T$ . (Or, in cross sections, we index cross-section units by  $i$  and write  $i = 1, \dots, N$ .)

In this linear regression model the expected value of  $y_t$  conditional upon  $x_t$  taking a particular value, say  $x_t^*$ , is

$$E(y_t | x_t = x_t^*) = x_t^{*\prime} \beta.$$

That is, the **regression function** is the **conditional expectation** of  $y_t$ .

---

<sup>2</sup>We speak of the **regression intercept** and the **regression slope**.

We assume that the linear model sketched is true in population; that is, it is the **data-generating process (DGP)**. But in practice, of course, we don't know the values of the model's parameters,  $\beta_1, \beta_2, \dots, \beta_K$  and  $\sigma^2$ . Our job is to *estimate* them using a sample of data from the population. We estimate the  $\beta$ 's precisely as before, using the computer to solve  $\min_{\beta} \sum_{t=1}^T \varepsilon_t^2$ .

### 3.2.2 Notation, Assumptions and Results: The Full Ideal Conditions

The discussion thus far was intentionally a bit loose, focusing on motivation and intuition. Let us now be more precise about what we assume and what results obtain.

#### A Bit of Matrix Notation

One reason that vector-matrix notation is useful is because the probabilistic regression model can be written very compactly using it. We have written the model as

$$y_t = \beta_1 + \beta_2 x_{2t} + \dots + \beta_K x_{Kt} + \varepsilon_t, \quad t = 1, \dots, T.$$

$$\varepsilon_t \sim iid N(0, \sigma^2)$$

Now stack  $\varepsilon_t, t = 1, \dots, T$ , into the vector  $\varepsilon$ , where  $\varepsilon' = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T)$ . Then we can write the complete model over all observations as

$$y = X\beta + \varepsilon \tag{3.2}$$

$$\varepsilon \sim N(\underline{0}, \sigma^2 I). \tag{3.3}$$

This concise representation is very convenient.

Indeed representation (3.2)-(3.3) is crucially important, not simply because it is concise, but because the various assumptions that we need to

make to get various statistical results are most naturally and simply stated on  $X$  and  $\varepsilon$  in equation (3.2).

The most restrictive set of assumptions is known as the “full ideal conditions” (FIC), which are so strict as to be nearly preposterous in economic contexts, and most of econometrics is devoted to confronting various *failures* of the FIC. But before we worry about FIC failures, it’s useful first to recall what happens when they hold.

#### Assumptions: The Full Ideal Conditions (FIC)

1. The DGP is (3.2)-(3.3), and the fitted model matches the DGP exactly.
2.  $X$  is fixed in repeated samples.
3.  $X$  is of full column rank ( $K$ ).

FIC 1 has many important sub-conditions embedded. For example:

1. Linear relationship,  $E(y) = X\beta$
2. Fixed coefficients,  $\beta$
3.  $\varepsilon \sim N$
4.  $\varepsilon$  has constant variance  $\sigma^2$
5. The  $\varepsilon$ ’s are uncorrelated.

FIC 2 says that re-running the world to generate a new sample  $y^*$  would entail simply generating new shocks  $\varepsilon^*$  and running them through equation (3.2):

$$y^* = X\beta + \varepsilon^*.$$

That is,  $X$  would stay fixed across replications.

FIC 3 just says “no multicollinearity” – i.e., no redundancy among the variables contained in  $X$  (more precisely, no regressor is a perfect linear combination of the others).

### Results Under the FIC

The least squares estimator remains

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y,$$

but in a probabilistic interpretation under the FIC, we can say a great deal about its statistical properties. Among other things, it is minimum-variance unbiased (MVUE) and normally distributed with covariance matrix  $\sigma^2(X'X)^{-1}$ . We write

$$\hat{\beta}_{OLS} \sim N(\beta, \sigma^2(X'X)^{-1}).$$

We estimate the covariance matrix  $\sigma^2(X'X)^{-1}$  using  $s^2(X'X)^{-1}$ , where

$$s^2 = \frac{\sum_{t=1}^T e_t^2}{T - K}.$$

## 3.3 A Typical Regression Analysis

Consider a typical regression output, which we show in Table 1. We do so dozens of times in this book, and the output format and interpretation are always the same, so it’s important to get comfortable with it quickly. The output is in Eviews format. Other software will produce more-or-less the same information, which is fundamental and standard.

The results begin by reminding us that we’re running a least-squares (LS) regression, and that the left-hand-side variable is  $y$ . It then shows us the sample range of the historical data, which happens to be 1960 to 2007, for a total of 48 observations.

**Table 1**  
**Regression of y on x and z**

LS // Dependent Variable is Y

Sample: 1960 2007

Included observations: 48

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.884732	0.190297	51.94359	0.0000
X	1.073140	0.150341	7.138031	0.0000
Z	-0.638011	0.172499	-3.698642	0.0006
R-squared	0.552928	Mean dependent var		10.08241
Adjusted R-squared	0.533059	S.D. dependent var		1.908842
S.E. of regression	1.304371	Akaike info criterion		3.429780
Sum squared resid	76.56223	Schwarz criterion		3.546730
Log likelihood	-79.31472	F-statistic		27.82752
Durbin-Watson stat	1.506278	Prob(F-statistic)		0.000000

Next comes a table listing each right-hand-side variable together with four statistics. The right-hand-side variables  $x$  and  $z$  need no explanation, but the variable  $c$  does.  $c$  is notation for the earlier-mentioned constant variable. The  $c$  variable always equals one, so the estimated coefficient on  $c$  is the estimated intercept of the regression line.<sup>3</sup>

### 3.3.1 Coefficient Estimates, Standard Errors, $t$ Statistics and $p$ -Values

The four statistics associated with each right-hand-side variable are the estimated coefficient (“Coefficient”), its standard error (“Std. Error”), a  $t$  statistic, and a corresponding probability value (“Prob.”).

The “coefficients” are simply the regression coefficient estimates. Per the OLS formula that we introduced earlier in equation (3.1), they are the elements of the  $(K \times 1)$  vector,  $(X'X)^{-1}X'y$ .

---

<sup>3</sup>Sometimes the population coefficient on  $c$  is called the **constant term**, and the regression estimate is called the estimated constant term.

The **standard errors** of the estimated coefficients indicate their sampling variability, and hence their reliability. In line with result (??) above, the  $i$ th standard error is

$$s\sqrt{(X'X)_{ii}^{-1}},$$

where  $(X'X)_{ii}^{-1}$  denotes the  $i$ th diagonal element of  $(X'X)^{-1}$ , and  $s$  is an estimate (defined below) of  $\sigma$ .

The estimated coefficient plus or minus one standard error is approximately a 68% confidence interval for the true but unknown population parameter (contribution to the conditional expectation), and the estimated coefficient plus or minus two standard errors is approximately a 95% confidence interval, assuming that the estimated coefficient is approximately normally distributed.<sup>4</sup> Thus large coefficient standard errors translate into wide confidence intervals.

Each  $t$  **statistic** provides a test of the hypothesis of variable irrelevance: that the true but unknown population parameter (contribution to the conditional expectation) is zero, so that the corresponding variable contributes nothing to the conditional expectation and can therefore be dropped. One way to test this variable irrelevance, with, say, a 5% probability of incorrect rejection, is to check whether zero is outside the 95% confidence interval for the parameter. If so, we reject irrelevance. The  $t$  statistic is just the ratio of the estimated coefficient to its standard error, so if zero is outside the 95% confidence interval, then the  $t$  statistic must be bigger than two in absolute value. Thus we can quickly test irrelevance at the 5% level by checking whether the  $t$  statistic is greater than two in absolute value.<sup>5</sup>

Finally, associated with each  $t$  statistic is a **probability value**, which is the probability of getting a value of the  $t$  statistic at least as large in

<sup>4</sup>Coefficients will be approximately normally distributed in large samples quite generally, and exactly normally distributed in samples of any size if the regression disturbance is normally distributed.

<sup>5</sup>In large samples the  $t$  statistic is distributed  $N(0, 1)$  quite generally. In samples of any size the  $t$  statistic follows a  $t$  distribution if the regression disturbances are Gaussian.

absolute value as the one actually obtained, assuming that the irrelevance hypothesis is true. Hence if a  $t$  statistic were two, the corresponding probability value would be approximately .05 (assuming large  $T$  and/or Gaussian disturbances). The smaller the probability value, the stronger the evidence against irrelevance. There's no magic cutoff, but typically probability values less than 0.1 are viewed as strong evidence against irrelevance, and probability values below 0.05 are viewed as very strong evidence against irrelevance. Probability values are useful because they eliminate the need for consulting tables of the  $t$  or  $z$  distributions. Effectively the computer does it for us and tells us the significance level at which the irrelevance hypothesis is just rejected.

Now let's interpret the actual estimated coefficients, standard errors,  $t$  statistics, and probability values. The estimated intercept is approximately 10, so that conditional on  $x$  and  $z$  both being zero, we expect  $y$  to be 10. Moreover, the intercept is very precisely estimated, as evidenced by the small standard error relative to the estimated coefficient. An approximate 95% confidence interval for the true but unknown population intercept is  $10 \pm 2(.19)$ , or [9.62, 10.38]. Zero is far outside that interval, so the corresponding  $t$  statistic is huge, with a probability value that's zero to four decimal places.

The estimated coefficient on  $x$  is 1.07, and the standard error is again small in relation to the size of the estimated coefficient, so the  $t$  statistic is large and its probability value small. Hence at conventional levels we reject the hypothesis that  $x$  contributes nothing to the conditional expectation  $E(y|x, z)$ . The estimated coefficient is positive, so that  $x$  contributes positively to the conditional expectation; that is,  $E(y|x, z)$  is larger for larger  $x$ , other things equal.

The estimated coefficient on  $z$  is -.64. Its standard error is larger relative to the estimated parameter; hence its  $t$  statistic is smaller than those of the other coefficients. The standard error is nevertheless small, and the absolute

value of the  $t$  statistic is still well above 2, with a small probability value of .06%. Hence at conventional levels we reject the hypothesis that  $z$  contributes nothing to the conditional expectation  $E(y|x, z)$ . The estimated coefficient is negative, so that  $z$  contributes negatively to the conditional expectation; that is,  $E(y|x, z)$  is smaller for larger  $z$ , other things equal.

### 3.3.2 Residual Plot

After running a time-series regression, it's usually a good idea to assess the adequacy of the model by plotting over time and examining the actual data ( $y_t$ 's), the fitted values ( $\hat{y}_t$ 's), and the residuals ( $e_t$ 's). Often we'll refer to such plots, shown together in a single graph, as a **residual plot**.<sup>6</sup> In Figure 4 we show the residual plot for the regression of  $y \rightarrow c, x, z$ . The actual (short dash) and fitted (long dash) values appear at the top of the graph; their scale is on the right. The fitted values track the actual values fairly well. The residuals appear at the bottom of the graph (solid line); their scale is on the left. It's important to note that the scales differ; the  $e_t$ 's are in fact substantially smaller and less variable than either the  $y_t$ 's or the  $\hat{y}_t$ 's. We draw the zero line through the residuals for visual comparison. There are no obvious patterns in the residuals.

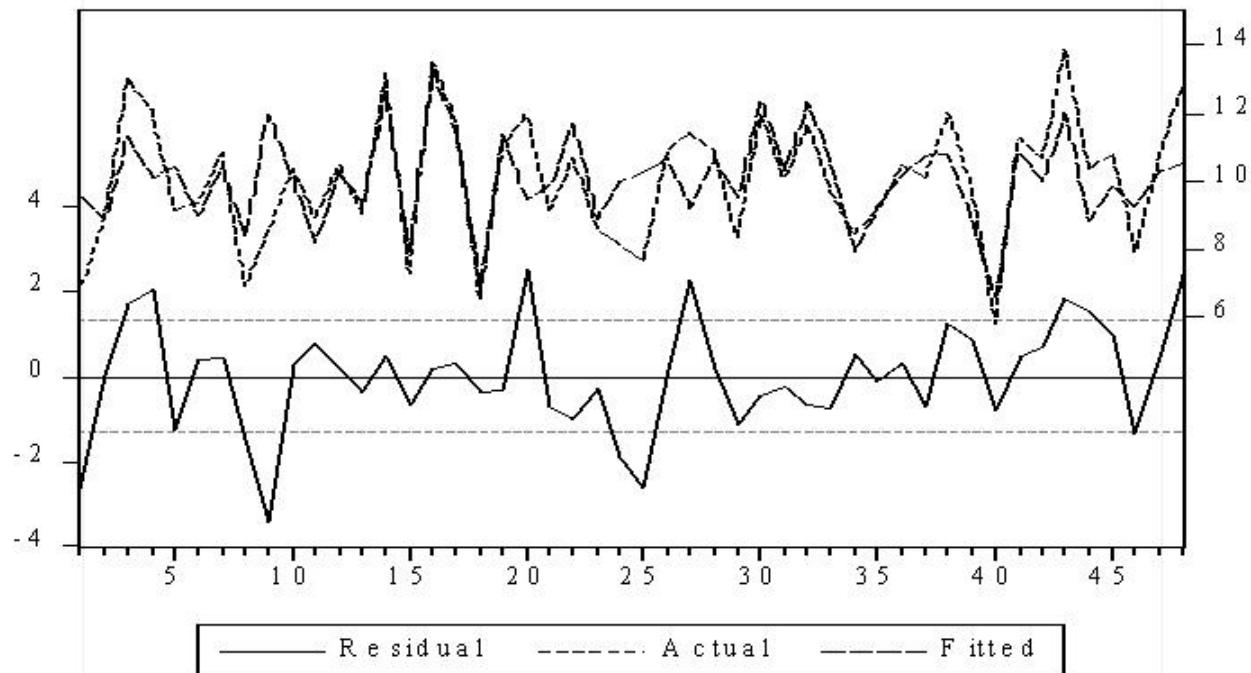
Residual plots are obviously useful in time-series perspective, and not useful in cross sections, for which there is no natural ordering of the data. In cross sections, however, we often examine **residual scatterplots**, that is, scatterplots of  $y$  vs.  $\hat{y}$  for all observations in the cross section, with special attention paid to the general pattern of deviations from the forty-five degree line.

A variety of diagnostic statistics follow; they help us to evaluate the adequacy of the regression. Here we review them very briefly.

---

<sup>6</sup>Sometimes, however, we'll use “residual plot” to refer to a plot of the residuals alone. The intended meaning will be clear from context.

**Figure 4**  
Residual Plot  
Regression of  $y$  on  $x$  and  $z$



### 3.3.3 Mean dependent var

The sample mean of the dependent variable is

$$\bar{y} = \frac{1}{T} \sum_{t=1}^T y_t$$

It measures the central tendency, or location, of  $y$ .

### 3.3.4 S.D. dependent var

The sample standard deviation of the dependent variable is

$$SD = \sqrt{\frac{\sum_{t=1}^T (y_t - \bar{y})^2}{T - 1}}.$$

It measures the dispersion, or scale, of  $y$ .

### 3.3.5 Sum squared resid

Minimizing the **sum of squared residuals** is the objective of least squares estimation. It's natural, then, to record the minimized value of the sum of squared residuals. In isolation it's not of much value, but it serves as an input to other diagnostics that we'll discuss shortly. Moreover, it's useful for comparing models and testing hypotheses. The formula is

$$SSR = \sum_{t=1}^T e_t^2.$$

### 3.3.6 F–statistic

We use the **F statistic** to test the hypothesis that the coefficients of all variables in the regression except the intercept are jointly zero.<sup>7</sup> That is, we test whether, taken jointly as a set, the variables included in the model contribute nothing to the expectation of  $y$  conditional on the variables. This contrasts with the  $t$  statistics, which we use to examine the contributions of the variables one at a time.<sup>8</sup> If no variables contribute, then if the regression disturbances are Gaussian the  $F$  statistic follows an  $F$  distribution with  $K - 1$  and  $T - K$  degrees of freedom. The formula is

$$F = \frac{(SSR_{res} - SSR)/(K - 1)}{SSR/(T - K)},$$

where  $SSR_{res}$  is the sum of squared residuals from a *restricted* regression that contains only an intercept. Thus the test proceeds by examining how much  $SSR$  increases when all the variables except the constant are dropped. If it increases by a great deal, there's evidence that at least one of the variables

---

<sup>7</sup>We don't want to restrict the intercept to be zero, because under the hypothesis that all the other coefficients are zero, the intercept would equal the mean of  $y$ , which in general is not zero.

<sup>8</sup>In the case of only one right-hand-side variable, the  $t$  and  $F$  statistics contain exactly the same information, and one can show that  $F = t^2$ . When there are two or more right-hand-side variables, however, the hypotheses tested differ, and  $F \neq t^2$ .

contributes to the conditional expectation.

### 3.3.7 Prob(F-statistic)

The probability value for the  $F$  statistic gives the significance level at which we can just reject the hypothesis that the set of right-hand-side variables makes no contribution to the conditional expectation. Here the value is indistinguishable from zero, so we reject the hypothesis overwhelmingly.

### 3.3.8 S.E. of regression

We'd like an estimate of  $\sigma^2$ , because  $\sigma^2$  tells us whether the regression "fit" is good. The observed residuals, the  $e_t$ 's, are effectively estimates of the unobserved population disturbances, the  $\varepsilon_t$ 's. Thus the sample variance of the  $e$ 's, which we denote  $s^2$  (read "s-squared"), is a natural estimator of  $\sigma^2$ :

$$s^2 = \frac{\sum_{t=1}^T e_t^2}{T - K}.$$

$s^2$  is an estimate of the dispersion of the regression disturbance and hence is used to assess goodness of fit. The larger is  $s^2$ , the worse the model's fit.  $s^2$  involves a degrees-of-freedom correction (division by  $T - K$  rather than by  $T$  or  $T - 1$ ), which, but whether one divides by  $T$  or  $T - K$  is of no asymptotic consequence.

The **standard error of the regression** (SER) conveys the same information; it's an estimator of  $\sigma$  rather than  $\sigma^2$ , so we simply use  $s$  rather than  $s^2$ . The formula is

$$SER = s = \sqrt{s^2} = \sqrt{\frac{\sum_{t=1}^T e_t^2}{T - K}}.$$

The standard error of the regression is easier to interpret than  $s^2$ , because its units are the same as those of the  $e$ 's, whereas the units of  $s^2$  are not. If

the  $e$ 's are in dollars, then the squared  $e$ 's are in dollars squared, so  $s^2$  is in dollars squared. By taking the square root at the end of it all,  $SER$  converts the units back to dollars.

### 3.3.9 R-squared

If an intercept is included in the regression, as is almost always the case,  $R$ -squared must be between zero and one. In that case,  $R$ -squared, usually written  $R^2$ , is the percentage of the variance of  $y$  explained by the variables included in the regression.  $R^2$  is widely used as an easily-interpreted check of **goodness of fit**. Here the  $R^2$  is about 55% – good but not great.

The formula is for  $R^2$  is

$$R^2 = 1 - \frac{\sum_{t=1}^T e_t^2}{\sum_{t=1}^T (y_t - \bar{y})^2}.$$

The key is the ratio on the right of the formula. First, note that the ratio must be positive (it exclusively involves sums of squares) and less than one (if the regression includes an intercept). Hence  $R^2$ , which is one *minus* the ratio, must be in  $[0, 1]$ . Second, note what the ratio involves. The numerator is SSR from a regression on *all variables*, and the denominator is the is SSR from a regression on an intercept alone. Hence the ratio is the fraction of variation in  $y$  *not* explained by the included  $x$ 's, so that  $R^2$  – which, again, is one *minus* the ratio – is the fraction of variation in  $y$  that *is* explained by the included  $x$ 's.

We can write  $R^2$  in a more roundabout way as

$$R^2 = 1 - \frac{\frac{1}{T} \sum_{t=1}^T e_t^2}{\frac{1}{T} \sum_{t=1}^T (y_t - \bar{y})^2}.$$

This proves useful for moving to *adjusted R*<sup>2</sup>, to which we now turn.

### 3.3.10 Adjusted R-squared

The interpretation is the same as that of  $R^2$ , but the formula is a bit different. Adjusted  $R^2$  incorporates adjustments for the  $K$  degrees of freedom used in fitting the full model to  $y$  (numerator of the ratio), and for the 1 degree of freedom used in fitting the a mean to  $y$  (denominator of the ratio). As long as there is more than one right-hand-side variable in the model fitted, adjusted  $\bar{R}^2$  is smaller than  $R^2$ ; here, however, the two are typically very close (in this case, 53% vs. 55%). The formula for  $\bar{R}^2$  is

$$\bar{R}^2 = 1 - \frac{\frac{1}{T-K} \sum_{t=1}^T e_t^2}{\frac{1}{T-1} \sum_{t=1}^T (y_t - \bar{y})^2},$$

where  $K$  is the number of right-hand-side variables, including the constant term. The numerator in the large fraction is precisely  $s_e^2$ , and the denominator is precisely  $s_y^2$ .

### 3.3.11 Durbin-Watson stat

We're always interested in examining whether there are patterns in residuals; if there are, the model somehow missed something systematic in the  $y$  data. The **Durbin-Watson statistic** tests for a certain kind of pattern, correlation over time, called **serial correlation**.

The Durbin-Watson test works within the context of the model

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 z_t + \varepsilon_t$$

$$\varepsilon_t = \phi \varepsilon_{t-1} + v_t$$

*iid*

$$v_t \sim N(0, \sigma^2).$$

The regression disturbance is serially correlated when  $\phi \neq 0$ . The hypothesis

of interest is  $\phi = 0$ . When  $\phi \neq 0$ , the disturbance is serially correlated. More specifically, when  $\phi \neq 0$ , we say that  $\varepsilon_t$  follows an autoregressive process of order one, or  $AR(1)$  for short.<sup>9</sup> If  $\phi > 0$  the disturbance is positively serially correlated, and if  $\phi < 0$  the disturbance is negatively serially correlated. **Positive serial correlation** is typically the relevant alternative in the economic and financial applications that will concern us.

The formula for the Durbin-Watson (DW) statistic is

$$DW = \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T e_t^2}.$$

$DW$  takes values in the interval  $[0, 4]$ , and if all is well,  $DW$  should be around 2. If  $DW$  is substantially less than 2, there is evidence of positive serial correlation. As a rough rule of thumb, if  $DW$  is less than 1.5, there may be cause for alarm, and we should consult the tables of the  $DW$  statistic, available in many statistics and econometrics texts. Here  $DW$  is very close to 1.5. A look at the tables of the  $DW$  statistic reveals, however, that we would not reject the null hypothesis at the five percent level. (Why Eviews neglects to print a p-value is something of a mystery.)

Note well that  $DW$  and its good properties involve several strict assumptions. Gaussian disturbances are required, and the  $AR(1)$  alternative is the only one explicitly entertained, whereas in reality much richer forms of serial correlation may arise, and disturbances may of course be non-Gaussian. Subsequently we will introduce much more flexible approaches to testing/assessing residual serial correlation.

### 3.3.12 Akaike info criterion and Schwarz criterion

The Akaike and Schwarz criteria are used for model selection, and in certain contexts they have provable optimality properties in that regard. The

---

<sup>9</sup>The Durbin-Watson test is designed to be very good at detecting serial correlation of the  $AR(1)$  type. Many other types of serial correlation are possible; we'll discuss them extensively later.

formulas are:

$$AIC = e^{\left(\frac{2K}{T}\right)} \frac{\sum_{t=1}^T e_t^2}{T}$$

and

$$SIC = T^{\left(\frac{K}{T}\right)} \frac{\sum_{t=1}^T e_t^2}{T}.$$

Both are penalized versions of the mean-squared residual, where the penalties are functions of the degrees of freedom used in fitting the model. For both  $AIC$  and  $SIC$ , “smaller is better.” We will have much more to say about them in section ?? below.

### 3.3.13 Log Likelihood

The **likelihood function** is tremendously important in statistics, as it summarizes all the information contained in the data. It is simply the joint density function of the data, viewed as a function of the model parameters.

The number reported is the maximized value of the log of the likelihood function under the assumption of Gaussian disturbances.<sup>10</sup> Like the sum of squared residuals,  $SIC$  and  $AIC$ , it’s not of much use alone, but it’s useful for comparing models and testing hypotheses. We will sometimes use the maximized log likelihood function directly, although we’ll often focus on the minimized sum of squared residuals.

## 3.4 Regression From a Forecasting Perspective

### 3.4.1 The Key to Everything (or at Least Many Things)

Linear least squares regression, by construction, is consistent under very general conditions for “the linear function of  $x_t$  that gives the best approximation

---

<sup>10</sup>Throughout, “log” refers to a natural (base e) logarithm.

to  $y_t$  under squared-error loss,” which is the linear projection,

$$P(y_t|x_t) = x_t' \beta.$$

If the conditional expectation  $E(y_t|x_t)$  is linear in  $x_t$ , then the linear projection and the conditional expectation coincide, and OLS is consistent the for conditional expectation  $E(y_t|x_t)$ .

Hence to forecast  $y_t$  for any given value of  $x_t$ , we can use the fitted line to find the value of  $y_t$  that corresponds to the given value of  $x_t$ . In large samples that “linear least squares forecast” of  $y_t$  will either be the conditional mean  $E(y_t|x_t)$ , which as we mentioned earlier in Chapter 2 is the optimal forecast under quadratic loss, or the best linear approximation to it,  $P(y_t|x_t)$ .

One leading case in which the linear projection and conditional mean coincide (that is,  $E(y_t|x_t)$  is linear in  $x_t$ ) is joint normality. In particular, suppose that

$$\begin{pmatrix} y \\ x \end{pmatrix} \sim N \left( \mu, \Sigma \right)$$

where

$$\mu = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix}.$$

Then it can be shown that:

$$y|x \sim N \left( \mu_{y|x}, \Sigma_{y|x} \right)$$

where

$$\begin{aligned} \mu_{y|x} &= \mu_y + \Sigma_{yx} \Sigma_{xx}^{-1} (x - \mu_x) \\ \Sigma_{y|x} &= \Sigma_{yy} - \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy} \end{aligned}$$

In what follows we’ll often casually speak of linear regression as estimating the conditional expectation function. You can think of a Gaussian situation, or you can just mentally replace “conditional expectation” with “linear

projection.” We’ll also implicitly assume quadratic loss, which is why we’re interested in the conditional mean in the first place.

### 3.4.2 Why Take a Probabilistic Approach to Regression, as Opposed to Pure Curve Fitting?

We want conditional mean point forecasts, and the conditional mean is a probabilistic concept. We also may want to test hypotheses regarding which variables actually enter in the determination of the conditional mean. Last and not at all least, we also want to quantify the *uncertainty* associated with our forecasts – that is, we want interval and density forecasts – and doing so requires probabilistic modeling.

### 3.4.3 Regression For Estimating Conditional Means is Regression for Forecasting

We already introduced this, but we repeat for emphasis: In our regression model, the expected value of  $y_t$  conditional on  $x_t$  is

$$E(y_t|x_t) = x_t' \beta.$$

That is, the **regression function** is the **conditional expectation** of  $y_t$  given  $x_t$ .

This is crucial for forecasting, because the expectation of future  $y$  conditional upon available information is a particularly good forecast. In fact, under quadratic loss, it is the best possible forecast (i.e., it minimizes expected loss). The intimate connection between regression and optimal forecasts makes regression an important tool for forecasting.

### 3.4.4 LS and Quadratic Loss

Quadratic loss is routinely invoked for prediction, in which case the conditional mean is the optimal forecast, as mentioned above. OLS optimizes quadratic loss in estimation, and it's good to have the model estimation criterion match the predictive criterion.

### 3.4.5 Estimated Coefficient Signs and Sizes

The “best fit” that OLS delivers is effectively a best (in-sample) forecast. Each estimated coefficient gives the weight put on the corresponding  $x$  variable in forming the best linear in-sample forecast of  $y$ .

### 3.4.6 Standard Errors, $t$ Statistics, $p$ -values, $F$ Statistic, Log Likelihood, etc.

These let us do formal statistical inference as to which regressors are relevant for forecasting  $y$ .

### 3.4.7 Fitted Values and Residuals

The fitted values are effectively in-sample forecasts:

$$\hat{y}_t = x_t' \hat{\beta},$$

$t = 1, \dots, T$ . The in-sample forecast is automatically unbiased if an intercept is included, because the residuals must then sum to 0 (see EPC 2).

The residuals are effectively in-sample forecast errors:

$$e_t = y_t - \hat{y}_t,$$

$t = 1, \dots, T$ .

Forecasters are keenly interested in studying the properties of their forecast errors. Systematic patterns in forecast errors indicate that the forecasting model is inadequate – as we will show and explore later in great depth, forecast errors from a good forecasting model must be unforecastable! And again, residuals are in-sample forecast errors.

### 3.4.8 Mean and Variance of Dependent Variable

An obvious benchmark forecast is the sample, or historical, mean of  $y$ , an estimate of the *unconditional* mean of  $y$ . It's obtained by regressing  $y$  on an intercept alone – no conditioning on other regressors!

The sample standard deviation of  $y$  is a measure of the (in-sample) accuracy of the unconditional mean forecast under quadratic loss.

It's natural to compare the accuracy of our conditional-mean forecasts to naive unconditional-mean forecasts.  $R^2$  and  $\bar{R}^2$ , to which we now turn, do precisely that.

### 3.4.9 $R^2$ and $\bar{R}^2$

Hopefully conditional-mean forecasts that condition on regressors other than just an intercept are better than naive unconditional-mean forecasts.  $R^2$  and  $\bar{R}^2$  effectively *compare* the in-sample accuracy of conditional-mean and unconditional-mean forecasts.

$$R^2 = 1 - \frac{\frac{1}{T} \sum_{t=1}^T e_t^2}{\frac{1}{T} \sum_{t=1}^T (y_t - \bar{y})^2}.$$

$$\bar{R}^2 = 1 - \frac{\frac{1}{T-K} \sum_{t=1}^T e_t^2}{\frac{1}{T-1} \sum_{t=1}^T (y_t - \bar{y})^2},$$

### 3.4.10 *SSR* (or *MSE*), *SER* (or Residual $s^2$ ), *AIC* and *SIC*

Each attempts an estimate of *out-of-sample* forecast accuracy (which is what we really care about) on the basis of in-sample forecast accuracy, with an eye toward selecting forecasting models. (That is, we'd like to select a forecasting model that will perform well for out-of-sample forecasting, quite apart from its in-sample fit.) Everything proceeds by inflating the in-sample mean-squared error (*MSE*), in various attempts to offset the deflation from regression fitting, to obtain a good estimate of out-of-sample mean-squared error.

We have:

$$\begin{aligned} MSE &= \frac{\sum_{t=1}^T e_t^2}{T} \\ s^2 &= \left( \frac{T}{T-K} \right) MSE \\ AIC &= \left( e^{(\frac{2K}{T})} \right) MSE \\ SIC &= \left( T^{(\frac{K}{T})} \right) MSE. \end{aligned}$$

We will have much more to say about *AIC* and *SIC* in section ?? below.

### 3.4.11 Durbin–Watson

We mentioned earlier that we're interested in examining whether there are patterns in our forecast errors, because errors from a good forecasting model should be unforecastable. The Durbin–Watson statistic tests for a particular and important such pattern, serial correlation. If the errors made by a forecasting model are serially correlated, then they are forecastable, and we could improve the forecasts by forecasting the forecast errors! We will subsequently discuss such issues at great length.

### 3.4.12 Residual Plots

Residual plots are useful for visually flagging neglected things that impact forecasting. Residual serial correlation indicates that point forecasts could be improved. Residual volatility clustering indicates that interval forecasts and density could be improved. (Why?) Evidence of structural change in residuals indicates that *everything* could be improved.

## 3.5 Exercises, Problems and Complements

1. Regression, regression diagnostics, and regression graphics in action.

At the end of each quarter, you forecast a series  $y$  for the next quarter. You do this using a regression model that relates the current value of  $y$  to the lagged value of a single predictor  $x$ . That is, you regress

$$y_t \rightarrow c, x_{t-1}.$$

(In your computer workfile,  $y_t$  is called Y, and  $x_{t-1}$  is called XLAG1. So you run

$$Y \rightarrow c, XLAG1.$$

- Why might include a *lagged*, rather than current, right-hand-side variable?
- Graph Y vs. XLAG1 and discuss.
- Régress Y on XLAG1 and discuss (including related regression diagnostics that you deem relevant).
- Consider as many variations as you deem relevant on the general theme. At a minimum, you will want to consider the following:
  - Does it appear necessary to include an intercept in the regression?

- ii. Does the functional form appear adequate? Might the relationship be nonlinear?
  - iii. Do the regression residuals seem completely random, and if not, do they appear serially correlated, heteroskedastic, or something else?
  - iv. Are there any outliers? If so, does the estimated model appear robust to their presence?
  - v. Do the regression disturbances appear normally distributed?
  - vi. How might you assess whether the estimated model is structurally stable?
2. Least-squares regression residuals have zero mean.

Prove that least-squares regression residuals must sum to zero, and hence must have zero mean, if an intercept is included in the regression. Hence in-sample regression “forecasts” are unbiased.

3. Conditional mean and variance

Consider the regression model,

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 x_t^2 + \beta_4 z_t + \varepsilon_t$$

$$\begin{aligned} & \text{iid} \\ \varepsilon_t & \sim (0, \sigma^2). \end{aligned}$$

- (a) Find the mean of  $y_t$  conditional upon  $x_t = x_t^*$  and  $z_t = z_t^*$ . Does the conditional mean vary with the conditioning information  $(x_t^*, z_t^*)$ ? Discuss.
- (b) Find the variance of  $y_t$  conditional upon  $x_t = x_t^*$  and  $z_t = z_t^*$ . Does the conditional variance vary with the conditioning information  $(x_t^*, z_t^*)$ ? Discuss.

4. Conditional means vs. linear projections.

Consider a scalar  $y$  and a vector  $x$ , with joint density  $f(y, x)$ .

- (a) The conditional mean  $E(y|x)$  is not *necessarily* linear in  $x$ . Give an example of a non-linear conditional mean function.
- (b) Consider such a non-linear conditional mean situation. You assume (incorrectly) that a linear regression model holds. You estimate the model by OLS. We say that you are estimating “**linear projection weights**.” Linear projection weights are best linear approximations (under quadratic loss) to conditional expectations.
- (c) Consider again such a non-linear conditional mean situation. In large samples and under quadratic loss, what can be said about the comparative merits of conditional-mean vs. linear-projection forecasts?
- (d) What factors influence whether your prediction will actually perform well?

5. Squared residual plots for time series.

Consider a time-series plot of *squared* residuals rather than “raw” residuals. Why might that be useful?

6. HAC Standard Errors

Recall that OLS linear regression is consistent for the linear projection under great generality. Also recall, however, that if regression disturbances are autocorrelated and/or heteroskedastic, then OLS standard errors are biased and inconsistent. This is an issue if we want to do inference as to which predictors ( $x$  variables) are of relevance. HAC methods (short for “heteroskedasticity and autocorrelation consistent”) provide a quick and sometimes-useful fix. The variance of the OLS esti-

mator is:

$$\Sigma = (X'X)^{-1}E(X'\varepsilon\varepsilon'X)(X'X)^{-1}.$$

For *iid*  $\varepsilon_t$  this collapses to the usual  $\sigma^2(X'X)^{-1}$ , but otherwise we need the full formula. Write it as:

$$\Sigma = (X'X)^{-1}E(X'\Omega X)(X'X)^{-1}.$$

The question is what estimator to use for  $E(X'\Omega X)$ . In the case of pure heteroskedasticity ( $\Omega$  diagonal but not scalar), we can use the White estimator,

$$E(\widehat{X'\Omega X}) = X'diag(e_1^2, \dots, e_T^2)X = \sum_{t=1}^T e_t^2 x'_t x_t.$$

In the case of heteroskedasticity and autocorrelation, we can use the Newey-West estimator,

$$E(\widehat{X'\Omega X}) = \sum_{t=1}^T e_t^2 x'_t x_t + \sum_{l=1}^m \left(1 - \frac{l}{m+1}\right) \sum_{t=l+1}^T e_t e_{t-l} (x'_t x_{t-l} + x'_{t-l} x_t),$$

where the so-called “truncation lag”  $m$  is chosen by the user. The first Newey-West term is the White estimator, and the second Newey-West term is an additional adjustment for autocorrelation.

There is a strong case against HAC estimators in forecasting contexts: They achieve robust inference for predictor relevance, but they don’t *exploit* any heteroskedasticity present (to improve interval forecasts) or serial correlation present (to improve point forecasts). Nevertheless we introduce them here because (1) they are often produced by regression software, and (2) they can be of use, as we will see, in exploratory modeling en route to arriving at a complete forecasting model.

# Chapter 4

## Forecast Model Building and Use

It has been said that “It’s difficult to make predictions, especially about the future.” This quip is funny insofar as *all* predictions are about the future. But actually they’re not. Prediction is a major topic even in cross-sections, in which there is no temporal aspect. In this chapter we consider cross-section prediction.

### 4.1 Cross-Section Prediction

The environment is:

$$y_i = x'_i \beta + \varepsilon_i, \quad i = 1, \dots, N$$
$$\varepsilon_i \sim \text{iid } D(0, \sigma^2).$$

In cross sections, everything is easy. That is, cross-section prediction simply requires evaluating the conditional expectation (regression relationship) at a *chosen* value of  $x$ ,  $x = x^*$ . Suppose, for example, that we know a regression relationship between expenditure on restaurant meals ( $y$ ) to income ( $x$ ). If we get new income data for a new household, we can use it to predict its restaurant expenditures.

### 4.1.1 Point Prediction

Continue to assume for the moment that we know the model parameters. That is, assume that we know  $\beta$  and all parameters governing  $D$ .<sup>1</sup>

We immediately obtain point forecasts as:

$$E(y_i|x_i = x^*) = x^{*\prime}\beta.$$

### 4.1.2 Density Prediction for $D$ Gaussian

Density forecasts, and hence interval forecasts, are a bit more involved, depending on what we're willing to assume. In any event the key is somehow to account for **disturbance uncertainty**, the part of forecast uncertainty arising from the fact that our forecasting models involve stochastic disturbances.

If  $D$  is Gaussian, then the density prediction is immediately

$$y_i|x_i = x^* \sim N(x^{*\prime}\beta, \sigma^2). \quad (4.1)$$

We can calculate any desired interval forecast from the density forecast. For example, a 95% interval would be  $x^{*\prime}\beta \pm 1.96\sigma$ .

Now let's calculate the density and interval forecasts by a more round-about simulation algorithm that will be very useful in more complicated (and realistic) cases.

1. Take  $R$  draws from the disturbance density  $N(0, \sigma^2)$ .
2. Add  $x^{*\prime}\beta$  to each disturbance draw.
3. Form a density forecast by fitting a density to the output from step 2.
4. Form an interval forecast (95%, say) by sorting the output from step 2 to get the empirical cdf, and taking the left and right interval endpoints

---

<sup>1</sup>Note that the mean and variance are in general insufficient to characterize a non-Gaussian  $D$ .

as the the .025% and .975% values, respectively.

As  $R \rightarrow \infty$ , the algorithmic results coincide with those of 4.1.

### 4.1.3 Density Prediction for $D$ Parametric Non-Gaussian

Our simulation algorithm still works for non-Gaussian  $D$ , so long as we can simulate from  $D$ .

1. Take  $R$  draws from the disturbance density  $D$ .
2. Add  $x^*'\beta$  to each disturbance draw.
3. Form a density forecast by fitting a density to the output from step 2.
4. Form a 95% interval forecast by sorting the output from step 2, and taking the left and right interval endpoints as the the .025% and .975% values, respectively.<sup>2</sup>

Again as  $R \rightarrow \infty$ , the algorithmic results become arbitrarily accurate.

### 4.1.4 Making the Forecasts Feasible

The approaches above are infeasible in that they assume known parameters. They can be made feasible by replacing unknown parameters with estimates. For example, the feasible version of the point prediction is  $x^*'\hat{\beta}$ . Similarly, to construct a feasible 95% interval forecast in the Gaussian case we can take  $x^*'\hat{\beta} \pm 1.96\hat{\sigma}$ , where  $\hat{\sigma}$  is the standard error of the regression (also earlier denoted  $s$ ).

---

<sup>2</sup>Note that, now that we have in general abandoned symmetry, the prescribed method no longer necessarily generates the shortest interval.

#### 4.1.5 Density Prediction for $D$ Non-Parametric

Now assume that we know nothing about distribution  $D$ , except that it has mean 0. In addition, now that we have introduced “feasible” forecasts, we will stay in that world.

1. Take  $R$  draws from the regression residual density (which is an approximation to the disturbance density) by assigning probability  $1/N$  to each regression residual and sampling with replacement.
2. Add  $x^* \hat{\beta}$  to each draw.
3. Form a density forecast by fitting a density to the output from step 2.
4. Form a 95% interval forecast by sorting the output from step 2, and taking the left and right interval endpoints as the .025% and .975% values, respectively.

As  $R \rightarrow \infty$  and  $N \rightarrow \infty$ , the algorithmic results become arbitrarily accurate.

#### 4.1.6 Density Forecasts for $D$ Nonparametric and Acknowledging Parameter Estimation Uncertainty

Thus far we have accounted only for disturbance uncertainty in our feasible density forecasts. Disturbance uncertainty reflects the fact that disturbance realizations are inherently unpredictable. There is simply nothing that we can do about disturbance uncertainty; it is present always and everywhere, even if we were somehow to know the DGP and its parameters.

We now consider **parameter estimation uncertainty**. The coefficients that we use to produce predictions are of course just *estimates*. That is, even if we somehow know the form of the DGP, we still have to estimate its parameters. Those estimates are subject to sampling variability, which makes

an additional contribution to prediction errors. The “feasible” approach to density forecasting sketched above still fails to acknowledge parameter estimation uncertainty, because it treats “plugged-in” parameter estimates as true values, ignoring the fact that they are only estimates and hence subject to sampling variability. Parameter estimation uncertainty is often ignored, as its contribution to overall forecast MSE can be shown to vanish unusually quickly as sample size grows (See EPC 1). But it impacts forecast uncertainty in small samples and hence should not be ignored in general.

1. Take  $R$  approximate disturbance draws by assigning probability  $1/N$  to each regression residual and sampling with replacement.
2. Take  $R$  draws from the large- $N$  sampling density of  $\hat{\beta}$ , namely

$$\hat{\beta}_{OLS} \sim N(\beta, \sigma^2(X'X)^{-1}),$$

as approximated by  $N(\hat{\beta}, \hat{\sigma}^2(X'X)^{-1})$ .

3. To each disturbance draw from 1 add the corresponding  $x^{*'}\hat{\beta}$  draw from 2.
4. Form a density forecast by fitting a density to the output from step 3.
5. Form a 95% interval forecast by sorting the output from step 3, and taking the left and right interval endpoints as the the .025% and .975% values, respectively.

As  $R \rightarrow \infty$  and  $N \rightarrow \infty$ , we get precisely correct results.

#### 4.1.7 Incorporating Heteroskedasticity

We will illustrate for the Gaussian case without parameter estimation uncertainty, using an approach that closely parallel’s White’s test for heteroskedasticity.

ticity. Recall the feasible density forecast,

$$y_i|x_i = x^* \sim N(x^{*\prime}\hat{\beta}, \hat{\sigma}^2).$$

Now we want to allow for the possibility that  $\hat{\sigma}^2$  varies with  $x_i$ .

1. Regress by OLS:  $y_i \rightarrow x_i$  and save the residuals  $e_i$ .
2. Regress  $e_i^2 \rightarrow x_i$ . Call the estimated coefficient vector  $\hat{\gamma}$ .
3. Form the density forecast as

$$y_i|x_i = x^* \sim N(x^{*\prime}\hat{\beta}, \hat{\sigma}^2(x^*)),$$

where  $\hat{\sigma}^2(x^*) = x^{*\prime}\hat{\gamma}$  is the fitted value from the regression in step 2 evaluated at  $x^*$ .

One could of course run regression 1 by weighted least squares (WLS) rather than OLS using the  $\hat{\sigma}^2(x^*)$  as weights, but the efficiency gains in estimating  $\beta$  are not likely to produce large additional improvements in calibration of density and interval forecasts. The key is to allow the disturbance variance to adapt to  $x^*$  when forming forecasts, quite apart from whether they are centered at  $x^{*\prime}\hat{\beta}_{OLS}$  or  $x^{*\prime}\hat{\beta}_{WLS}$ .

## 4.2 Wage Prediction Conditional on Education and Experience

### 4.2.1 The CPS Dataset

We will examine the CPS wage dataset, containing data on a large cross section of individuals on wages, education, experience, sex, race and union status. For a detailed description see Appendix E. For now we will use only wage, education and experience.

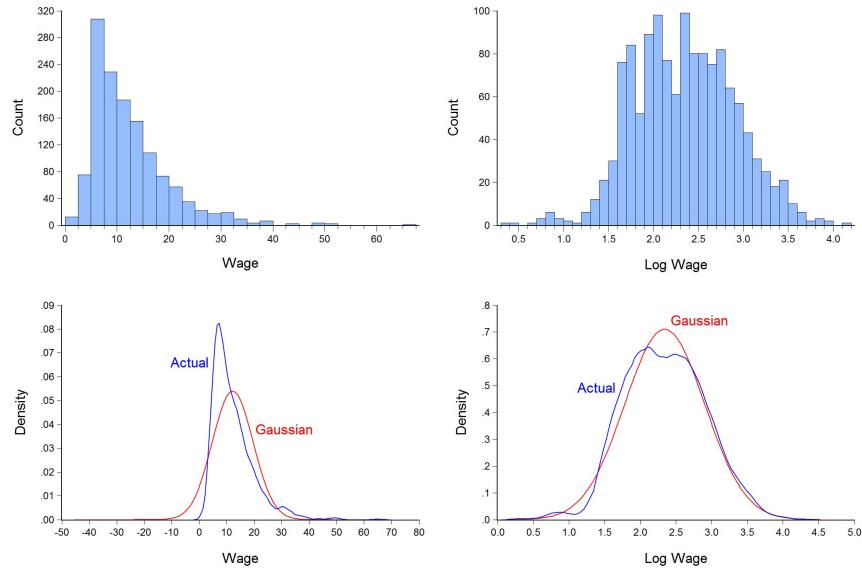


Figure 4.1: Distributions of Wages and Log Wages

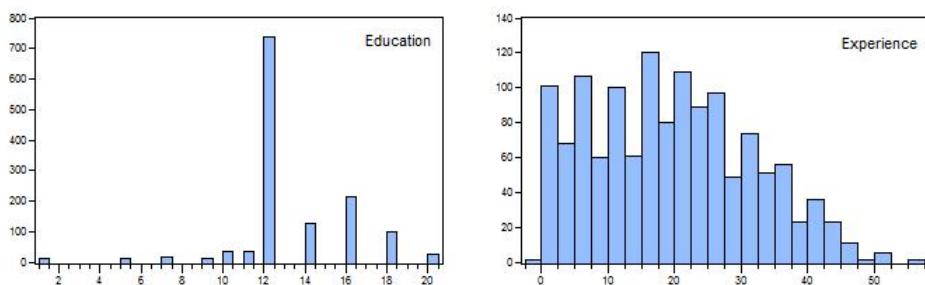


Figure 4.2: Histograms for Wage Covariates

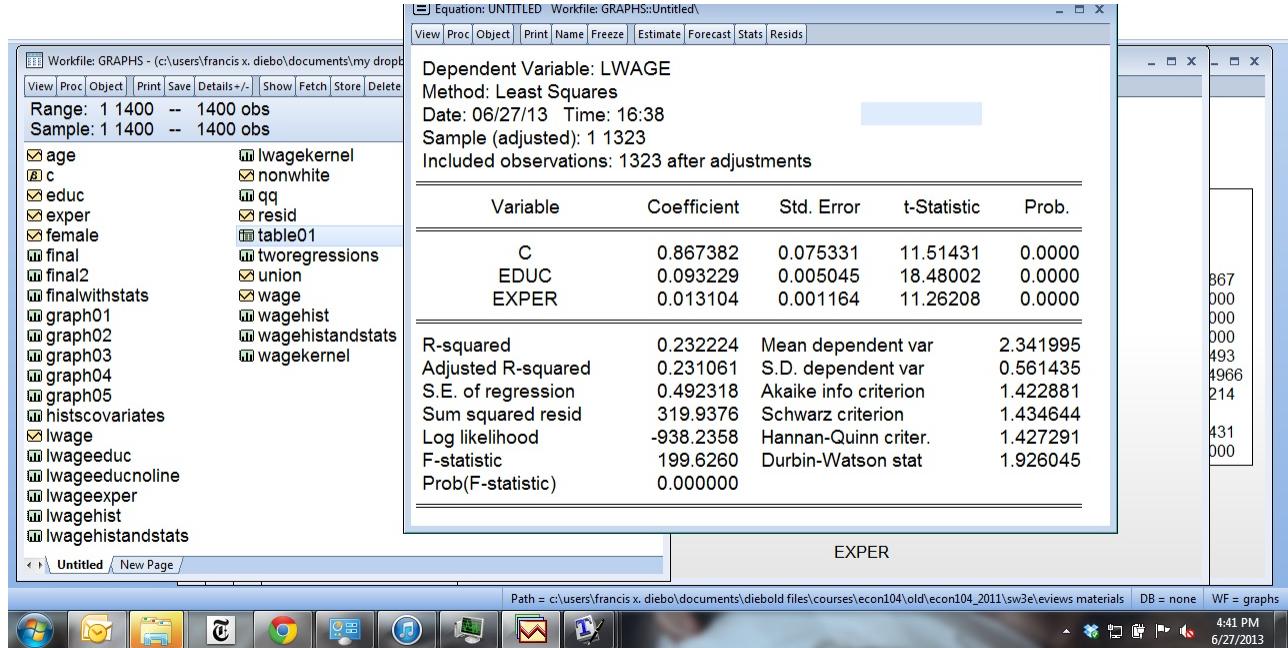


Figure 4.3: Wage Regression on Education and Experience

- Basic features of wage, education and experience data.

In Figures 4.1 and 4.2 we show histograms and statistics for potential determinants of wages. Education (EDUC) and experience (EXPER) are standard continuous variables, although we measure them only discretely (in years).

### 4.2.2 Regression

- Linear regression of log wage on predictors (education and experience).

Recall our basic wage regression,

$$LWAGE \rightarrow c, EDUC, EXPER,$$

shown in Figure 4.3. Both explanatory variables are highly significant, with expected signs. In table ?? consider a linear versus a quadratic model.

Even though the quadratic regression coefficients are statistically significant, we see only an extremely small improvement in adj.  $R^2$  and RMSE. We

also consider the histograms for the two models, in Figure ??.

We can see that the densities of residuals are almost identical, perhaps that those from the quadratic model are ever so slightly less skewed. Since we believe in the parsimony principle however, we will restrict ourselves to a linear model in the absence of overwhelming evidence in favor of a non-linear model. NOTE: There are many more nonlinear models to try besides quadratic! See section 4.3 for possible further extensions.

Throughout we will use the “best” estimated log wage model for feasible prediction of wage, for  $(\text{EDUC}, \text{EXPER}) = (10, 1)$  and  $(\text{EDUC}, \text{EXPER}) = (14, 20)$ . (NOTE: The model is for log wage, but the forecasts are for wage.)

### 4.2.3 Point Prediction by Exponentiating vs. Simulation

An obvious point forecast of  $WAGE$  can be obtained by exponentiating a forecast of  $LWAGE$ . But there are issues. In particular, if  $\ln y_{t+h,t}$  is an unbiased forecast of  $\ln y_{t+h}$ , then  $\exp(\ln y_{t+h,t})$  is a *biased* forecast of  $y_{t+h}$ .<sup>3</sup> More generally, if  $(f(y))_{t+h,t}$  is an unbiased forecast of  $(f(y))_{t+h}$ , then  $f^{-1}((f(y))_{t+h,t})$  is a biased forecast of  $y_{t+h}$ , for arbitrary nonlinear function  $f$ , because the expected value of a nonlinear function of a random variable does not equal the nonlinear function of the expected value, a result known as Jensen’s inequality.<sup>4</sup>

Various analytic “bias corrections” have been proposed, but they rely on strong and unnecessary assumptions. The modern approach is simulation-based. Using simulation, simply build up the density forecast of the object of interest (e.g.,  $WAGE$  rather than  $\ln WAGE$ ), the sample mean of which across simulations is consistent for the population conditional mean. The bias correction is done automatically!

---

<sup>3</sup>A forecast is unbiased if its mean error is zero. Other things equal, unbiasedness is desirable.

<sup>4</sup>As the predictive regression  $R^2 \rightarrow 1$ , however, the bias vanishes. Why?

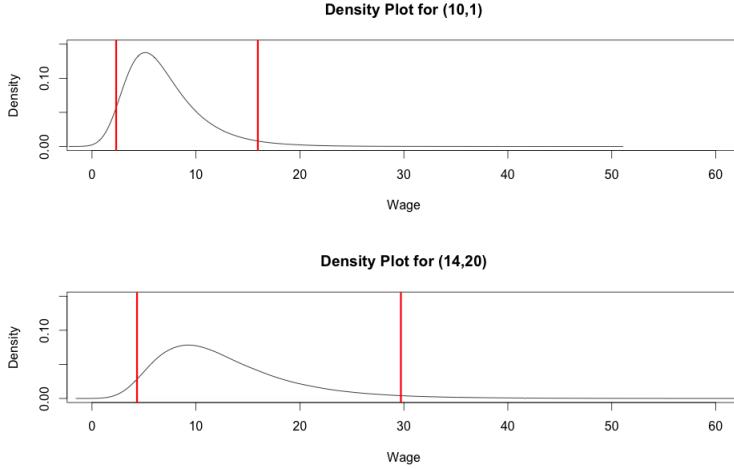


Figure 4.4: Predicted densities for wage under the assumption that residuals are homoskedastic and Gaussian, abstracting from parameter uncertainty. The model is in logs and then exponentiated. The top graph is estimated density of wage with 10 years of education, 1 year of experience. The bottom graph is the same for 14 years of education, 20 years of experience. The red vertical lines indicate 95% CI.

#### 4.2.4 Density Prediction for $D$ Gaussian

We now apply the methods from section 4.1.2 to the linear model. To operationalize that algorithm, we must first make an estimator of  $\sigma^2$  and  $\beta$ .  $\hat{\beta}$  is taken directly as the OLS regression coefficients, and  $\hat{\sigma}$  can be taken as the residual standard error. With those plug-in estimators found, we can follow the algorithm directly. Since we are in a Gaussian environment, recall we could find a 95% CI by taking  $x^*\beta \pm 1.96\hat{\sigma}$ . However, in more complex environments, we will have to take the CI directly from the simulated data, so we will do that here by sorting the sample draws and taking the left and right endpoints to be the .025% and .975% values, respectively. This yields the output from figure 4.4.

Two things are of particular note here. First is that, as expected, the density prediction for the individual with more education and experience has a much higher mean. Second is that the CI for individual 2 is much wider than that of individual 1, or similarly that the density prediction has much higher variance. This is perhaps surprising, since we were in a case with

homoskedasticity. In fact this is one of the costs of working with a log-linear model for wages:

$$\begin{aligned}\log(y) &= x'\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \\ y &= \exp(x'\beta) \exp(\varepsilon) \\ \Rightarrow \mathbb{V}[y] &= [\exp(x'\beta)]^2 \mathbb{V}[\exp(\varepsilon)]\end{aligned}$$

Thus even with homoskedasticity in logs, the variance of the level  $y_t$  will depend on  $x$ .

#### 4.2.5 Density Forecasts for $D$ Gaussian and Acknowledging Parameter Estimation Uncertainty

We are still in a sufficiently simple world that we may follow directly the algorithm above. A quick way to think about the algorithm of the previous section is the following: Since residuals are Gaussian,  $y$  is Gaussian. So to compute a density prediction of  $y$ , all we really need is to estimate its mean and covariance. The mean is given directly as the conditional mean from the model. For the covariance:

$$\begin{aligned}\mathbb{V}[y] &= \mathbb{V}[x'\beta + \varepsilon] \\ &= \mathbb{V}[x'\beta] + \mathbb{V}[\varepsilon]\end{aligned}$$

Since the previous section did not allow for parameter estimation uncertainty, the first term in that sum was zero. We will now accurately estimate that first term and include it in our density prediction. This idea is explored more in the EPC's.

Having Gaussian disturbances means that the distribution of  $\hat{\beta}$  is precisely

Gaussian, and as above we know its mean and covariance:  $\beta$  and  $\sigma^2(X'X)^{-1}$ . To operationalize this, we will make draws from  $N(\hat{\beta}, \hat{\sigma}^2(X'X)^{-1})$ . Concretely, our algorithm is the following:

1. Take  $R$  draws from the estimated disturbance density  $N(0, \hat{\sigma}^2)$ .
2. Take  $R$  draws of  $\beta$  from the estimated parameter sampling distribution  $N(\hat{\beta}, \hat{\sigma}^2(X'X)^{-1})$ .
3. Add the disturbance draw from step 1 to the draw of  $x^*\beta$ , where  $\beta$  is drawn as in step 2.
4. Exponentiate each draw to turn the draw of log wage into a draw for wage.
5. Form the density forecast by fitting a density to the output.
6. Form a 95% interval forecast by sorting the output, and taking the left and right interval endpoints as the .025% and .975% values, respectively.

Following this algorithm yields the output of figure 4.5. We see that these density forecasts are nearly identical to those without parameter uncertainty. This is to be expected once we consider the estimated covariance matrix of  $\beta$ , which we find has very small variance:

$$\mathbb{V}[\hat{\beta}] = \begin{bmatrix} 0.00567 & -0.000357 & -4.19e-05 \\ -0.000357 & 2.55e-05 & 1.22e-06 \\ -4.19e-05 & 1.22e-06 & 1.35e-06 \end{bmatrix}$$

#### 4.2.6 Density Forecasts for $D$ Gaussian, Acknowledging Parameter Estimation Uncertainty, and Allowing for Heteroskedasticity

For this section we find that we must work a little bit harder. There are two separate difficulties that are important to get correct: The first is an

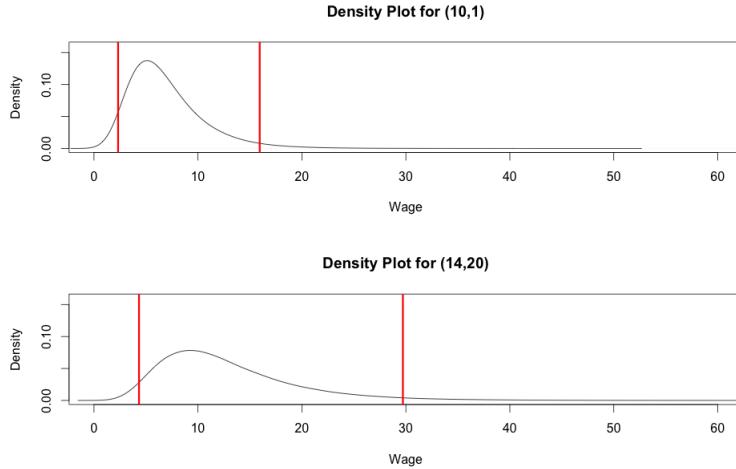


Figure 4.5: Predicted densities for wage under the assumption that residuals are homoskedastic and Gaussian. Here parameter uncertainty is accounted for in the density of wage. The model is in log wage and then exponentiated. The top graph is estimated density of wage with 10 years of education, 1 year of experience. The bottom graph is the same for 14 years of education, 20 years of experience. The red vertical lines indicate 95% CI.

appropriate drawing of the sampled residuals, the second is an appropriate drawing of the parameters.

First, the parameter estimation uncertainty. The covariance matrix we estimated above,  $\sigma^2(X'X)^{-1}$ , is no longer valid in the presence of heteroskedasticity. Rather:

$$\begin{aligned}\hat{\beta}_{OLS} &= (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\varepsilon \\ \Rightarrow \hat{\beta} &\sim N(\beta, (X'X)^{-1}X'\Omega X(X'X)^{-1})\end{aligned}\tag{4.2}$$

Under homoskedasticity,  $\Omega = \sigma^2 I$  and so this covariance matrix dramatically simplifies. This is no longer the case under heteroskedasticity. In this environment we will find the distinction to be of little numerical importance, but for other datasets it will be of dramatic importance.

Of course, in the presence of heteroskedasticity, we may prefer to instead conduct weighted least squares instead of OLS. Recall the WLS estimator is

$$\hat{\beta}_{WLS} = (X'\Sigma X)^{-1}X'\Sigma Y$$

Here  $\Sigma$  is any diagonal weighting matrix. A popular choice is of course  $\Omega^{-1}$ , as this choice is efficient, where this matrix can be estimated by a number of two-stage processes. The asymptotic covariance matrix of the WLS estimator is then

$$\begin{aligned}\hat{\beta}_{WLS} &= \beta + (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\varepsilon \\ \Rightarrow \hat{\beta}_{WLS} &\sim N(\beta, (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\Omega\Omega^{-1}X(X'\Omega^{-1})^{-1}) \\ \Rightarrow \hat{\beta}_{WLS} &\sim N(\beta, (X'\Omega^{-1}X)^{-1})\end{aligned}\quad (4.3)$$

Thus  $\hat{\beta}_{WLS}$  is simultaneously a better estimator than  $\hat{\beta}_{OLS}$  and with an easier covariance matrix to estimate. For this reason we will proceed using  $\hat{\beta}_{WLS}$ , and make draws from the above. To do this, we must select a specific two-stage process, as discussed above. We will discuss this in the course of the estimation of the residual density. This is done via the following algorithm:

1. Regress by OLS:  $y_i \rightarrow x_i$  and save the residuals.
2. Regress  $e_i^2 \rightarrow x_i$ . Call the estimated coefficient vector  $\hat{\gamma}$ .
3. Construct the vector of heteroskedasticities  $\hat{\sigma}^2(x_i) = x_i' \hat{\gamma}$ , and set  $\hat{\Omega} = diag(\hat{\sigma}^2(x_i))$ .
  - Use  $\hat{\Omega}$  to conduct WLS regression.
4. Take R draws of the residuals from  $N(0, \hat{\sigma}^2(x^*))$
5. Take R draws of  $\beta$  from  $N(\hat{\beta}, (X'\hat{\Omega}^{-1}X)^{-1})$
6. Add the disturbance draw from step 4 to the draw of  $x^*\beta$ , where  $\beta$  is drawn as in step 5.
7. Exponentiate each draw to turn the draw of log wage into a draw for wage.
8. Form the density forecast by fitting a density to the output.

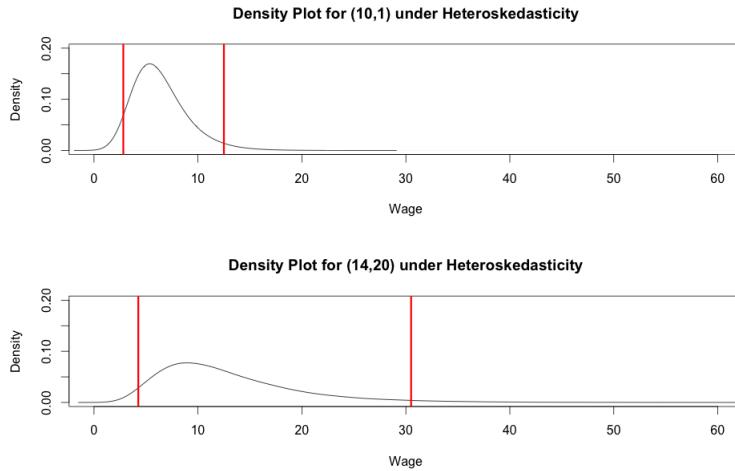


Figure 4.6: Predicted densities for wage under the assumption that residuals are heteroskedastic and Gaussian. Here parameter uncertainty is accounted for in the density of wage. The model is in log wage and then exponentiated. The top graph is estimated density of wage with 10 years of education, 1 year of experience. The bottom graph is the same for 14 years of education, 20 years of experience. The red vertical lines indicate 95% CI.

9. Form a 95% interval forecast by sorting the output, and taking the left and right interval endpoints as the .025% and .975% values, respectively.

Notice one could argue with our above procedure: We took residuals from the OLS regression to make the density prediction. There is certainly an argument to be made from re-taking residuals from the *WLS* regression and re-estimating the heteroskedasticity covariance matrix from there. However, the above will still be a consistent procedure (since the covariance matrix estimated is HAC-consistent), and since we have a surplus of observations we are unlikely to see a numerical difference between the two. The outputs from this procedure can be found in figure 4.6.

The CI interval for the lower wage, lower educated individual shrunk dramatically, while the CI for the (14, 20) individual did not change noticeably. This is explainable by the fact that the average number of years of education for our dataset is 13.1, and the average number of years of experience is 19.2. Thus the (14, 20) individual is close to the average. Since the form of heteroskedasticity is measured to be linear in this algorithm, and homoskedas-

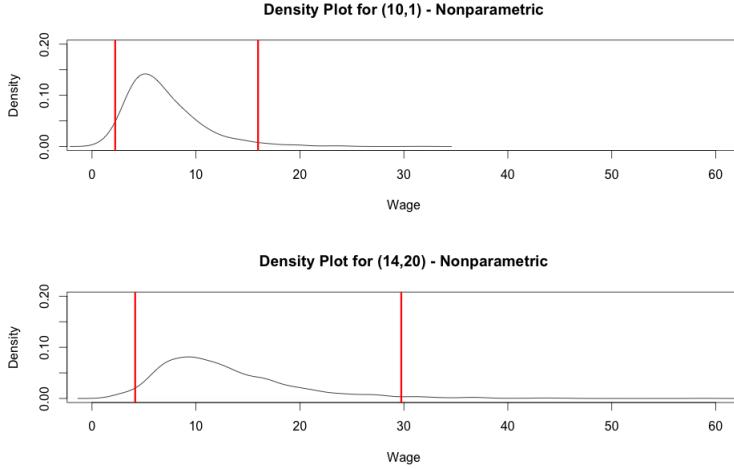


Figure 4.7: Predicted densities for wage under the assumption that residuals are homoskedastic, abstracting from parameter uncertainty. The density of residuals is now estimated nonparametrically. The model is in log wages and then exponentiated. The top graph is estimated density of wage with 10 years of education, 1 year of experience. The bottom graph is the same for 14 years of education, 20 years of experience. The red vertical lines indicate 95% CI.

ticity will set the variance of each individual to be approximately the mean variance of the dataset, it is expected that the mean individual will have approximately the same variance under homoskedasticity and heteroskedasticity.

#### 4.2.7 Density Prediction for $D$ Nonparametric

In this section we will make density predictions for our dataset dropping the assumption that disturbances are Gaussian. For now we will assume that we can estimate parameters with no uncertainty and that disturbances are homoskedastic. Here we may follow the exact algorithm of 4.1.5, with the added step that we exponentiate each draw to make a draw for wage from a draw for log wage. The yielded output can be found in 4.7.

Here we find that the nonparametric density estimates are quite similar to those found assuming  $D$  Gaussian. This suggests that our assumption of Gaussian disturbances was well-grounded. We can examine this further by

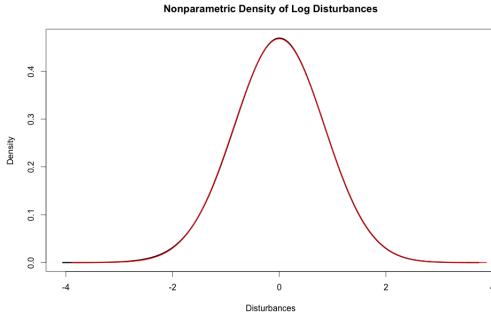


Figure 4.8: Nonparametric Density of Disturbances. Red overlaid line is Gaussian.

directly observing the nonparametric density of the disturbances to the log wages, seen in figure 4.8. These are some very Gaussian disturbances.

#### 4.2.8 Density Forecasts for $D$ Nonparametric and Acknowledging Parameter Estimation Uncertainty

Here we will blend the algorithms of the previous sections. Even though having non-Gaussian disturbances no longer assures that  $\hat{\beta}$  is precisely Gaussian, by CLT the normal distribution remains the large-N approximation. Thus, we may construct the algorithm exactly as in section 4.1.6, as before with the added step of exponentiating each log wage draw to get a draw for wage. This yields the output in figure 4.9.

Comparing the results from nonparametric estimation, and the results from just incorporating parameter estimation uncertainty, we should be unsurprised by this: nonparametric estimation told us that our Gaussian disturbances assumption was well-grounded, and our initial parameter uncertainty estimation results told us our parameter estimates were being measured quite accurately. Thus the output from this section very closely resembles that of our very first density predictions, with D Gaussian and no parameter uncertainty.

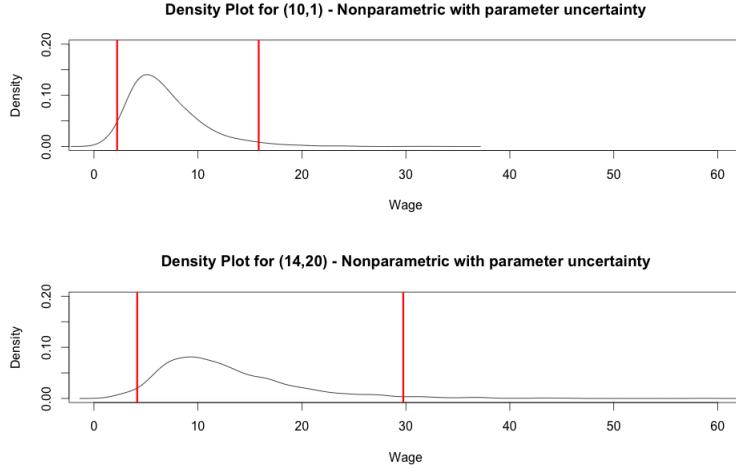


Figure 4.9: Predicted densities for wage under the assumption that residuals are homoskedastic. Here parameter uncertainty is accounted for in the density of wage, and the density of residuals is now estimated nonparametrically. The model is in log wage and then exponentiated. The top graph is estimated density of wage with 10 years of education, 1 year of experience. The bottom graph is the same for 14 years of education, 20 years of experience. The red vertical lines indicate 95% CI.

#### 4.2.9 Modeling Directly in Levels

Up until now our model has been

$$\log(y) = x'\beta + \varepsilon$$

From this model we construct density predictions for  $y$  by making draws of  $\log(y)$  and exponentiating. We now switch to the following model:

$$y = x'\beta + \varepsilon$$

We will now re-explore the above analysis in this context. The first thing we will notice is that density predictions under the assumption of Gaussian disturbances will generally perform quite poorly, because the disturbances to the level model are quite clearly non-Gaussian. See figure 4.10.

We therefore skip to a nonparametric density prediction, incorporating parameter uncertainty - although as before we will find that parameter un-

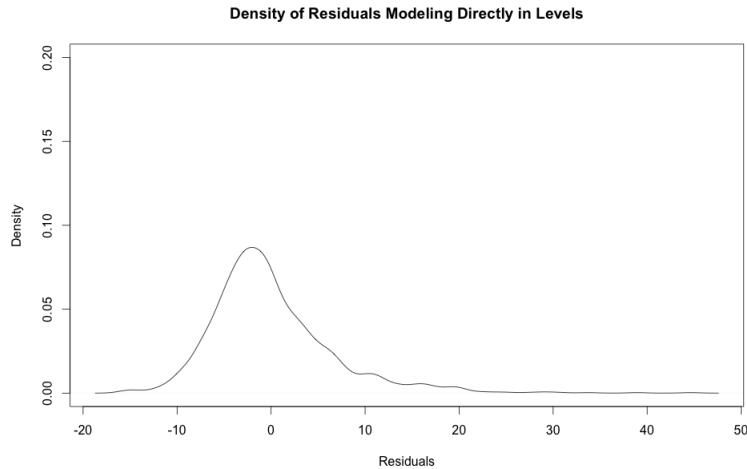


Figure 4.10: Non-Gaussian Residuals

certainty is quite small. The resulting output is in figure 4.11

We immediately notice a problem for individual 1: The 95% CI includes negative values for wage! This is inherently a problem of working directly in levels when modeling a variable for which only positive values make sense.

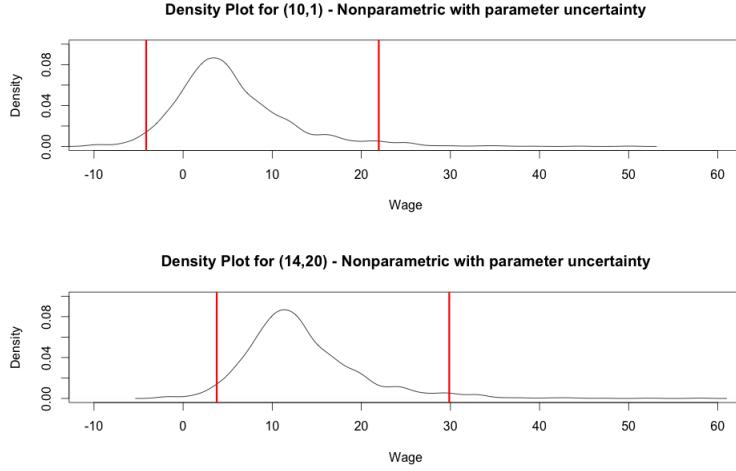


Figure 4.11: Predicted densities for wage under the assumption that residuals are homoskedastic. Here parameter uncertainty is accounted for in the density of wage, and the density of residuals is now estimated nonparametrically. The model is directly in wages. The top graph is estimated density of wage with 10 years of education, 1 year of experience. The bottom graph is the same for 14 years of education, 20 years of experience. The red vertical lines indicate 95% CI.

## 4.3 Non-Parametric Estimation of Conditional Mean Functions

### 4.3.1 Global Nonparametric Regression: Series

In the bivariate case we can think of the relationship as

$$y_t = g(x_t, \varepsilon_t),$$

or slightly less generally as

$$y_t = f(x_t) + \varepsilon_t.$$

First consider **Taylor series expansions** of  $f(x_t)$ . The linear (first-order) approximation is

$$f(x_t) \approx \beta_1 + \beta_2 x,$$

and the quadratic (second-order) approximation is

$$f(x_t) \approx \beta_1 + \beta_2 x_t + \beta_3 x_t^2.$$

In the multiple regression case, the Taylor approximations also involves interaction terms. Consider, for example,  $f(x_t, z_t)$ :

$$f(x_t, z_t) \approx \beta_1 + \beta_2 x_t + \beta_3 z_t + \beta_4 x_t^2 + \beta_5 z_t^2 + \beta_6 x_t z_t + \dots$$

Such **interaction effects** are also relevant in situations involving dummy variables. There we capture interactions by including products of dummies.<sup>5</sup>

Now consider **Fourier series expansions**. We have

$$f(x_t) \approx \beta_1 + \beta_2 \sin(x_t) + \beta_3 \cos(x_t) + \beta_4 \sin(2x_t) + \beta_5 \cos(2x_t) + \dots$$

One can also mix Taylor and Fourier approximations by regressing not only on powers and cross products (“Taylor terms”), but also on various sines and cosines (“Fourier terms”). Mixing may facilitate parsimony.

The ultimate point is that so-called “intrinsically non-linear” models are themselves linear when viewed from the series-expansion perspective. In principle, of course, an infinite number of series terms are required, but in practice nonlinearity is often quite gentle so that only a few series terms are required (e.g., quadratic).

### The Curse of Dimensionality

Let  $p$  be the adopted expansion order. Things quickly get out of hand as  $p$  grows, for fixed  $N$ .

---

<sup>5</sup>Notice that a product of dummies is one if and only if both individual dummies are one.

### Bandwidth Selection and the Bias-Variance Tradeoff

For fixed  $N$ , smaller  $p$  reduces variance but increases bias, larger  $p$  reduces bias but inflates variance.

Good things happen as  $p \rightarrow \infty$  while  $p/N \rightarrow 0$ .

$p$  can be chosen by any of the criteria introduced earlier.

#### 4.3.2 Local Nonparametric Regression: Nearest-Neighbor

Here we introduce the idea of local regression based on “nearest neighbors”. It is a leading example of a local smoother. The basic model is

$$y_t = g(x_t) + \varepsilon_t.$$

##### Unweighted Locally-Constant Regression

We want to fit (predict)  $y$  for an arbitrary  $x^*$ . We use the  $x$  variables in a neighborhood of  $x^*$ ,  $n(x^*)$ . In particular we use the  $P_T$  nearest neighbors.  $P_T$  can be chosen by CV. We find the  $P_T$  nearest neighbors using the Euclidean norm:

$$\lambda(x^*, x_{P_N}^*) = [\sum_{k=1}^K (x_{P_N k}^* - x_k^*)^2]^{\frac{1}{2}}.$$

The fitted value is then

$$\hat{y}(x^*) = \frac{1}{P_N} \sum_{j \in n(x^*)} y_j$$

This “nearest-neighbor” idea is not only simple, but tremendously important for prediction. If we want to predict  $y$  for an arbitrary  $x^*$ , it is natural to examine and average the  $y$ ’s that happened for close  $x$ ’s.

### Weighted Locally-Linear Regression

We will use the “tri-cube” neighborhood weight function:

$$v_i(x_i, x^*, x_{P_N}^*) = C \left( \frac{\lambda(x_i, x^*)}{\lambda(x^*, x_{P_N}^*)} \right),$$

where

$$C(u) = \begin{cases} (1 - u^3)^3 & \text{for } u < 1 \\ 0 & \text{otherwise} \end{cases}$$

We then obtain the fitted value by weighted linear regression:

$$\hat{y}^* = \hat{g}(x^*) = x^{*\prime} \hat{\beta}$$

where

$$\hat{\beta} = \operatorname{argmin} [\sum_{i=1}^N v_i(y_i - x'_i \beta)^2]$$

Good things happen as  $P_N \rightarrow \infty$  while  $P_N/N \rightarrow 0$ .

Figure 4.12: Locally Weighted Regression

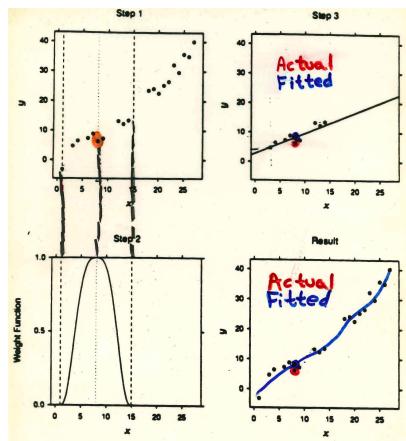
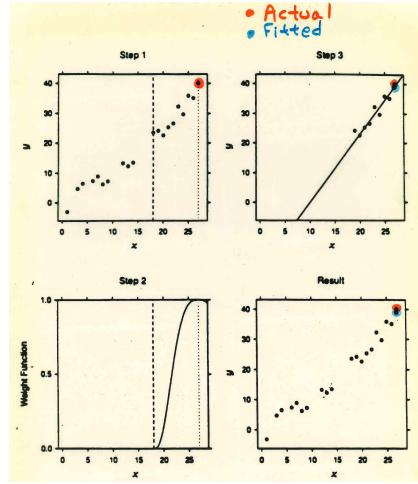


Figure 4.13: Locally Weighted Regression, Near the Edge



### “Robustness Iterations”

Consider the initial fit to be “robustness iteration 0”. Then define the robustness weight at iteration 1:

$$\rho_i^{(1)} = S \left( \frac{e_i^{(0)}}{6h} \right)$$

where

$$e_i^{(0)} = y_i - \hat{y}_i^{(0)}$$

$$h = \text{med} |e_i^{(0)}|$$

$$S(u) = \begin{cases} (1 - u^2)^2 & \text{for } |u| < 1 \\ 0 & \text{otherwise} \end{cases}$$

That is, we use bi-square robustness weighting, with bigger observations with bigger absolute residuals at iteration (0) downweighted progressively more, and observations with absolute residuals greater than six times the median absolute residual completely eliminated. We then obtain the fitted value by doubly-weighted linear regression:

$$\hat{y}_i^{*(1)} = \hat{g}^{(1)}(x^*) = x^{*\prime} \hat{\beta}^{(1)}$$

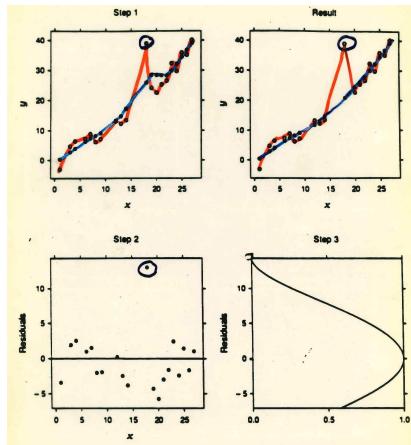
where

$$\hat{\beta}^{(1)} = \operatorname{argmin} [\sum_{i=1}^N \rho_i^{(1)} v_i (y_i - x_i' \beta)^2].$$

Then we continue iterating if desired.

We speak of “robust weighted locally-linear regression”. Extensions to locally-polynomial regression are immediate.

Figure 4.14: Locally Weighted Regression, Robustness Weighting for Outliers



### 4.3.3 Forecasting Perspectives

#### On Global vs. Local Smoothers for Forecasting

In cross-section environments, both global and local smoothers are useful for prediction. Local smoothers are perhaps more flexible and more popular in cross sections.  $x^*$  is usually interior to the observed  $x$ 's, so nearest-neighbor approaches feel natural.

In time-series environments both global and local smoothers can be useful for prediction. But there's a twist. Economic time-series data tend to trend, so that  $x^*$  can often be exterior to the observed  $x$ 's. That can create serious issues for local smoothers, as, for example, there may be no nearby “nearest neighbors”! Polynomial and Fourier global smoothers, in contrast, can be readily extrapolated for short-horizon out-of-sample forecasts. They have

issues of their own for long-horizon forecasts, however, as, for example, all polynomials diverge either to  $+\infty$  or  $-\infty$  when extrapolated far enough.

### Nearest Neighbors as a General Forecasting Method

Notice how natural and general NN is for forecasting. If we want to know what  $y$  is likely to go with  $x^*$  an obvious strategy is to look at the  $y$ 's that went with  $x$ 's nearest  $x^*$ . And the NN idea can be used to produce not just point forecasts (e.g., by fitting a constant to the  $y$ 's, but moreover to produce density forecasts (by fitting a distribution to the  $y$ 's). The NN idea is also equally relevant and useful in time series.

## 4.4 Wage Prediction, Continued

### 4.4.1 Point Wage Prediction

### 4.4.2 Density Wage Prediction

## 4.5 Exercises, Problems and Complements

1. Additional insight on parameter-estimation uncertainty.

Consider a simple homogeneous linear regression with zero-mean variables and Gaussian disturbances

$$y_t = \beta x_t + \varepsilon_t$$

$$\begin{aligned} & \text{ iid} \\ \varepsilon_t & \sim N(0, \sigma^2). \end{aligned}$$

It can be shown that

$$\text{var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{t=1}^T x_t^2},$$

and that  $\hat{\beta}$  and  $\varepsilon$  are independent. Now consider an operational point prediction of  $y$  given that  $x = x^*$ ,  $\hat{y} = \hat{\beta}x^*$ , and consider the variance of the corresponding forecast error. We have

$$\text{var}(e) = \text{var}(y - \hat{y}) = \text{var}((\beta x^* + \varepsilon) - \hat{\beta}x^*) = \sigma^2 + \frac{\sigma^2}{\sum_{i=1}^N x_i^2} x^{*2}.$$

In this expression, the first term accounts for the usual disturbance uncertainty, and the second accounts for parameter estimation uncertainty. Taken together, the results suggest an operational density forecast that accounts for parameter uncertainty,

$$y_i | x_i = x^* \sim N\left(\hat{\beta}x^*, \hat{\sigma}^2 + \frac{\hat{\sigma}^2}{\sum_{i=1}^N x_i^2} x^{*2}\right),$$

from which interval forecasts may be constructed as well. Note that when parameter uncertainty exists, the closer  $x^*$  is to the mean  $x(0)$ , the smaller is the prediction-error variance. Note also that as the sample size gets large,  $\sum_{i=1}^N x_i^2$  gets large as well, so the adjustment for parameter estimation uncertainty vanishes (in fact very quickly, like  $1/N$ ), and the density forecast collapses to the feasible Gaussian density forecast introduced in the text.

The ideas sketched here can be shown to carry over to more complicated situations (e.g., non-Gaussian,  $y$  and  $x$  don't necessarily have zero means, more than one regressor, etc.); it remains true that the closer is  $x$  to its mean, the tighter is the prediction interval.

## 2. Prediction intervals via quantile regression.

Granger, C.W.J., H. White, and M. Kamstra (1987), "Interval Forecasting: An Analysis Based Upon ARCH – Quantile Estimators," *Journal of Econometrics*. White (1990) allows for nonlinear conditional quantile regression via neural nets.

### 3. In-sample vs. out-of-sample prediction.

In cross sections all prediction has an “in-sample” flavor insofar as the  $X^*$  for which we want to forecast  $y$  is typically interior to the historical  $X$ . In time series, in contrast, future times are by definition exterior to past times.

### 4. Model uncertainty.

We have thus far emphasized disturbance uncertainty and parameter estimation uncertainty (which is due in part to data uncertainty, which in turn has several components).

A third source of prediction error is **model uncertainty**. All our models are intentional simplifications, and the fact is that different models produce different forecasts. Despite our best intentions, and our use of powerful tools such as information criteria, we never know the DGP, and surely any model that we use is *not* the DGP.

### 5. “Data-rich” environments.

“Big data.” “Wide data,” for example, corresponds to  $K$  large relative to  $T$ . In extreme cases we might even have  $K$  much larger than  $T$ . How to get a sample covariance matrix for the variables in  $X$ ? How to run a regression? One way or another, we need to recover degrees of freedom, so dimensionality reduction is key, which leads to notions of variable selection and “sparsity”, or shrinkage and “regularization”.

### 6. Neural Networks

**Neural networks** amount to a particular non-linear functional form associated with repeatedly running linear combinations of inputs through non-linear ”squashing” functions. The 0-1 squashing function is useful in classification, and the logistic function is useful for regression. The neural net literature is full of biological jargon, which serves to obfuscate

rather than clarify. We speak, for example, of a “single-output feedforward neural network with  $n$  inputs and 1 hidden layer with  $q$  neurons.” But the idea is simple. If the output is  $y$  and the inputs are  $x$ ’s, we write

$$y_t = \Phi(\beta_0 + \sum_{i=1}^q \beta_i h_{it}),$$

where

$$h_{it} = \Psi(\gamma_{i0} + \sum_{j=1}^n \gamma_{ij} x_{jt}), i = 1, \dots, q$$

are the “neurons” (“hidden units”), and the “activation functions”  $\Psi$  and  $\Phi$  are arbitrary, except that  $\Psi$  (the squashing function) is generally restricted to be bounded. (Commonly  $\Phi(x) = x$ .) Assembling it all, we write

$$y_t = \Phi \left( \beta_0 + \sum_{i=1}^q \beta_i \Psi \left( \gamma_{i0} + \sum_{j=1}^n \gamma_{ij} x_{jt} \right) \right) = f(x_t; \theta),$$

which makes clear that a neural net is just a particular non-linear functional form for a regression model.

- 7. Trees
- 8. Kernel Regression
- 9. Regression Splines

Polynomial are global. Unattractive in that the fit at the end is influenced by the data at beginning (for example).

Move to piecewise cubic (say). But it’s discontinuous at the join point(s) (“knots”).

Move to continuous piecewise cubic; i.e., force continuity at the knots. But it might have an unreasonable kink.

Move to cubic spline. Forces continuity *and* continuity of first and second derivatives at the knots. Nice! A polynomial of degree  $p$  spline has continuous  $d^{\text{th}}$ -order derivatives,  $d = 0, \dots, p - 1$ . So, for example, linear spline is piecewise linear, continuous but not differentiable at the knots.

- Linear Splines
- Constructing Cubic Splines
- Natural Cubic Splines

Extrapolates linearly beyond the left and right boundary knots. This adds constraints (two at each end), recovering degrees of freedom and hence allowing for more knots.

A cubic spline with  $K$  knots uses  $K + 4$  degrees of freedom. A natural spline with  $K$  knots uses  $K$  degrees of freedom.

- Knot Placement

You'd like more knots in rough areas of the function being estimated, but of course you don't know where those areas are, so it's tricky.

Smoothing splines avoid that issue.

## 10. Smoothing Splines

$$\min_{\{f \in F\}} \sum_{t=1}^T (y_t - f(x_t))^2 + \lambda \int f''(z)^2 dz$$

HP Trend does that:

$$\min_{\{s_t\}_{t=1}^T} \sum_{t=1}^T (y_t - s_t)^2 + \lambda \sum_{t=2}^{T-1} ((s_{t+1} - s_t) - (s_t - s_{t-1}))^2$$

The smoothing spline is a natural cubic spline. It has a knot at each unique  $x$  value, but smoothness is imposed via  $\lambda$ . No need to choose knot locations; instead just choose a single  $\lambda$ . Can be done by CV.

There is an analytic formula giving effective degrees of freedom, so we can specify d.f. rather than  $\lambda$ .

## 4.6 Notes

“LOWESS”



# Part III

## Time Series: A Components Perspective



# Chapter 5

## Trend and Seasonality

Consider a time-series situation:

$$y_t = x'_t \beta + \varepsilon_t, \quad t = 1, \dots, T$$

$$\varepsilon_t \sim \text{iid } N(0, \sigma^2).$$

Note that the disturbance density is assumed Gaussian for simplicity (for example in calculating density forecasts), but we could of course relax that assumption just as we did for cross-section forecasts.

### 5.1 The Forecasting the Right-Hand-Side Variables (FRV) Problem

In the future period of interest,  $T + h$ , it must be true that

$$y_{T+h} = x'_{T+h} \beta + \varepsilon_{T+h}.$$

Under quadratic loss the conditional mean forecast is optimal, and we immediately have

$$E(y_{T+h}|x_{T+h}) = x'_{T+h} \beta.$$

Suppose for the moment that we know the regression parameters. Forming the conditional expectation still requires knowing  $x_{T+h}$ , which we don't, so it

seems that we're stuck.

We call this the “**forecasting-the-right-hand-side-variables (FRV) problem**.” It *is* a problem, but it’s not nearly as damaging as you might fear.

- We can abandon time series and only work in cross sections, where the FRV problem doesn’t exist! But of course that’s throwing out the baby with the bathwater and hardly a useful or serious prescription.
- We can move to scenario forecasts. **Time-series scenario forecasts** (also called **stress tests**, or **contingency analyses**), help us answer the “what if” questions that often arise. As with cross-section prediction, there is no FRV problem, and for precisely the same reason. For any given “scenario”  $x^*$ , we immediately have

$$E(y_{T+h}|x_{T+h} = x^*) = x^{*'}_{T+h}\beta.$$

However, notwithstanding the occasional usefulness of scenario analyses, we generally don’t want to make forecasts of  $y$  conditional upon assumptions about  $x$ ; rather, we just simply want the best possible forecast of  $y$ .

- We can work with models involving lagged rather than current  $x$ , that is, models that relate  $y_t$  to  $x_{t-h}$  rather than relating  $y_t$  to  $x_t$ . This sounds ad hoc, but it’s actually not, and we will have much more to say about it later.
- We can work with models for which we actually *do* know how to forecast  $x$ . In some important cases, the FRV problem doesn’t arise at all, because the regressors are perfectly deterministic, so we know *exactly* what they’ll be at any future time. The trend and seasonality models that we now discuss are leading examples.

## 5.2 Deterministic Trend

Time series fluctuate over time, and we often mentally allocate those fluctuations to unobserved underlying components, such as trends, seasonals, and cycles. In this section we focus on **trends**.<sup>1</sup> More precisely, in our general unobserved-components model,

$$y_t = T_t + S_t + C_t + \varepsilon_t,$$

we now include only trend and noise,

$$y_t = T_t + \varepsilon_t.$$

Trend is obviously pervasive. It involves slow, long-run, evolution in the variables that we want to model and forecast. In business, finance, and economics, trend is produced by slowly evolving preferences, technologies, institutions, and demographics.

We will study both **deterministic trend**, evolving in a perfectly predictable way, and **stochastic trend**, evolving in an approximately predictable way. We treat the deterministic case here, and we treat the stochastic case later in Chapter 5.3.

### 5.2.1 Trend Models

Sometimes series increase or decrease like a straight line. That is, sometimes a simple linear function of time,

$$T_t = \beta_0 + \beta_1 TIME_t,$$

provides a good description of the trend, in which case we speak of **linear trend**. We construct the variable *TIME* artificially; it is called a “time

---

<sup>1</sup>Later we'll define and study **seasonals** and **cycles**. Not all components need be present in all observed series.

trend” or “**time dummy**.” Time equals 1 in the first period of the sample, 2 in the second period, and so on. Thus, for a sample of size  $T$ ,  $TIME = (1, 2, 3, \dots, T - 1, T)$ ; put differently,  $TIME_t = t$ .  $\beta_0$  is the **intercept**; it’s the value of the trend at time  $t = 0$ .  $\beta_1$  is the **slope**; it’s positive if the trend is increasing and negative if the trend is decreasing. The larger the absolute value of  $\beta_1$ , the steeper the trend’s slope. In business, finance, and economics, linear trends are typically (but not necessarily) increasing, corresponding to growth.

Sometimes trend appears nonlinear, or curved, as for example when a variable increases at an increasing or decreasing rate. Ultimately, we don’t require that trends be linear, only that they be smooth. **Quadratic trend** models can potentially capture nonlinearities. Such trends are simply quadratic, as opposed to linear, functions of time,

$$T_t = \beta_0 + \beta_1 TIME_t + \beta_2 TIME_t^2.$$

Linear trend emerges as a special (and potentially restrictive) case when  $\beta_2 = 0$ .<sup>2</sup>

A variety of different nonlinear quadratic trend shapes are possible, depending on the signs and sizes of the coefficients. In particular, if  $\beta_1 > 0$  and  $\beta_2 > 0$ , the trend is monotonically, but nonlinearly, increasing. Conversely, if  $\beta_1 < 0$  and  $\beta_2 < 0$ , the trend is monotonically decreasing. If  $\beta_1 < 0$  and  $\beta_2 > 0$  the trend is U-shaped, and if  $\beta_1 > 0$  and  $\beta_2 < 0$  the trend has an inverted U shape. See Figure 5.1. Keep in mind that quadratic trends are used to provide local approximations; one rarely has a U-shaped trend, for example. Instead, all of the data may lie on one or the other side of the “U.”

Other types of nonlinear trend are sometimes appropriate. Sometimes, in particular, trend is nonlinear in levels but linear in logarithms. That’s

---

<sup>2</sup>Higher-order polynomial trends are sometimes entertained, but it’s important to use low-order polynomials to maintain smoothness.

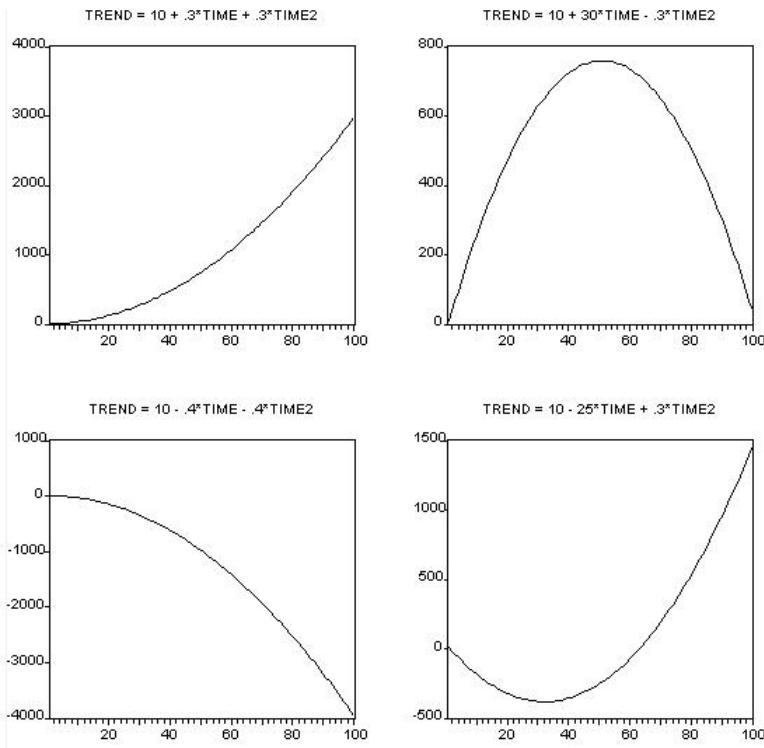


Figure 5.1: Various Shapes of Quadratic Trends

called **exponential trend**, or **log-linear trend**, and is very common in business, finance and economics.<sup>3</sup> It arises because economic variables often display roughly constant growth rates (e.g., three percent per year). If trend is characterized by constant growth at rate  $\beta_1$ , then we can write

$$T_t = \beta_0 e^{\beta_1 \text{TIME}_t}. \quad (5.1)$$

The trend is a nonlinear (exponential) function of time in levels, but in logarithms we have

$$\ln(T_t) = \ln(\beta_0) + \beta_1 \text{TIME}_t.$$

Thus,  $\ln(T_t)$  is a linear function of time. As with quadratic trend, depending on the signs and sizes of the parameter values, exponential trend can achieve a variety of patterns, increasing or decreasing at an increasing or decreasing rate. See Figure 5.2.

<sup>3</sup>Throughout this book, logarithms are *natural* (base e) logarithms.

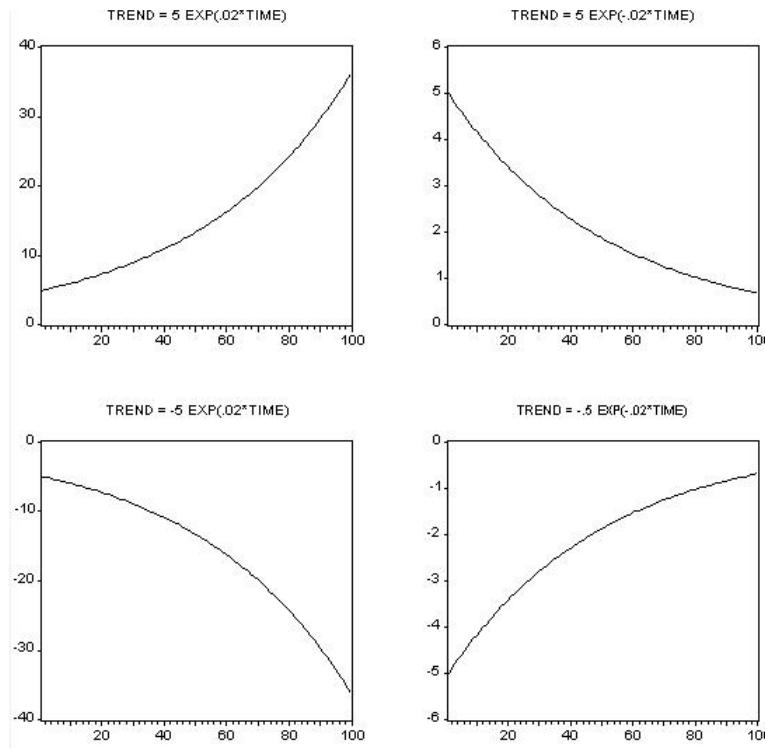


Figure 5.2: Various Shapes of Exponential Trends

It's important to note that, although qualitatively similar trend shapes can be achieved with quadratic and exponential trend, there are subtle differences between them. The nonlinear trends in some series are well approximated by quadratic trend, while the trends in other series are better approximated by exponential trend. Neither is necessarily “better” in general; rather, they're simply different, and which is better in any particular situation is ultimately an empirical matter.

### 5.2.2 Trend Estimation

Before we can estimate trend models, we need to create and store on the computer variables such as *TIME* and its square. Fortunately we don't have to type the trend values (1, 2, 3, 4, ...) in by hand; rather, in most software packages a command exists to create *TIME* automatically, after which we can immediately compute derived variables such as *TIME*<sup>2</sup>. Because, for

example,  $TIME = 1, 2, \dots, T$ ,  $TIME^2 = 1, 4, \dots, T^2$ .

For the most part we fit our various trend models to data on a time series  $y$  using **ordinary least-squares regression**. In the linear and quadratic trend cases, the regressions are just simple OLS regressions. In an obvious notation, we run

$$y \rightarrow c, TIME$$

and

$$y \rightarrow c, TIME, TIME^2,$$

respectively, where  $c$  denotes inclusion of an intercept.

The exponential trend, in contrast, is a bit more nuanced. We can estimate it in two ways. First, because the nonlinear exponential trend is nevertheless linear in logs, we can estimate it by regressing  $\ln y$  on an intercept and  $TIME$ ,

$$\ln y \rightarrow c, TIME.$$

Note that  $c$  provides an estimate of  $\ln \beta_0$  in equation (5.1) and so must be exponentiated to obtain an estimate of  $\beta_0$ . Similarly the fitted values from this regression are the fitted values of  $\ln y$ , so they must be exponentiated to get the fitted values of  $y$ .

Alternatively, we can proceed directly from the exponential representation and let the computer use numerical algorithms to find<sup>4</sup>

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{t=1}^T [y_t - \beta_0 e^{\beta_1 TIME_t}]^2.$$

This is called **nonlinear least squares**, or NLS.

NLS can be used to perform least-squares estimation for any model, including linear models, but in the linear case it's more sensible simply to use

---

<sup>4</sup>“Argmin” just means “the argument that minimizes.”

OLS. Some intrinsically nonlinear forecasting models can't be estimated using OLS, however, but they can be estimated using nonlinear least squares. We resort to nonlinear least squares in such cases.<sup>5</sup> (We will encounter several in this book.) Even for models like exponential trend, which as we have seen can be transformed to linearity, estimation in levels by NLS is useful, because statistics like AIC and SIC can then be directly compared to those of linear and quadratic trend models.

### 5.2.3 Forecasting Trends

Suppose we're presently at time  $T$ , and we have a sample of historical data,  $\{y_1, y_2, \dots, y_T\}$ . We want to use a trend model to forecast the  $h$ -step-ahead value of  $y$ . For illustrative purposes, we'll work with a linear trend, but the procedures are identical for quadratic and exponential trends.

First consider point forecasts. The linear trend model, which holds for any time  $t$ , is

$$y_t = \beta_0 + \beta_1 TIME_t + \varepsilon_t.$$

In particular, at time  $T + h$ , the future time of interest,

$$y_{T+h} = \beta_0 + \beta_1 TIME_{T+h} + \varepsilon_{T+h}.$$

Two future values of series appear on the right side of the equation,  $TIME_{T+h}$  and  $\varepsilon_{T+h}$ . If  $TIME_{T+h}$  and  $\varepsilon_{T+h}$  were known at time  $T$ , we could immediately crank out the forecast. In fact,  $TIME_{T+h}$  is known at time  $T$ , because the artificially-constructed time variable is perfectly predictable; specifically,  $TIME_{T+h} = T + h$ . Unfortunately  $\varepsilon_{T+h}$  is not known at time  $T$ , so we replace it with an optimal forecast of  $\varepsilon_{T+h}$  constructed using information only up

---

<sup>5</sup>When we estimate by NLS, we use a computer to find the minimum of the sum of squared residual function directly, using numerical methods, by literally trying many (perhaps hundreds or even thousands) of different  $(\beta_0, \beta_1)$  values until those that minimize the sum of squared residuals are found. This is not only more laborious (and hence slow), but also less numerically reliable, as, for example, one may arrive at a minimum that is local but not global.

to time  $T$ .<sup>6</sup> Under the assumption that  $\varepsilon$  is simply independent zero-mean random noise, the optimal forecast of  $\varepsilon_{T+h}$  for any future period is 0, yielding the point forecast,<sup>7</sup>

$$y_{T+h,T} = \beta_0 + \beta_1 \text{TIME}_{T+h}.$$

The subscript “ $T + h, T$ ” on the forecast reminds us that the forecast is for time  $T + h$  and is made at time  $T$ . Note that the point forecast formula at which we arrived is not of practical use, because it assumes known values of the trend parameters  $\beta_0$  and  $\beta_1$ . But it’s a simple matter to make it operational – we just replace unknown parameters with their least squares estimates, yielding

$$\hat{y}_{T+h,T} = \hat{\beta}_0 + \hat{\beta}_1 \text{TIME}_{T+h}.$$

Now consider density forecasts under normality and ignoring parameter estimation uncertainty. We immediately have the density forecast,  $N(y_{T+h,T}, \sigma^2)$ , where  $\sigma$  is the standard deviation of the disturbance in the trend regression. To make this operational, we use the density forecast  $N(\hat{y}_{T+h,T}, \hat{\sigma}^2)$ , where  $\hat{\sigma}^2$  is the square of the standard error of the regression. Armed with the density forecast, we can construct any desired interval forecast. For example, the 95% interval forecast ignoring parameter estimation uncertainty is  $y_{T+h,T} \pm 1.96\sigma$ , where  $\sigma$  is the standard deviation of the disturbance in the trend regression. To make this operational, we use  $\hat{y}_{T+h,T} \pm 1.96\hat{\sigma}$ , where  $\hat{\sigma}$  is the standard error of the regression.

We can use the simulation-based methods of Chapter 4 to dispense with the normality assumption and/or account for parameter-estimation uncertainty.

---

<sup>6</sup>More formally, we say that we’re “projecting  $\varepsilon_{T+h}$  on the time- $T$  information set.”

<sup>7</sup>“Independent zero-mean random noise” is just a fancy way of saying that the regression disturbances satisfy the usual assumptions – they are identically and independently distributed.

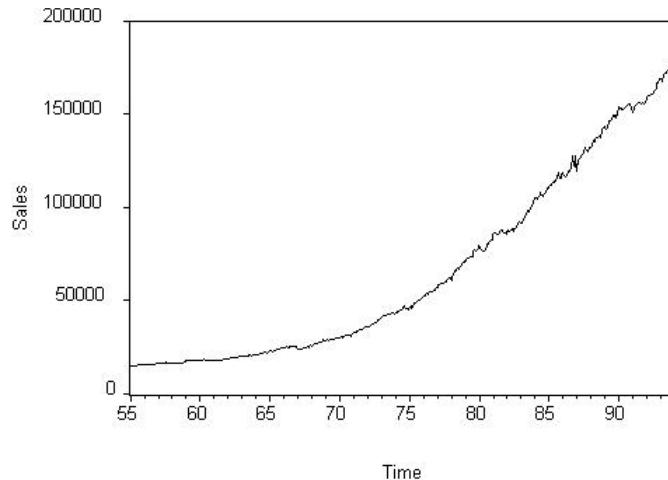


Figure 5.3: Retail Sales

#### 5.2.4 Forecasting Retail Sales

We'll illustrate trend modeling with an application to forecasting U.S. current-dollar retail sales. The data are monthly from 1955.01 through 1994.12 and have been seasonally adjusted.<sup>8</sup> We'll use the period 1955.01-1993.12 to estimate our forecasting models, and we'll use the "holdout sample" 1994.01-1994.12 to examine their out-of-sample forecasting performance.

In Figure 5.3 we provide a time series plot of the retail sales data, which display a clear nonlinear trend and not much else. Cycles are probably present but are not easily visible, because they account for a comparatively minor share of the series' variation.

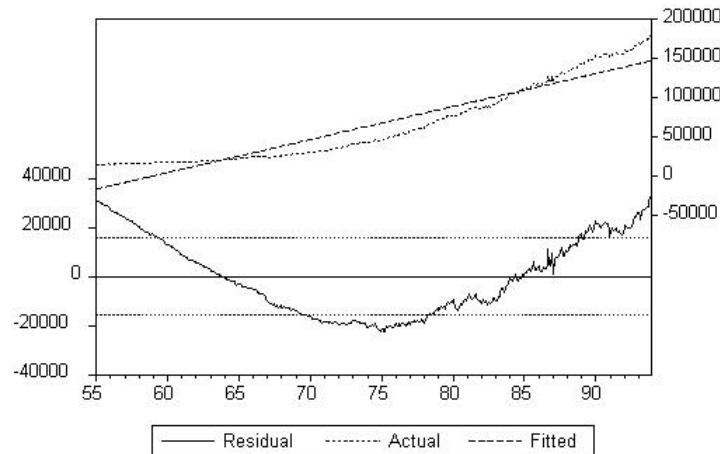
In Table 5.4a we show the results of fitting a linear trend model by regressing retail sales on a constant and a linear time trend. The trend appears highly significant as judged by the  $p$ -value of the  $t$  statistic on the time trend, and the regression's  $R^2$  is high. Moreover, the Durbin-Watson statistic indicates that the disturbances are positively serially correlated, so that the disturbance at any time  $t$  is positively correlated with the disturbance at time  $t - 1$ . In later chapters we'll show how to model such residual serial

---

<sup>8</sup>When we say that the data have been "seasonally adjusted," we simply mean that they have been smoothed in a way that eliminates seasonal variation. We'll discuss seasonality in detail in Section 5.3.

Dependent Variable is RTRR				
Sample: 1955:01 1993:12				
Included observations: 468				
Variable	Coefficient	Std. Error	T-Statistic	Prob.
C	-16391.25	1469.177	-11.15676	0.0000
TIME	349.7731	5.428670	64.43073	0.0000
R-squared	0.899076		Mean dependent var	65630.56
Adjusted R-squared	0.898859		S.D. dependent var	49889.26
S.E. of regression	15866.12		Akaike info criterion	19.34815
Sum squared resid	1.17E+11		Schwarz criterion	19.36587
Log likelihood	-5189.529		F-statistic	4151.319
Durbin-Watson stat	0.004682		Prob(F-statistic)	0.000000

(a) Retail Sales: Linear Trend Regression



(b) Retail Sales: Linear Trend Residual Plot

Figure 5.4: Retail Sales: Linear Trend

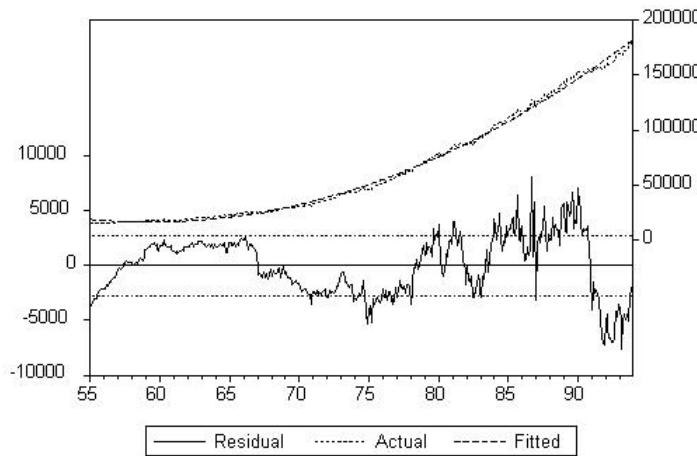
correlation and exploit it for forecasting purposes, but for now we'll ignore it and focus only on the trend.<sup>9</sup>

The residual plot in Figure 5.4b makes clear what's happening. The linear trend is simply inadequate, because the actual trend is nonlinear. That's one key reason why the residuals are so highly serially correlated – first the data are all above the linear trend, then below, and then above. Along with the residuals, we plot plus-or-minus one standard error of the regression, for visual reference.

<sup>9</sup>Such **residual serial correlation** may, however, render the standard errors of estimated coefficients (and the associated  $t$  statistics) untrustworthy, and robust standard errors (e.g., Newey-West) can be used. In addition, *AIC* and *SIC* remain valid.

Dependent Variable is RTRR				
Sample: 1955:01 1993:12				
Included observations: 468				
Variable	Coefficient	Std. Error	T-Statistic	Prob.
C	18708.70	379.9566	49.23905	0.0000
TIME	-98.31130	3.741388	-26.27669	0.0000
TIME2	0.955404	0.007725	123.6754	0.0000
R-squared	0.997022		Mean dependent var	65630.56
Adjusted R-squared	0.997010		S.D. dependent var	49889.26
S.E. of regression	2728.205		Akaike info criterion	15.82919
Sum squared resid	3.46E+09		Schwarz criterion	15.85578
Log likelihood	-4365.093		F-statistic	77848.80
Durbin-Watson stat	0.151089		Prob(F-statistic)	0.000000

(a) Retail Sales: Quadratic Trend Regression



(b) Retail Sales: Quadratic Trend Residual Plot

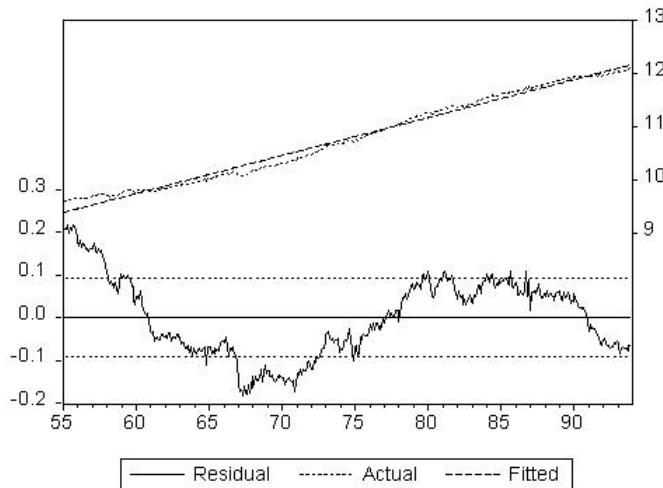
Figure 5.5: Retail Sales: Quadratic Trend

Table 5.5a presents the results of fitting a quadratic trend model. Both the linear and quadratic terms appear highly significant.  $R^2$  is now almost 1. Figure 5.5b shows the residual plot, which now looks very nice, as the fitted nonlinear trend tracks the evolution of retail sales well. The residuals still display persistent dynamics (indicated as well by the still-low Durbin-Watson statistic) but there's little scope for explaining such dynamics with trend, because they're related to the business cycle, not the growth trend.

Now let's estimate a different type of nonlinear trend model, the exponential trend. First we'll do it by OLS regression of the log of retail sales on a constant and linear time trend variable. We show the estimation results and

Dependent Variable is LRTRR				
Sample: 1955:01 1993:12				
Included observations: 468				
Variable	Coefficient	Std Error	T-Statistic	Prob.
C	9.389975	0.008508	1103.684	0.0000
TIME	0.005931	3.14E-05	188.6541	0.0000
R-squared	0.987076		Mean dependent var	10.78072
Adjusted R-squared	0.987048		S.D. dependent var	0.807325
S.E. of regression	0.091879		Akaike info criterion	-4.770302
Sum squared resid	3.933853		Schwarz criterion	-4.752573
Log likelihood	454.1874		F-statistic	35590.36
Durbin-Watson stat	0.019949		Prob(F-statistic)	0.000000

(a) Retail Sales: Log Linear Trend Regression



(b) Retail Sales: Log Linear Trend Residual Plot

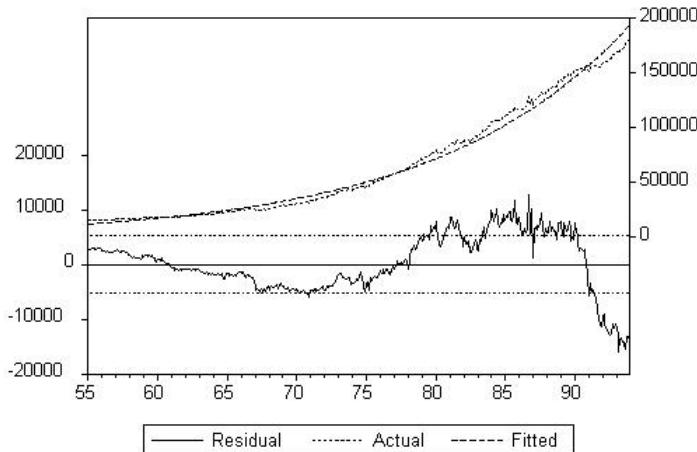
Figure 5.6: Retail Sales: Log Linear Trend

residual plot in Table 5.6a and Figure 5.6b. As with the quadratic nonlinear trend, the exponential nonlinear trend model seems to fit well, apart from the low Durbin-Watson statistic.

In sharp contrast to the results of fitting a linear trend to retail sales, which were poor, the results of fitting a linear trend to the *log* of retail sales seem much improved. But it's hard to compare the log-linear trend model to the linear and quadratic models because they're in levels, not logs, which renders diagnostic statistics like  $R^2$  and the standard error of the regression incomparable. One way around this problem is to estimate the exponential trend model directly in levels, using nonlinear least squares. In Table 5.7a

Dependent Variable is RTRR				
Sample: 1955:01 1993:12				
Included observations: 468				
Convergence achieved after 1 iterations				
RTRR=C(1)*EXP(C(2)*TIME)				
	Coefficient	Std. Error	T-Statistic	Prob.
C(1)	11967.80	177.9598	67.25003	0.0000
C(2)	0.005944	3.77E-05	157.7469	0.0000
R-squared	0.988796		Mean dependent var	65630.56
Adjusted R-squared	0.988772		S.D. dependent var	49889.26
S.E. of regression	5286.406		Akaike info criterion	17.15005
Sum squared resid	1.30E+10		Schwarz criterion	17.16778
Log likelihood	-4675.175		F-statistic	41126.02
Durbin-Watson stat	0.040527		Prob(F-statistic)	0.000000

(a) Retail Sales: Exponential Trend Regression - Nonlinear Least Squares



(b) Retail Sales: Exponential Trend Residual Plot

Figure 5.7: Retail Sales: Exponential Trend

and Figure 5.7b we show the nonlinear least squares estimation results and residual plot for the exponential trend model. The diagnostic statistics and residual plot indicate that the exponential trend fits better than the linear but worse than the quadratic.

Thus far we've been informal in our comparison of the linear, quadratic and exponential trend models for retail sales. We've noticed, for example, that the quadratic trend seems to fit the best. The quadratic trend model, however, contains one more parameter than the other two, so it's not surprising that it fits a little better, and there's no guarantee that its better fit on historical data will translate into better out-of-sample forecasting performance. (Recall

	Linear Trend	Quadratic Trend	Exponential Trend
AIC	19.35	15.83	17.15
SIC	19.37	15.86	17.17

Figure 5.8: Model Selection Criteria: Linear, Quadratic, and Exponential Trend Models

the parsimony principle.) To settle upon a final model, we examine the *AIC* or *SIC*, which we summarize in Table 5.8 for the three trend models.<sup>10</sup> Both the *AIC* and *SIC* indicate that nonlinearity is important in the trend, as both rank the linear trend last. Both, moreover, favor the quadratic trend model. So let's use the quadratic trend model.

Figure 5.9 shows the history of retail sales, 1990.01-1993.12, together with out-of-sample point and 95% interval extrapolation forecasts, 1994.01-1994.12. The point forecasts look reasonable. The interval forecasts are computed under the (incorrect) assumption that the deviation of retail sales from trend is random noise, which is why they're of equal width throughout. Nevertheless, they look reasonable.

In Figure 5.10 we show the history of retail sales through 1993, the quadratic trend forecast for 1994, and the realization for 1994. The forecast is quite good, as the realization hugs the forecasted trend line quite closely. All of the realizations, moreover, fall inside the 95% forecast interval.

For comparison, we examine the forecasting performance of a simple linear trend model. Figure 5.11 presents the history of retail sales and the out-of-sample point and 95% interval extrapolation forecasts for 1994. The point forecasts look very strange. The huge drop forecasted relative to the historical sample path occurs because the linear trend is far below the sample path by the end of the sample. The confidence intervals are very wide, reflecting the large standard error of the linear trend regression relative to the quadratic trend regression.

---

<sup>10</sup>It's important that the exponential trend model be estimated in levels, in order to maintain comparability of the exponential trend model *AIC* and *SIC* with those of the other trend models.

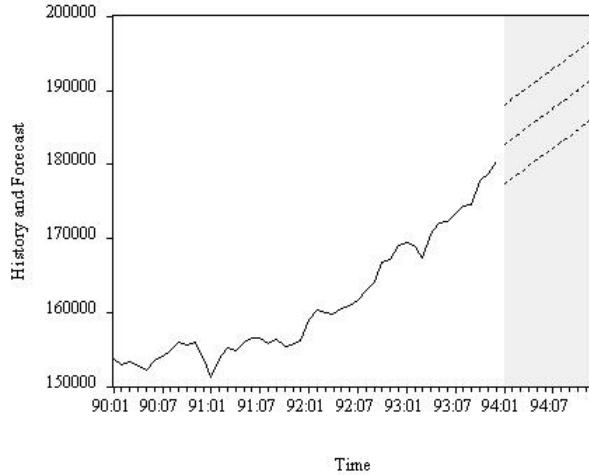


Figure 5.9: Retail Sales: Quadratic Trend Forecast

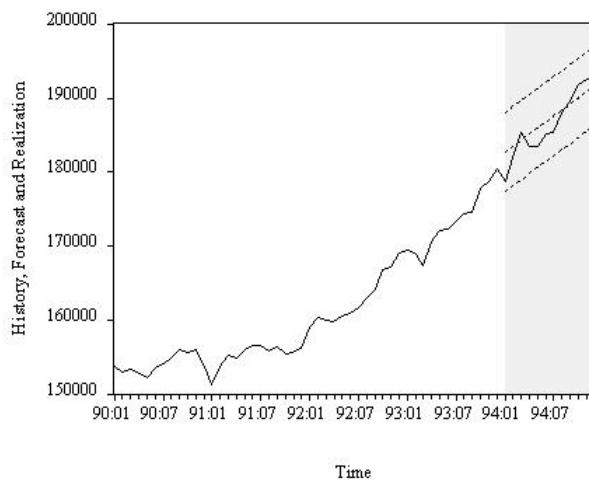


Figure 5.10: Retail Sales: Quadratic Trend Forecast and Realization

Finally, Figure 5.12 shows the history, the linear trend forecast for 1994, and the realization. The forecast is terrible – far below the realization. Even the very wide interval forecasts fail to contain the realizations. The reason for the failure of the linear trend forecast is that the forecasts (point and interval) are computed under the assumption that the linear trend model is actually the true DGP, whereas in fact the linear trend model is a very poor approximation to the trend in retail sales.

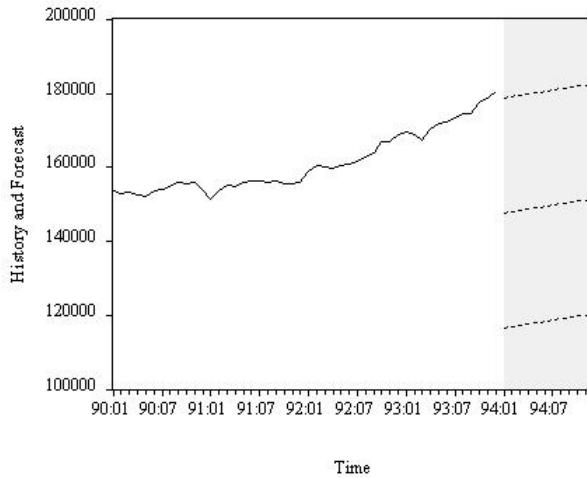


Figure 5.11: Retail Sales: Linear Trend Forecast

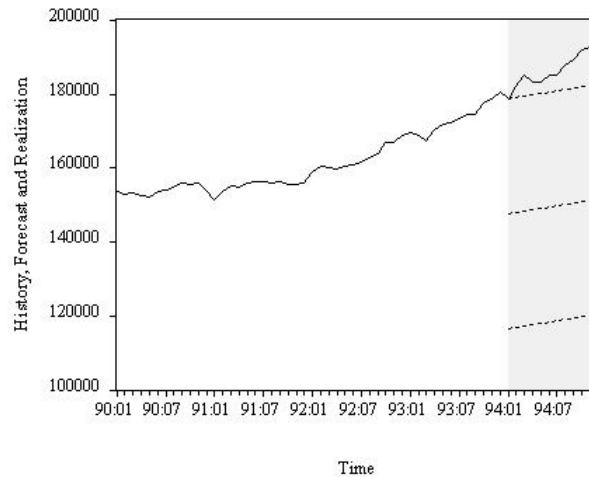


Figure 5.12: Retail Sales: Linear Trend Forecast and Realization

### 5.3 Deterministic Seasonality

Time series fluctuate over time, and we often mentally allocate those fluctuations to unobserved underlying components, such as trends, seasonals, and cycles. In this section we focus on **seasonals**.<sup>11</sup> More precisely, in our general unobserved-components model,

$$y_t = T_t + S_t + C_t + \varepsilon_t,$$

---

<sup>11</sup>Later we'll define and study cycles. Not all components need be present in all observed series.

we now include only the seasonal and noise components,

$$y_t = S_t + \varepsilon_t.$$

Seasonality involves patterns that repeat every year.<sup>12</sup> Seasonality is produced by aspects of technologies, preferences and institutions that are linked to the calendar, such as holidays that occur at the same time each year.

You might imagine that, although certain series are seasonal for obvious reasons, seasonality is nevertheless uncommon. On the contrary, and perhaps surprisingly, seasonality is pervasive in business and economics. Any technology that involves the weather, such as production of agricultural commodities, is likely to be seasonal. Preferences may also be linked to the calendar. For example, people want to do more vacation travel in the summer, which tends to increase both the price and quantity of summertime gasoline sales. Finally, social institutions that are linked to the calendar, such as holidays, are responsible for seasonal variation in many series. Purchases of retail goods skyrocket, for example, every Christmas season.

We will introduce both **deterministic seasonality** and **stochastic seasonality**. We treat the deterministic case here, and we treat the stochastic case later in Chapter .

### 5.3.1 Seasonal Models

A key technique for modeling seasonality is **regression on seasonal dummies**. Let  $s$  be the number of seasons in a year. Normally we'd think of four seasons in a year, but that notion is too restrictive for our purposes. Instead, think of  $s$  as the number of observations on a series in each year. Thus  $s = 4$  if we have quarterly data,  $s = 12$  if we have monthly data,  $s = 52$  if we have weekly data, and so forth.

---

<sup>12</sup>Note therefore that seasonality is impossible, and therefore not an issue, in data recorded once per year, or less often than once per year.

Now let's construct **seasonal dummy variables**, which indicate which season we're in. If, for example, there are four seasons ( $s = 4$ ), we create  $D_1 = (1, 0, 0, 0, \dots)$ ,  $D_2 = (0, 1, 0, 0, \dots)$ ,  $D_3 = (0, 0, 1, 0, \dots)$  and  $D_4 = (0, 0, 0, 1, \dots)$ .  $D_1$  indicates whether we're in the first quarter (it's 1 in the first quarter and zero otherwise),  $D_2$  indicates whether we're in the second quarter (it's 1 in the second quarter and zero otherwise), and so on. At any given time, we can be in only one of the four quarters, so only one seasonal dummy is nonzero.

The deterministic seasonal component is

$$S_t = \sum_{i=1}^s \gamma_i D_{it}.$$

It is an intercept that varies in a deterministic manner over throughout the seasons within each year. Those different intercepts, the  $\gamma_i$ 's, are called the **seasonal factors**; they summarize the seasonal pattern over the year.

In the absence of seasonality, the  $\gamma_i$ 's are all the same, so we drop all the seasonal dummies and instead include an intercept in the usual way.

Crucially, note that the deterministic seasonal variation is perfectly predictable, just as with our earlier-studied deterministic trend variation.

### 5.3.2 Seasonal Estimation

Before we can estimate seasonal models we need to create and store on the computer the seasonal dummies  $D_i$ ,  $i = 1, \dots, s$ . Most software packages have a command to do it instantly.

We fit our seasonal models to data on a time series  $y$  using **ordinary least-squares regression**. We simply run

$$y \rightarrow D_1, \dots, D_s.$$

We can also blend models to capture trend and seasonality simultaneously.

For example, we capture quadratic trend plus seasonality by running

$$y \rightarrow \text{TIME}, \text{TIME}^2, D_1, \dots, D_s.$$

Note that whenever we include a full set of seasonal dummies, we drop the intercept, to avoid perfect multicollinearity.<sup>13</sup>

### 5.3.3 Forecasting Seasonals

Consider constructing an  $h$ -step-ahead point forecast,  $y_{T+h,T}$  at time  $T$ . As with the pure trend model, there's no problem of forecasting the right-hand side variables, due to the special (perfectly predictable) nature of seasonal dummies, so point forecasts are easy to generate. The model is

$$y_t = \sum_{i=1}^s \gamma_i D_{it} + \varepsilon_t,$$

so that at time  $T + h$ ,

$$y_{T+h} = \sum_{i=1}^s \gamma_i D_{i,T+h} + \varepsilon_{T+h}.$$

As with the trend model discussed earlier, we project the right side of the equation on what's known at time  $T$  (that is, the time- $T$  information set,  $\Omega_T$ ) to obtain the forecast

$$y_{T+h,T} = \sum_{i=1}^s \gamma_i D_{i,T+h}.$$

There is no FRV problem, because  $D_{i,T+h}$  is known with certainty, for all  $i$  and  $h$ . As always, we make the point forecast operational by replacing

---

<sup>13</sup>See also EPC ??.

unknown parameters with estimates,

$$\hat{y}_{T+h,T} = \sum_{i=1}^s \hat{\gamma}_i D_{i,T+h}.$$

To form density forecasts we again proceed precisely as in the trend model. If we assume that the regression disturbance is normally distributed, then the density forecast ignoring parameter estimation uncertainty is  $N(\hat{y}_{T+h,T}, \sigma^2)$ , where  $\sigma$  is the standard deviation of the regression disturbance. The operational density forecast is then  $N(\hat{y}_{T+h,T}, \hat{\sigma}^2)$ , and the corresponding 95% interval forecast is  $\hat{y}_{T+h,T} \pm 1.96\hat{\sigma}$ .

We can use simulation-based methods from Chapter 4 to dispense with the normality assumption or account for parameter-estimation uncertainty.

### 5.3.4 Forecasting Housing Starts

We'll use the seasonal modeling techniques that we've developed in this chapter to build a forecasting model for housing starts. Housing starts are seasonal because it's usually preferable to start houses in the spring, so that they're completed before winter arrives. We have monthly data on U.S. housing starts; we'll use the 1946.01-1993.12 period for estimation and the 1994.01-1994.11 period for out-of-sample forecasting. We show the entire series in Figure 5.13, and we zoom in on the 1990.01-1994.11 period in Figure 5.14 in order to reveal the seasonal pattern in better detail.

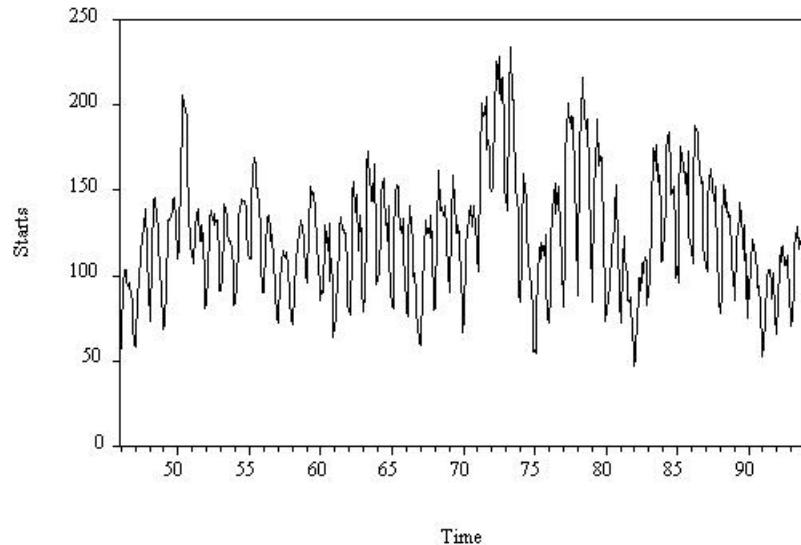


Figure 5.13: Housing Starts, 1946-1994

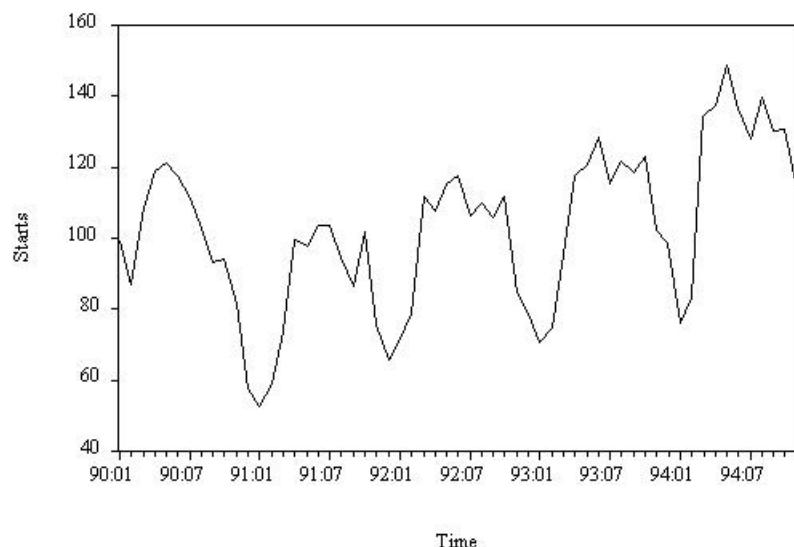


Figure 5.14: Housing Starts, 1946-1994 - Zoom on 1990-1994

The figures reveal that there is no trend, so we'll work with the pure seasonal model,

$$y_t = \sum_{i=1}^s \gamma_i D_{it} + \varepsilon_t.$$

Table 5.15a shows the estimation results. The twelve seasonal dummies account for more than a third of the variation in housing starts, as  $R^2 = .38$ . At least some of the remaining variation is cyclical, which the model is not designed to capture. (Note the very low Durbin-Watson statistic.)

The residual plot in Figure 5.15b makes clear the strengths and limitations of the model. First compare the actual and fitted values. The fitted values go through the same seasonal pattern every year – there's nothing in the model other than deterministic seasonal dummies – but that rigid seasonal pattern picks up a lot of the variation in housing starts. It doesn't pick up *all* of the variation, however, as evidenced by the serial correlation that's apparent in the residuals. Note the dips in the residuals, for example, in recessions (e.g., 1990, 1982, 1980, and 1975), and the peaks in booms.

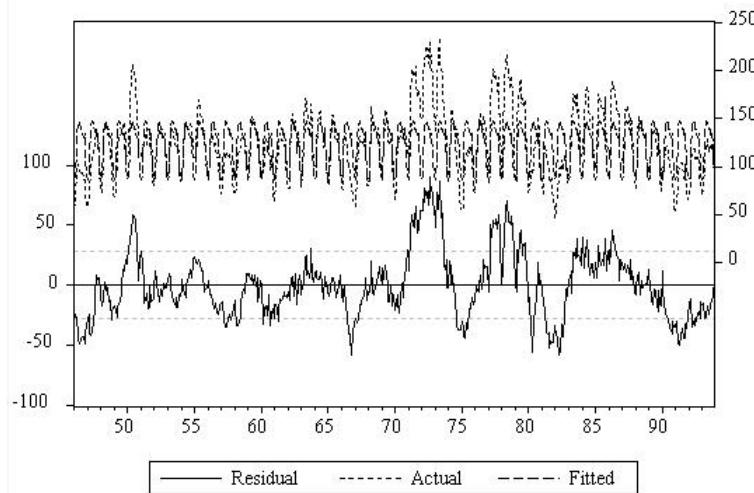
The estimated seasonal factors are just the twelve estimated coefficients on the seasonal dummies; we graph them in Figure 5.16. The seasonal effects are very low in January and February, and then rise quickly and peak in May, after which they decline, at first slowly and then abruptly in November and December.

In Figure 5.17 we see the history of housing starts through 1993, together with the out-of-sample point and 95% interval extrapolation forecasts for the first eleven months of 1994. The forecasts look reasonable, as the model has evidently done a good job of capturing the seasonal pattern. The forecast intervals are quite wide, however, reflecting the fact that the seasonal effects captured by the forecasting model are responsible for only about a third of the variation in the variable being forecast.

In Figure 5.18, we include the 1994 realization. The forecast appears highly accurate, as the realization and forecast are quite close throughout. Moreover, the realization is everywhere well within the 95% interval.

LS // Dependent Variable is STARTS				
Sample: 1946:01 1993:12				
Included observations: 576				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D1	86.50417	4.029055	21.47009	0.0000
D2	89.50417	4.029055	22.21468	0.0000
D3	122.8833	4.029055	30.49929	0.0000
D4	142.1687	4.029055	35.28588	0.0000
D5	147.5000	4.029055	36.60908	0.0000
D6	145.9979	4.029055	36.23627	0.0000
D7	139.1125	4.029055	34.52733	0.0000
D8	138.4167	4.029055	34.35462	0.0000
D9	130.5625	4.029055	32.40524	0.0000
D10	134.0917	4.029055	33.28117	0.0000
D11	111.8333	4.029055	27.75671	0.0000
D12	92.15833	4.029055	22.87344	0.0000
R-squared	0.383780		Mean dependent var	123.3944
Adjusted R-squared	0.371762		S.D. dependent var	35.21775
S.E. of regression	27.91411		Akaike info criterion	6.678878
Sum squared resid	439467.5		Schwarz criterion	6.769630
Log likelihood	-2728.825		F-statistic	31.93250
Durbin-Watson stat	0.154140		Prob(F-statistic)	0.0000000

(a) Housing Starts: Seasonal Dummy Variables



(b) Housing Starts: Seasonal Dummy Variables, Residual Plot

Figure 5.15: Housing Starts: Seasonal Dummy Model

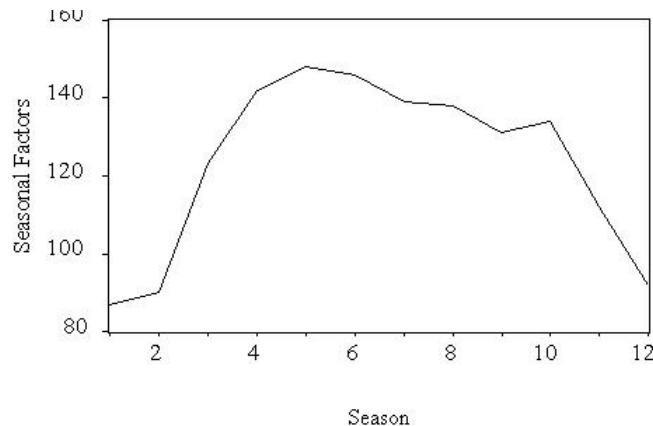


Figure 5.16: Housing Starts: Estimated Seasonal Factors

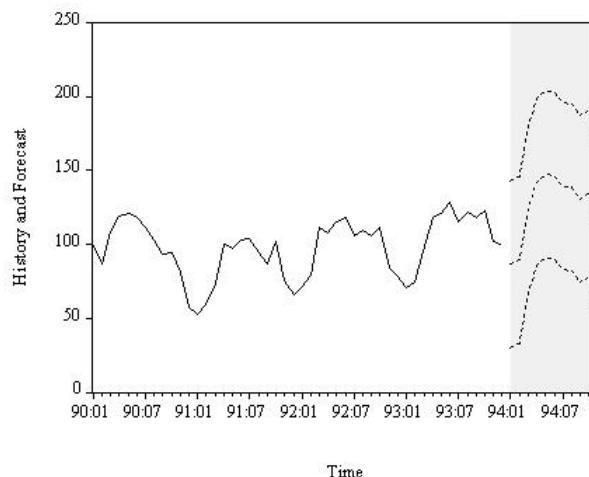


Figure 5.17: Housing Starts: Seasonal Model Forecast

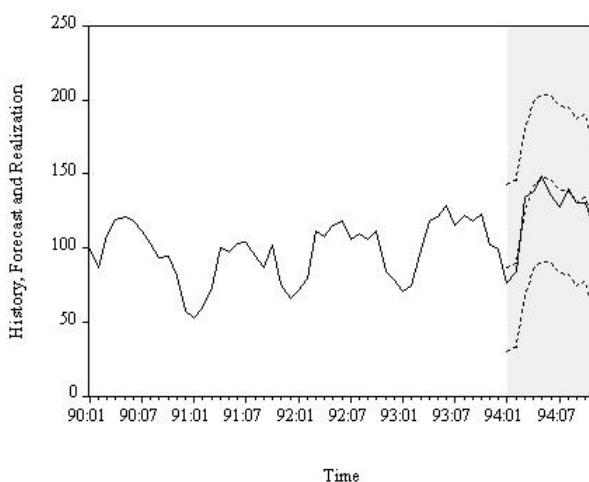


Figure 5.18: Housing Starts: Seasonal Model Forecast and Realization

## 5.4 Exercises, Problems and Complements

1. Calculating forecasts from trend models.

You work for the International Monetary Fund in Washington DC, monitoring Singapore's real consumption expenditures. Using a sample of quarterly real consumption data (measured in billions of 2005 Singapore dollars),  $y_t$ ,  $t = 1990.1, \dots, 2006.4$ , you estimate the linear consumption trend model,  $y_t = \beta_0 + \beta_1 TIME_t + \varepsilon_t$ , where  $\varepsilon_t \sim iidN(0, \sigma^2)$ , obtaining the estimates  $\hat{\beta}_0 = 0.51$ ,  $\hat{\beta}_1 = 2.30$ , and  $\hat{\sigma}^2 = 16$ . Based upon your estimated trend model, construct feasible point, interval and density forecasts for 2010.1.

2. Calendar span vs. observation count in trend estimation.

Suppose it's the last day of the year. You are using a trend model to produce a 1-year-ahead (end-of-year) forecast of a stock (as opposed to flow) variable observed daily. Would you prefer to estimate your forecasting model using the most recent 500 daily observations (and then forecast 365 steps ahead) or 50 annual end-of-year observations (and then forecast 1 step ahead)? Discuss. In particular, if you prefer to use the 50 annual observations, why is that? Isn't 500 a much larger sample size than 50, so shouldn't you prefer to use it?

3. Mechanics of trend estimation and forecasting.

Obtain from the web an upward-trending monthly series that interests you. Choose your series such that it spans at least ten years, and such that it ends at the end of a year (i.e., in December).

- a. What is the series and why does it interest you? Produce a time series plot of it. Discuss.
- b. Fit linear, quadratic and exponential trend models to your series. Discuss the associated diagnostic statistics and residual plots.

- c. Select a trend model using the AIC and using the SIC. Do the selected models agree? If not, which do you prefer?
  - d. Use your preferred model to forecast each of the twelve months of the next year. Discuss.
  - e. The *residuals* from your fitted model are effectively a *detrended* version of your original series. Why? Plot them and discuss.
4. Properties of **polynomial trends**.

Consider a tenth-order deterministic polynomial trend:

$$T_t = \beta_0 + \beta_1 TIME_t + \beta_2 TIME_t^2 + \dots + \beta_{10} TIME_t^{10}.$$

- a. How many local maxima or minima may such a trend display?
  - b. Plot the trend for various values of the parameters to reveal some of the different possible trend shapes.
  - c. Is this an attractive trend model in general? Why or why not?
  - d. How do you expect this trend to fit in-sample?
  - e. How do you expect this trend to forecast out-of-sample?
5. Seasonal adjustment.

One way to deal with seasonality in a series is simply to remove it, and then to model and forecast the **seasonally adjusted series**.<sup>14</sup> This strategy is perhaps appropriate in certain situations, such as when interest centers explicitly on forecasting **nonseasonal fluctuations**, as is often the case in macroeconomics. Seasonal adjustment is often inappropriate in business forecasting situations, however, precisely because interest typically centers on forecasting *all* the variation in a series, not just the nonseasonal part. If seasonality is responsible for a large part

---

<sup>14</sup>Removal of seasonality is called **seasonal adjustment**.

of the variation in a series of interest, the last thing a forecaster wants to do is discard it and pretend it isn't there.

- a. Discuss in detail how you'd use dummy variable regression methods to seasonally adjust a series. (Hint: the seasonally adjusted series is closely related to the residual from the seasonal dummy variable regression.)
- b. Search the Web (or the library) for information on the latest U.S. Census Bureau seasonal adjustment procedure, and report what you learned.
6. Fourier seasonality.

Thus far we have used seasonal dummies. We can also take a Fourier series approach, the benefits of which are two-fold. First, it produces a smooth seasonal pattern, which accords with the basic intuition that the progression through different seasons is gradual rather than discontinuous. Second, it promotes parsimony, which not only respects the parsimony principle but also enhances numerical stability in estimation.

The Fourier approach may be especially useful with high-frequency data. Consider, for example, seasonal daily data. For a variety of reasons, regression on more than three hundred daily dummies may not be appealing! So instead of using

$$S_t = \sum_{s=1}^{365} \gamma_i D_{it},$$

we can use

$$S_t = \sum_{p=1}^P \left( \delta_{c,p} \cos \left( 2\pi p \frac{d_t}{365} \right) + \delta_{s,p} \sin \left( 2\pi p \frac{d_t}{365} \right) \right),$$

where  $d_t$  is a repeating step function that cycles through 1, ..., 365. We can choose  $P$  using the usual model selection criteria. (Note that for simplicity we have dropped February 29 in leap years.)

## 5.5 Notes

# Chapter 6

## Cycles I: Autoregressions and Wold’s Chain Rule

We’ve already considered models with trend and seasonal components. In this chapter we consider a crucial third component, **cycles**. When you think of a “cycle,” you probably think of the sort of rigid up-and-down pattern depicted in Figure 6.1. Such cycles can sometimes arise, but cyclical fluctuations in business, finance, economics and government are typically much less rigid. In fact, when we speak of cycles, we have in mind a much more general notion of cyclicality: any sort of stable, mean-reverting dynamics not captured by trends or seasonals.

Cycles, according to our broad interpretation, may display the sort of back-and-forth movement characterized in Figure 6.1, but they need not. All we require is that there be some stable dynamics (“covariance stationary” dynamics, in the jargon that we’ll shortly introduce) that link the present to the past, and hence the future to the present. Cycles are present in most of the series that concern us, and it’s crucial that we know how to model and forecast them, because their history conveys information regarding their future.

Trend and seasonal dynamics are simple, so we can capture them with simple models. Cyclical dynamics, however, are a bit more complicated, and

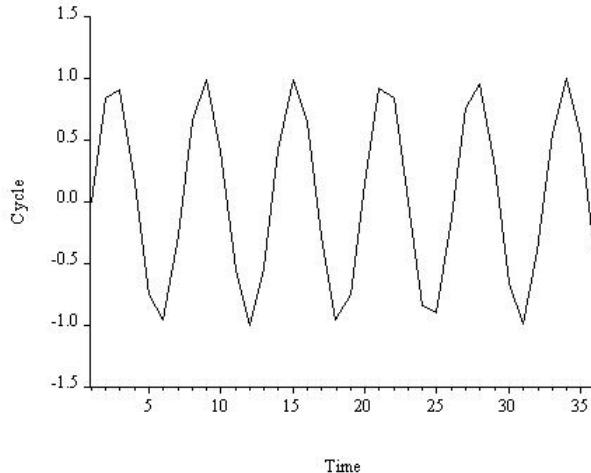


Figure 6.1: A Rigid Cyclical Pattern

consequently the cycle models we need are a bit more involved. We will emphasize autoregressive models.

Let's jump in.

## 6.1 Characterizing Cycles

Here we introduce methods for characterizing cyclical dynamics in model-free fashion.

### 6.1.1 Covariance Stationary Time Series

#### Basic Ideas

A **realization** of a time series is an ordered set,

$$\{ \dots, y_{-2}, y_{-1}, y_0, y_1, y_2, \dots \}.$$

Typically the observations are ordered in time – hence the name **time series** – but they don't have to be. We could, for example, examine a spatial series, such as office space rental rates as we move along a line from a point in

midtown Manhattan to a point in the New York suburbs thirty miles away. But the most important case, by far, involves observations ordered in time, so that's what we'll stress.

In theory, a time series realization begins in the infinite past and continues into the infinite future. This perspective may seem abstract and of limited practical applicability, but it will be useful in deriving certain very important properties of the models we'll be using shortly. In practice, of course, the data we observe is just a finite subset of a realization,  $\{y_1, \dots, y_T\}$ , called a **sample path**.

Shortly we'll be building models for cyclical time series. If the underlying probabilistic structure of the series were changing over time, we'd be doomed – there would be no way to relate the future to the past, because the laws governing the future would differ from those governing the past. At a minimum we'd like a series' mean and its covariance structure (that is, the covariances between current and past values) to be stable over time, in which case we say that the series is **covariance stationary**. Let's discuss covariance stationarity in greater depth. The first requirement for a series to be covariance stationary is that the mean of the series be stable over time. The mean of the series at time  $t$  is  $Ey_t = \mu_t$ . If the mean is stable over time, as required by covariance stationarity, then we can write  $Ey_t = \mu$ , for all  $t$ . Because the mean is constant over time, there's no need to put a time subscript on it.

The second requirement for a series to be covariance stationary is that its covariance structure be stable over time. Quantifying stability of the covariance structure is a bit tricky, but tremendously important, and we do it using the **autocovariance function**. The autocovariance at displacement  $\tau$  is just the covariance between  $y_t$  and  $y_{t-\tau}$ . It will of course depend on  $\tau$ , and it may also depend on  $t$ , so in general we write

$$\gamma(t, \tau) = cov(y_t, y_{t-\tau}) = E(y_t - \mu)(y_{t-\tau} - \mu).$$

If the covariance structure is stable over time, as required by covariance stationarity, then the autocovariances depend only on displacement,  $\tau$ , not on time,  $t$ , and we write  $\gamma(t, \tau) = \gamma(\tau)$ , for all  $t$ .

The autocovariance function is important because it provides a basic summary of cyclical dynamics in a covariance stationary series. By examining the autocovariance structure of a series, we learn about its dynamic behavior. We graph and examine the autocovariances as a function of  $\tau$ . Note that the autocovariance function is symmetric; that is,  $\gamma(\tau) = \gamma(-\tau)$ , for all  $\tau$ . Typically, we'll consider only non-negative values of  $\tau$ . Symmetry reflects the fact that the autocovariance of a covariance stationary series depends only on displacement; it doesn't matter whether we go forward or backward. Note also that  $\gamma(0) = \text{cov}(y_t, y_t) = \text{var}(y_t)$ .

There is one more technical requirement of covariance stationarity: we require that the variance of the series – the autocovariance at displacement 0,  $\gamma(0)$ , be finite. It can be shown that no autocovariance can be larger in absolute value than  $\gamma(0)$ , so if  $\gamma(0) < \infty$ , then so too are all the other autocovariances.

It may seem that the requirements for covariance stationarity are quite stringent, which would bode poorly for our models, almost all of which invoke covariance stationarity in one way or another. It is certainly true that many economic, business, financial and government series are not covariance stationary. An upward trend, for example, corresponds to a steadily increasing mean, and seasonality corresponds to means that vary with the season, both of which are violations of covariance stationarity.

But appearances can be deceptive. Although many series are not covariance stationary, it is frequently possible to work with models that give special treatment to nonstationary components such as trend and seasonality, so that the cyclical component that's left over is likely to be covariance stationary. We'll often adopt that strategy. Alternatively, simple transformations often

appear to transform nonstationary series to covariance stationarity. For example, many series that are clearly nonstationary in levels appear covariance stationary in growth rates.

In addition, note that although covariance stationarity requires means and covariances to be stable and finite, it places no restrictions on other aspects of the distribution of the series, such as skewness and kurtosis.<sup>1</sup> The upshot is simple: whether we work directly in levels and include special components for the nonstationary elements of our models, or we work on transformed data such as growth rates, the covariance stationarity assumption is not as unrealistic as it may seem.

Recall that the correlation between two random variables  $x$  and  $y$  is defined by

$$\text{corr}(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}.$$

That is, the correlation is simply the covariance, “normalized,” or “standardized,” by the product of the standard deviations of  $x$  and  $y$ . Both the correlation and the covariance are measures of linear association between two random variables. The correlation is often more informative and easily interpreted, however, because the construction of the correlation coefficient guarantees that  $\text{corr}(x, y) \in [-1, 1]$ , whereas the covariance between the same two random variables may take any value. The correlation, moreover, does not depend on the units in which  $x$  and  $y$  are measured, whereas the covariance does. Thus, for example, if  $x$  and  $y$  have a covariance of ten million, they’re not necessarily very strongly associated, whereas if they have a correlation of .95, it is unambiguously clear that they are very strongly associated.

In light of the superior interpretability of correlations as compared to covariances, we often work with the correlation, rather than the covariance, between  $y_t$  and  $y_{t-\tau}$ . That is, we work with the **autocorrelation function**,

---

<sup>1</sup>For that reason, covariance stationarity is sometimes called **second-order stationarity** or **weak stationarity**.

$\rho(\tau)$ , rather than the autocovariance function,  $\gamma(\tau)$ . The autocorrelation function is obtained by dividing the autocovariance function by the variance,

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)}, \tau = 0, 1, 2, \dots$$

The formula for the autocorrelation is just the usual correlation formula, specialized to the correlation between  $y_t$  and  $y_{t-\tau}$ . To see why, note that the variance of  $y_t$  is  $\gamma(0)$ , and by covariance stationarity, the variance of  $y$  at any other time  $y_{t-\tau}$  is also  $\gamma(0)$ . Thus,

$$\rho(\tau) = \frac{cov(y_t, y_{t-\tau})}{\sqrt{var(y_t)} \sqrt{var(y_{t-\tau})}} = \frac{\gamma(\tau)}{\sqrt{\gamma(0)} \sqrt{\gamma(0)}} = \frac{\gamma(\tau)}{\gamma(0)},$$

as claimed. Note that we always have  $\rho(0) = \frac{\gamma(0)}{\gamma(0)} = 1$ , because any series is perfectly correlated with itself. Thus the autocorrelation at displacement 0 isn't of interest; rather, only the autocorrelations *beyond* displacement 0 inform us about a series' dynamic structure.

Finally, the **partial autocorrelation function**,  $p(\tau)$ , is sometimes useful.  $p(\tau)$  is just the coefficient of  $y_{t-\tau}$  in a population linear regression of  $y_t$  on  $y_{t-1}, \dots, y_{t-\tau}$ .<sup>2</sup> We call such regressions **autoregressions**, because the variable is regressed on lagged values of itself. It's easy to see that the autocorrelations and partial autocorrelations, although related, differ in an important way. The autocorrelations are just the "simple" or "regular" correlations between  $y_t$  and  $y_{t-\tau}$ . The partial autocorrelations, on the other hand, measure the association between  $y_t$  and  $y_{t-\tau}$  after *controlling* for the effects of  $y_{t-1}, \dots, y_{t-\tau+1}$ ; that is, they measure the partial correlation between  $y_t$  and  $y_{t-\tau}$ .

As with the autocorrelations, we often graph the partial autocorrelations

---

<sup>2</sup>To get a feel for what we mean by "population regression," imagine that we have an infinite sample of data at our disposal, so that the parameter estimates in the regression are not contaminated by sampling variation – that is, they're the true population values. The thought experiment just described is a population regression.

as a function of  $\tau$  and examine their qualitative shape, which we'll do soon. Like the autocorrelation function, the partial autocorrelation function provides a summary of a series' dynamics, but as we'll see, it does so in a different way.<sup>3</sup>

All of the covariance stationary processes that we will study subsequently have autocorrelation and partial autocorrelation functions that approach zero, one way or another, as the displacement gets large. In Figure 6.2 we show an autocorrelation function that displays gradual one-sided damping, and in Figure 6.3 we show a constant autocorrelation function; the latter could not be the autocorrelation function of a stationary process, whose autocorrelation function must eventually decay. The precise decay patterns of autocorrelations and partial autocorrelations of a covariance stationary series, however, depend on the specifics of the series. In Figure 6.4, for example, we show an autocorrelation function that displays damped oscillation – the autocorrelations are positive at first, then become negative for a while, then positive again, and so on, while continuously getting smaller in absolute value. Finally, in Figure 6.5 we show an autocorrelation function that differs in the way it approaches zero – the autocorrelations drop abruptly to zero beyond a certain displacement.

---

<sup>3</sup>Also in parallel to the autocorrelation function, the partial autocorrelation at displacement 0 is always one and is therefore uninformative and uninteresting. Thus, when we graph the autocorrelation and partial autocorrelation functions, we'll begin at displacement 1 rather than displacement 0.

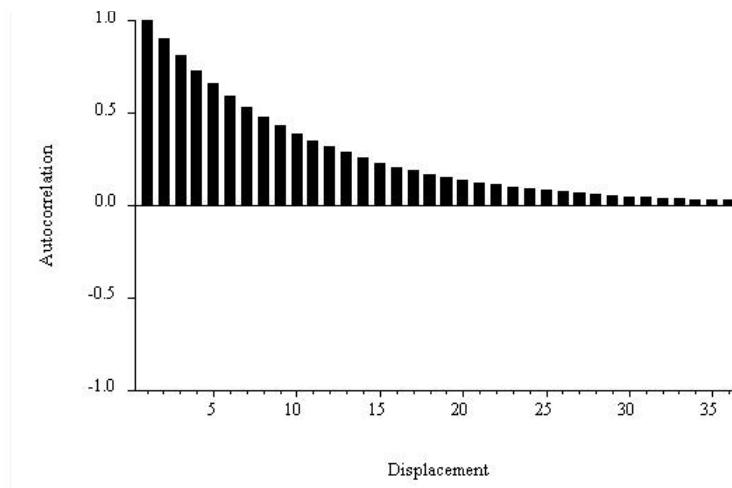


Figure 6.2: Autocorrelation Function: One-sided Gradual Damping

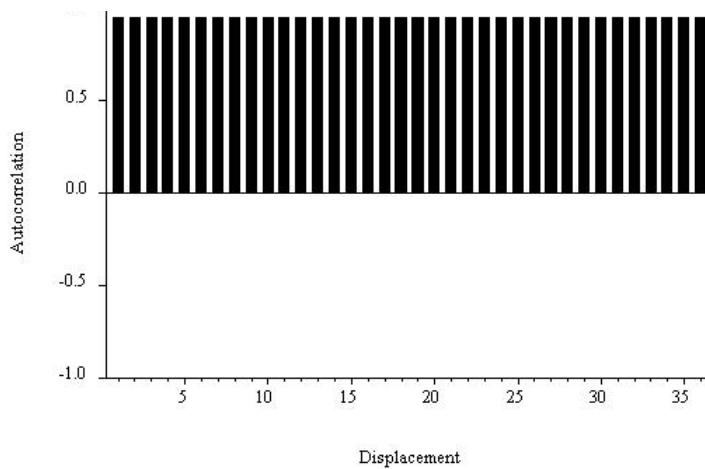


Figure 6.3: Constant Autocorrelation

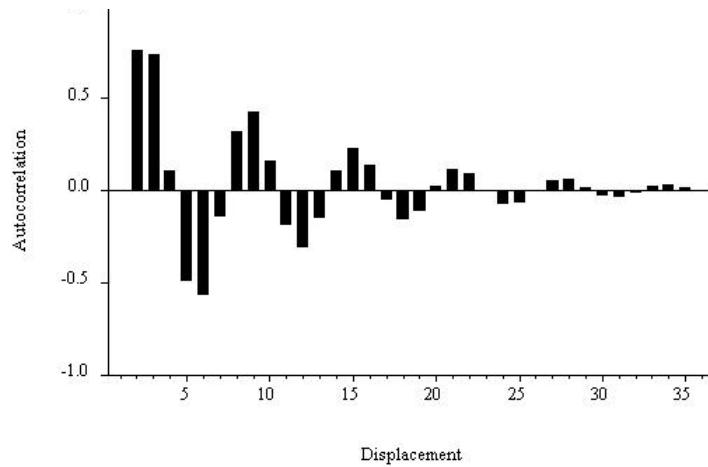


Figure 6.4: Autocorrelation Function: Gradual Damped Oscillation

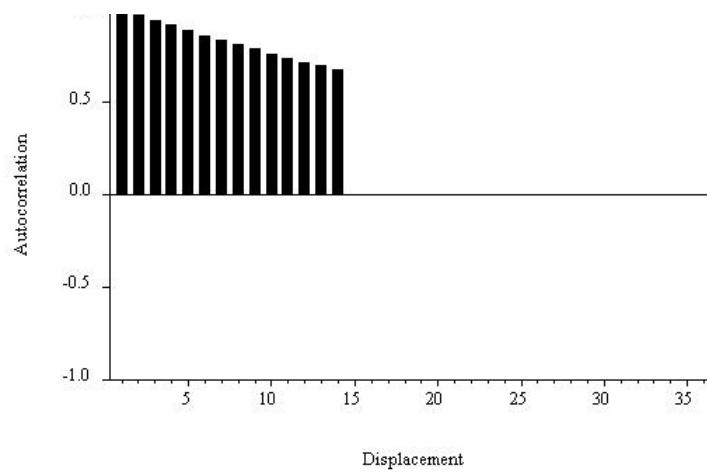


Figure 6.5: Autocorrelation Function: Sharp Cutoff

## 6.2 White Noise

### 6.2.1 Basic Ideas

Later in this chapter we'll study the population properties of certain important time series models, or **time series processes**. Before we estimate time series models, we need to understand their population properties, assuming that the postulated model is true. The simplest of all such time series processes is the fundamental building block from which all others are constructed. In fact, it's so important that we introduce it now. We use  $y$  to denote the observed series of interest. Suppose that

$$y_t = \varepsilon_t$$

$$\varepsilon_t \sim (0, \sigma^2),$$

where the “shock,”  $\varepsilon_t$ , is uncorrelated over time. We say that  $\varepsilon_t$ , and hence  $y_t$ , is **serially uncorrelated**. Throughout, unless explicitly stated otherwise, we assume that  $\sigma^2 < \infty$ . Such a process, with zero mean, constant variance, and no serial correlation, is called **zero-mean white noise**, or simply **white noise**.<sup>4</sup> Sometimes for short we write

$$\varepsilon_t \sim WN(0, \sigma^2)$$

and hence

$$y_t \sim WN(0, \sigma^2).$$

Note that, although  $\varepsilon_t$  and hence  $y_t$  are serially uncorrelated, they are not necessarily serially independent, because they are not necessarily normally distributed.<sup>5</sup> If in addition to being serially uncorrelated,  $y$  is serially

---

<sup>4</sup>It's called white noise by analogy with white light, which is composed of all colors of the spectrum, in equal amounts. We can think of white noise as being composed of a wide variety of cycles of differing periodicities, in equal amounts.

<sup>5</sup>Recall that zero correlation implies independence only in the normal case.

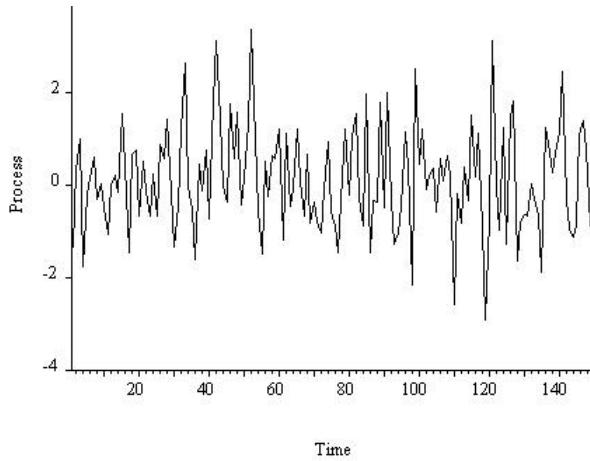


Figure 6.6: Realization of White Noise Process

independent, then we say that  $y$  is **independent white noise**.<sup>6</sup> We write

$$y_t \sim iid(0, \sigma^2),$$

and we say that “ $y$  is independently and identically distributed with zero mean and constant variance.” If  $y$  is serially uncorrelated and normally distributed, then it follows that  $y$  is also serially independent, and we say that  $y$  is **normal white noise**, or **Gaussian white noise**.<sup>7</sup> We write

$$y_t \sim iidN(0, \sigma^2).$$

We read “ $y$  is independently and identically distributed as normal, with zero mean and constant variance,” or simply “ $y$  is Gaussian white noise.” In Figure 6.6 we show a sample path of Gaussian white noise, of length  $T = 150$ , simulated on a computer. There are no patterns of any kind in the series due to the independence over time.

You’re already familiar with white noise, although you may not realize

---

<sup>6</sup>Another name for independent white noise is **strong white noise**, in contrast to standard serially uncorrelated **weak white noise**.

<sup>7</sup>Carl Friedrich Gauss, one of the greatest mathematicians of all time, discovered the normal distribution some 200 years ago; hence the adjective “Gaussian.”

it. Recall that the disturbance in a regression model is typically assumed to be white noise of one sort or another. There's a subtle difference here, however. Regression disturbances are not observable, whereas we're working with an observed series. Later, however, we'll see how all of our models for observed series can be used to model unobserved variables such as regression disturbances.

Let's characterize the dynamic stochastic structure of white noise,  $y_t \sim WN(0, \sigma^2)$ . By construction the unconditional mean of  $y$  is  $E(y_t) = 0$ , and the unconditional variance of  $y$  is  $\text{var}(y_t) = \sigma^2$ . Note in particular that the unconditional mean and variance are constant. In fact, the unconditional mean and variance must be constant for any covariance stationary process. The reason is that constancy of the unconditional mean was our first explicit requirement of covariance stationarity, and that constancy of the unconditional variance follows implicitly from the second requirement of covariance stationarity, that the autocovariances depend only on displacement, not on time.<sup>8</sup>

To understand fully the linear dynamic structure of a covariance stationary time series process, we need to compute and examine its mean and its autocovariance function. For white noise, we've already computed the mean and the variance, which is the autocovariance at displacement 0. We have yet to compute the rest of the autocovariance function; fortunately, however, it's very simple. Because white noise is, by definition, uncorrelated over time, all the autocovariances, and hence all the autocorrelations, are zero beyond displacement 0.<sup>9</sup> Formally, then, the autocovariance function for a white noise process is

$$\gamma(\tau) = \begin{cases} \sigma^2, & \tau = 0 \\ 0, & \tau \geq 1, \end{cases}$$

---

<sup>8</sup>Recall that  $\sigma^2 = \gamma(0)$ .

<sup>9</sup>If the autocovariances are all zero, so are the autocorrelations, because the autocorrelations are proportional to the autocovariances.

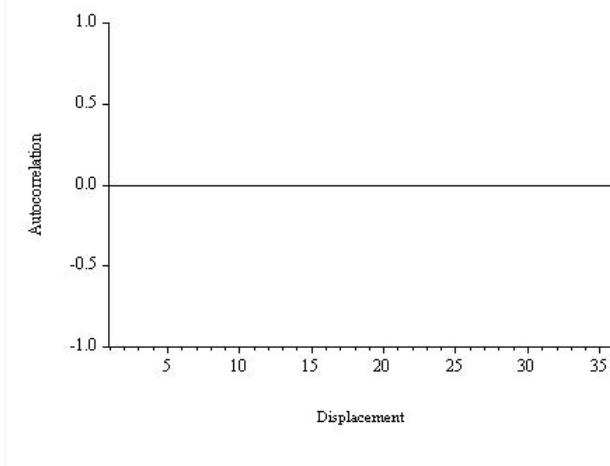


Figure 6.7: White Noise Autocorrelation Function

and the autocorrelation function for a white noise process is

$$\rho(\tau) = \begin{cases} 1, & \tau = 0 \\ 0, & \tau \geq 1. \end{cases}$$

In Figure 6.7 we plot the white noise autocorrelation function.

Finally, consider the partial autocorrelation function for a white noise series. For the same reason that the autocorrelation at displacement 0 is always one, so too is the partial autocorrelation at displacement 0. For a white noise process, all partial autocorrelations beyond displacement 0 are zero, which again follows from the fact that white noise, by construction, is serially uncorrelated. Population regressions of  $y_t$  on  $y_{t-1}$ , or on  $y_{t-1}$  and  $y_{t-2}$ , or on any other lags, produce nothing but zero coefficients, because the process is serially uncorrelated. Formally, the partial autocorrelation function of a white noise process is

$$p(\tau) = \begin{cases} 1, & \tau = 0 \\ 0, & \tau \geq 1. \end{cases}$$

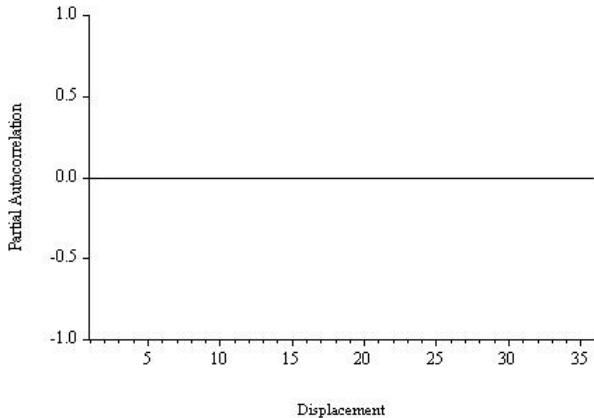


Figure 6.8: White Noise Partial Autocorrelation Function

We show the partial autocorrelation function of a white noise process in Figure 6.8. Again, it's degenerate, and exactly the same as the autocorrelation function!

White noise is very special, indeed degenerate in a sense, as what happens to a white noise series at any time is uncorrelated with anything in the past, and similarly, what happens in the future is uncorrelated with anything in the present or past. But understanding white noise is tremendously important for at least two reasons. First, as already mentioned, processes with much richer dynamics are built up by taking simple transformations of white noise.

Second, the goal of all time series modeling (and 1-step-ahead forecasting) is to reduce the data (or 1-step-ahead forecast errors) to white noise. After all, if such forecast errors aren't white noise, then they're serially correlated, which means that they're forecastable, and if forecast errors are forecastable then the forecast can't be very good. Thus it's important that we understand and be able to recognize white noise.

Thus far we've characterized white noise in terms of its mean, variance, autocorrelation function and partial autocorrelation function. Another characterization of dynamics involves the mean and variance of a process, *conditional* upon its past. In particular, we often gain insight into the dynamics in

a process by examining its conditional mean.<sup>10</sup> In fact, throughout our study of time series, we'll be interested in computing and contrasting the **unconditional mean and variance** and the **conditional mean and variance** of various processes of interest. Means and variances, which convey information about location and scale of random variables, are examples of what statisticians call **moments**. For the most part, our comparisons of the conditional and unconditional moment structure of time series processes will focus on means and variances (they're the most important moments), but sometimes we'll be interested in higher-order moments, which are related to properties such as skewness and kurtosis.

For comparing conditional and unconditional means and variances, it will simplify our story to consider independent white noise,  $y_t \sim iid(0, \sigma^2)$ . By the same arguments as before, the unconditional mean of  $y$  is 0 and the unconditional variance is  $\sigma^2$ . Now consider the conditional mean and variance, where the information set  $\Omega_{t-1}$  upon which we condition contains either the past history of the observed series,  $\Omega_{t-1} = y_{t-1}, y_{t-2}, \dots$ , or the past history of the shocks,  $\Omega_{t-1} = \varepsilon_{t-1}, \varepsilon_{t-2}, \dots$  (They're the same in the white noise case.) In contrast to the unconditional mean and variance, which must be constant by covariance stationarity, the conditional mean and variance need not be constant, and in general we'd expect them *not* to be constant. The unconditionally expected growth of laptop computer sales next quarter may be ten percent, but expected sales growth may be much higher, *conditional* upon knowledge that sales grew this quarter by twenty percent. For the independent white noise process, the conditional mean is

$$E(y_t | \Omega_{t-1}) = 0,$$

---

<sup>10</sup>If you need to refresh your memory on conditional means, consult any good introductory statistics book, such as Wonnacott and Wonnacott (1990).

and the conditional variance is

$$\text{var}(y_t|\Omega_{t-1}) = E[(y_t - E(y_t|\Omega_{t-1}))^2|\Omega_{t-1}] = \sigma^2.$$

Conditional and unconditional means and variances are identical for an independent white noise series; there are no dynamics in the process, and hence no dynamics in the conditional moments.

## 6.3 Estimation and Inference for the Mean, Autocorrelation and Partial Autocorrelation Functions

Now suppose we have a sample of data on a time series, and we don't know the true model that generated the data, or the mean, autocorrelation function or partial autocorrelation function associated with that true model. Instead, we want to use the data to estimate the mean, autocorrelation function, and partial autocorrelation function, which we might then use to help us learn about the underlying dynamics, and to decide upon a suitable model or set of models to fit to the data.

### 6.3.1 Sample Mean

The mean of a covariance stationary series is

$$\mu = E y_t.$$

A fundamental principle of estimation, called the **analog principle**, suggests that we develop estimators by replacing expectations with sample averages. Thus our estimator for the population mean, given a sample of size  $T$ , is the **sample mean**,

$$\bar{y} = \frac{1}{T} \sum_{t=1}^T y_t.$$

Typically we're not directly interested in the estimate of the mean, but it's needed for estimation of the autocorrelation function.

### 6.3.2 Sample Autocorrelations

The autocorrelation at displacement  $\tau$  for the covariance stationary series  $y$  is

$$\rho(\tau) = \frac{E[(y_t - \mu)(y_{t-\tau} - \mu)]}{E[(y_t - \mu)^2]}.$$

Application of the analog principle yields a natural estimator,

$$\hat{\rho}(\tau) = \frac{\frac{1}{T} \sum_{t=\tau+1}^T [(y_t - \bar{y})(y_{t-\tau} - \bar{y})]}{\frac{1}{T} \sum_{t=1}^T (y_t - \bar{y})^2} = \frac{\sum_{t=\tau+1}^T [(y_t - \bar{y})(y_{t-\tau} - \bar{y})]}{\sum_{t=1}^T (y_t - \bar{y})^2}.$$

This estimator, viewed as a function of  $\tau$ , is called the **sample autocorrelation function**, or **correlogram**. Note that some of the summations begin at  $t = \tau + 1$ , not at  $t = 1$ ; this is necessary because of the appearance of  $y_{t-\tau}$  in the sum. Note that we divide those same sums by  $T$ , even though only  $T - \tau$  terms appear in the sum. When  $T$  is large relative to  $\tau$  (which is the relevant case), division by  $T$  or by  $T - \tau$  will yield approximately the same result, so it won't make much difference for practical purposes, and moreover there are good mathematical reasons for preferring division by  $T$ .

It's often of interest to assess whether a series is reasonably approximated as white noise, which is to say whether all its autocorrelations are zero in population. A key result, which we simply assert, is that if a series is white noise, then the distribution of the sample autocorrelations in large samples is

$$\hat{\rho}(\tau) \sim N\left(0, \frac{1}{T}\right).$$

Note how simple the result is. The sample autocorrelations of a white noise series are approximately normally distributed, and the normal is always a convenient distribution to work with. Their mean is zero, which is to say the

sample autocorrelations are unbiased estimators of the true autocorrelations, which are in fact zero. Finally, the variance of the sample autocorrelations is approximately  $1/T$  (equivalently, the standard deviation is  $1/\sqrt{T}$ ), which is easy to construct and remember. Under normality, taking plus or minus two standard errors yields an approximate 95% confidence interval. Thus, if the series is white noise, approximately 95% of the sample autocorrelations should fall in the interval  $0 \pm 2/\sqrt{T}$ . In practice, when we plot the sample autocorrelations for a sample of data, we typically include the “two standard error bands,” which are useful for making informal graphical assessments of whether and how the series deviates from white noise.

The two-standard-error bands, although very useful, only provide 95% bounds for the sample autocorrelations taken one at a time. Ultimately, we’re often interested in whether a series is white noise, that is, whether *all* its autocorrelations are *jointly* zero. A simple extension lets us test that hypothesis. Rewrite the expression

$$\hat{\rho}(\tau) \sim N\left(0, \frac{1}{T}\right)$$

as

$$\sqrt{T}\hat{\rho}(\tau) \sim N(0, 1).$$

Squaring both sides yields<sup>11</sup>

$$T\hat{\rho}^2(\tau) \sim \chi_1^2.$$

It can be shown that, in addition to being approximately normally distributed, the sample autocorrelations at various displacements are approximately independent of one another. Recalling that the sum of independent  $\chi^2$  variables is also  $\chi^2$  with degrees of freedom equal to the sum of the degrees

---

<sup>11</sup>Recall that the square of a standard normal random variable is a  $\chi^2$  random variable with one degree of freedom. We square the sample autocorrelations  $\hat{\rho}(\tau)$  so that positive and negative values don’t cancel when we sum across various values of  $\tau$ , as we will soon do.

of freedom of the variables summed, we have shown that the **Box-Pierce Q-statistic**,

$$Q_{BP} = T \sum_{\tau=1}^m \hat{\rho}^2(\tau),$$

is approximately distributed as a  $\chi_m^2$  random variable under the null hypothesis that  $y$  is white noise.<sup>12</sup> A slight modification of this, designed to follow more closely the  $\chi^2$  distribution in small samples, is

$$Q_{LB} = T(T+2) \sum_{\tau=1}^m \left( \frac{1}{T-\tau} \right) \hat{\rho}^2(\tau).$$

Under the null hypothesis that  $y$  is white noise,  $Q_{LB}$  is approximately distributed as a  $\chi_m^2$  random variable. Note that the **Ljung-Box Q-statistic** is the same as the Box-Pierce  $Q$  statistic, except that the sum of squared autocorrelations is replaced by a weighted sum of squared autocorrelations, where the weights are  $(T+2)/(T-\tau)$ . For moderate and large  $T$ , the weights are approximately 1, so that the Ljung-Box statistic differs little from the Box-Pierce statistic.

Selection of  $m$  is done to balance competing criteria. On one hand, we don't want  $m$  too small, because after all, we're trying to do a joint test on a large part of the autocorrelation function. On the other hand, as  $m$  grows relative to  $T$ , the quality of the distributional approximations we've invoked deteriorates. In practice, focusing on  $m$  in the neighborhood of  $\sqrt{T}$  is often reasonable.

### 6.3.3 Sample Partial Autocorrelations

Recall that the partial autocorrelations are obtained from population linear regressions, which correspond to a thought experiment involving linear regression using an infinite sample of data. The sample partial autocorrelations

---

<sup>12</sup> $m$  is a maximum displacement selected by the user. Shortly we'll discuss how to choose it.

correspond to the same thought experiment, except that the linear regression is now done on the (feasible) sample of size  $T$ . If the fitted regression is

$$\hat{y}_t = \hat{c} + \hat{\beta}_1 y_{t-1} + \dots + \hat{\beta}_\tau y_{t-\tau},$$

then the **sample partial autocorrelation** at displacement  $\tau$  is

$$\hat{p}(\tau) \equiv \hat{\beta}_\tau.$$

Distributional results identical to those we discussed for the sample autocorrelations hold as well for the sample *partial* autocorrelations. That is, if the series is white noise, approximately 95% of the sample partial autocorrelations should fall in the interval  $\pm 2/\sqrt{T}$ . As with the sample autocorrelations, we typically plot the sample partial autocorrelations along with their two-standard-error bands.

A “**correlogram analysis**” simply means examination of the sample autocorrelation and partial autocorrelation functions (with two standard error bands), together with related diagnostics, such as  $Q$  statistics.

We don’t show the sample autocorrelation or partial autocorrelation at displacement 0, because as we mentioned earlier, they equal 1.0, by construction, and therefore convey no useful information. We’ll adopt this convention throughout.

Note that the sample autocorrelation and partial autocorrelation are identical at displacement 1. That’s because at displacement 1, there are no earlier lags to control for when computing the sample partial autocorrelation, so it equals the sample autocorrelation. At higher displacements, of course, the two diverge.

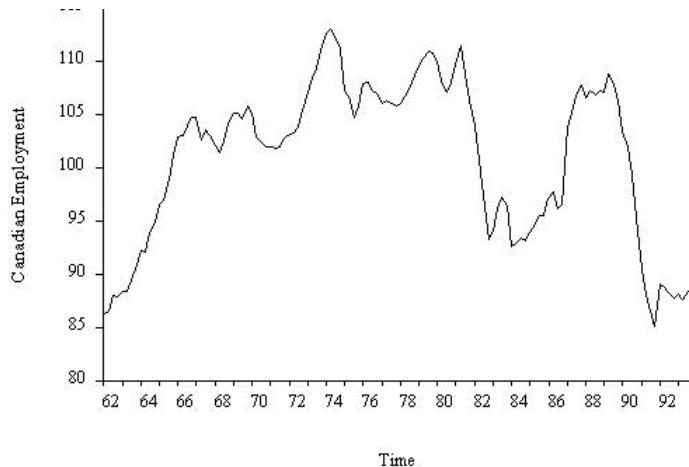


Figure 6.9: Canadian Employment Index

## 6.4 Canadian Employment I: Characterizing Cycles

To illustrate the ideas we've introduced, we examine a quarterly, seasonally-adjusted index of Canadian employment, 1962.1 - 1993.4, which we plot in Figure 6.9. The series displays no trend, and of course it displays no seasonality because it's seasonally adjusted. It does, however, appear highly serially correlated. It evolves in a slow, persistent fashion – high in business cycle booms and low in recessions.

To get a feel for the dynamics operating in the employment series we perform a correlogram analysis.<sup>13</sup> The results appear in Table 1. Consider first the  $Q$  statistic.<sup>14</sup> We compute the  $Q$  statistic and its  $p$ -value under the null hypothesis of white noise for values of  $m$  (the number of terms in the sum that underlies the  $Q$  statistic) ranging from one through twelve. The  $p$ -value is consistently zero to four decimal places, so the null hypothesis of white noise is decisively rejected.

Now we examine the sample autocorrelations and partial autocorrelations. The sample autocorrelations are very large relative to their standard errors

---

<sup>13</sup>A “correlogram analysis” simply means examination of the sample autocorrelation and partial autocorrelation functions (with two standard error bands), together with related diagnostics, such as  $Q$  statistics.

<sup>14</sup>We show the Ljung-Box version of the  $Q$  statistic.

and display slow one-sided decay.<sup>15</sup> The sample partial autocorrelations, in contrast, are large relative to their standard errors at first (particularly for the 1-quarter displacement) but are statistically negligible beyond displacement 2.<sup>16</sup> In Figure 6.10 we plot the sample autocorrelations and partial autocorrelations along with their two standard error bands.

It's clear that employment has a strong cyclical component; all diagnostics reject the white noise hypothesis immediately. Moreover, the sample autocorrelation and partial autocorrelation functions have particular shapes – the autocorrelation function displays slow one-sided damping, while the partial autocorrelation function cuts off at displacement 2. Such patterns, which summarize the dynamics in the series, can be useful for suggesting candidate forecasting models. Such is indeed the case.

---

<sup>15</sup> We don't show the sample autocorrelation or partial autocorrelation at displacement 0, because as we mentioned earlier, they equal 1.0, by construction, and therefore convey no useful information. We'll adopt this convention throughout.

<sup>16</sup> Note that the sample autocorrelation and partial autocorrelation are identical at displacement 1. That's because at displacement 1, there are no earlier lags to control for when computing the sample partial autocorrelation, so it equals the sample autocorrelation. At higher displacements, of course, the two diverge.

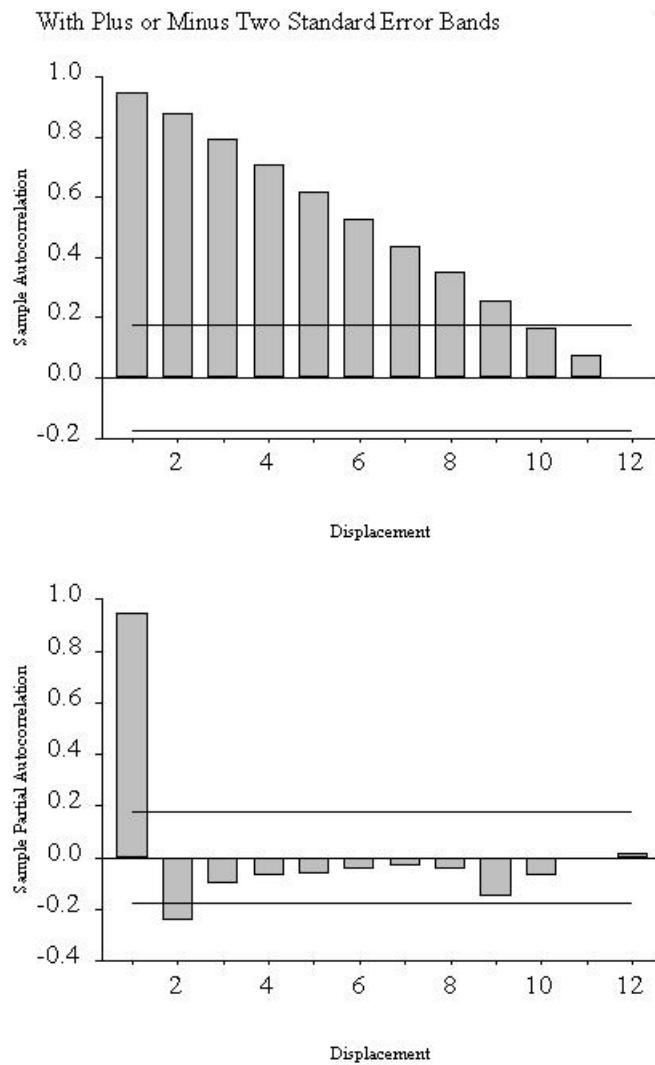


Figure 6.10: Sample Autocorrelation and Sample Partial Autocorrelation

## 6.5 Modeling Cycles With Autoregressions

### 6.5.1 Some Preliminary Notation: The Lag Operator

The **lag operator** and related constructs are the natural language in which time series models are expressed. If you want to understand and manipulate time series models – indeed, even if you simply want to be able to read the software manuals – you have to be comfortable with the lag operator. The lag operator,  $L$ , is very simple: it “operates” on a series by lagging it. Hence  $Ly_t = y_{t-1}$ . Similarly,  $L^2y_t = L(L(y_t)) = L(y_{t-1}) = y_{t-2}$ , and so on. Typically we’ll operate on a series not with the lag operator but with a **polynomial in the lag operator**. A lag operator polynomial of degree  $m$  is just a linear function of powers of  $L$ , up through the  $m$ -th power,

$$B(L) = b_0 + b_1L + b_2L^2 + \dots + b_mL^m.$$

To take a very simple example of a lag operator polynomial operating on a series, consider the  $m$ -th order lag operator polynomial  $L^m$ , for which

$$L^m y_t = y_{t-m}.$$

A well-known operator, the first-difference operator  $\Delta$ , is actually a first-order polynomial in the lag operator; you can readily verify that

$$\Delta y_t = (1 - L)y_t = y_t - y_{t-1}.$$

As a final example, consider the second-order lag operator polynomial  $1 + .9L + .6L^2$  operating on  $y_t$ . We have

$$(1 + .9L + .6L^2)y_t = y_t + .9y_{t-1} + .6y_{t-2},$$

which is a weighted sum, or **distributed lag**, of current and past values. All time-series models, one way or another, must contain such distributed

lags, because they've got to quantify how the past evolves into the present and future; hence lag operator notation is a useful shorthand for stating and manipulating time-series models.

Thus far we've considered only finite-order polynomials in the lag operator; it turns out that infinite-order polynomials are also of great interest. We write the infinite-order lag operator polynomial as

$$B(L) = b_0 + b_1 L + b_2 L^2 + \dots = \sum_{i=0}^{\infty} b_i L^i.$$

Thus, for example, to denote an infinite distributed lag of current and past shocks we might write

$$B(L)\varepsilon_t = b_0\varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \dots = \sum_{i=0}^{\infty} b_i\varepsilon_{t-i}.$$

At first sight, infinite distributed lags may seem esoteric and of limited practical interest, because models with infinite distributed lags have infinitely many parameters ( $b_0, b_1, b_2, \dots$ ) and therefore can't be estimated with a finite sample of data. On the contrary, and surprisingly, it turns out that models involving infinite distributed lags are central to time series modeling, as we shall soon see in detail.

### 6.5.2 Autoregressive Processes

Here we emphasize a very important model of cycles, the **autoregressive (AR) model**.

We begin by characterizing the autocorrelation function and related quantities under the assumption that the *AR* model is the DGP.<sup>17</sup> These characterizations have nothing to do with data or estimation, but they're crucial for developing a basic understanding of the properties of the models, which

---

<sup>17</sup>Sometimes we call time series models of cycles “time series processes,” which is short for **stochastic processes**.

is necessary to perform intelligent modeling. They enable us to make statements such as “If the data were really generated by an autoregressive process, then we’d expect its autocorrelation function to have property x.” Armed with that knowledge, we use the *sample* autocorrelations and partial autocorrelations, in conjunction with the *AIC* and the *SIC*, to suggest candidate models, which we then estimate.

The autoregressive process is a natural time-series model of cycles. It’s simply a *stochastic difference equation*, a simple mathematical model in which the current value of a series is linearly related to its past values, plus an additive stochastic shock. Stochastic difference equations are a natural vehicle for discrete-time stochastic dynamic modeling.

### 6.5.3 Autoregressive Disturbances and Lagged Dependent Variables

You already know the first-order autoregressive (*AR(1)*) model as a model of cyclical dynamics in regression disturbances. Recall, in particular the Durbin-Watson environment that we introduced earlier in Chapter 3:

$$y_t = x'_t \beta + \varepsilon_t$$

$$\varepsilon_t = \phi \varepsilon_{t-1} + v_t$$

$$\begin{aligned} & \text{iid} \\ v_t & \sim N(0, \sigma^2). \end{aligned}$$

To strip things to their essentials, suppose that the only regressor is an intercept.<sup>18</sup> Then we have:

$$y_t = \mu + \varepsilon_t \tag{6.1}$$

$$\varepsilon_t = \phi \varepsilon_{t-1} + v_t$$

---

<sup>18</sup>In later chapters we’ll bring in trends, seasonals, and other standard “*x* variables.”

$$v_t \sim iid(0, \sigma^2).$$

Now let us manipulate this “regression with serially-correlated disturbances” as follows. Because

$$y_t = \mu + \varepsilon_t,$$

we have

$$y_{t-1} = \mu + \varepsilon_{t-1},$$

so

$$\phi y_{t-1} = \phi\mu + \phi\varepsilon_{t-1}. \quad (6.2)$$

Subtracting 6.2 from 6.1 produces

$$y_t - \phi y_{t-1} = \mu(1 - \phi) + (\varepsilon_t - \phi\varepsilon_{t-1}),$$

or

$$y_t = \mu(1 - \phi) + \phi y_{t-1} + v_t.$$

Hence we have arrived at a model of “regression a lagged dependent variable with iid disturbances.” The two models are mathematically identical. LDV with classical disturbances does the same thing as no LDV with serially-correlated disturbances. Each approach “mops up” serial correlation not explained by other regressors. (And in this extreme case, there are no other regressors.)

In this chapter we’ll focus on univariate models with LDV’s, and again, to isolate the relevant issues we’ll focus on models with *only* LDV’s. Later, in Chapter 16, we’ll add  $x$ ’s as well.

### The AR(1) Process for Observed Series

The first-order autoregressive process, *AR*(1) for short, is

$$y_t = \phi y_{t-1} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

In lag operator form, we write

$$(1 - \phi L)y_t = \varepsilon_t.$$

In Figure 6.11 we show simulated realizations of length 150 of two  $AR(1)$  processes; the first is

$$y_t = .4y_{t-1} + \varepsilon_t,$$

and the second is

$$y_t = .95y_{t-1} + \varepsilon_t,$$

where in each case

$$\varepsilon_t \sim iidN(0, 1),$$

and the same innovation sequence underlies each realization. The fluctuations in the  $AR(1)$  with parameter  $\phi = .95$  appear much more persistent than those of the  $AR(1)$  with parameter  $\phi = .4$ . Thus the  $AR(1)$  model is capable of capturing highly persistent dynamics.

A certain condition involving the autoregressive lag operator polynomial must be satisfied for an autoregressive process to be covariance stationary. The condition is that all roots of the autoregressive lag operator polynomial must be outside the unit circle. In the  $AR(1)$  case we have

$$(1 - \phi L)y_t = \varepsilon_t,$$

so the autoregressive lag operator polynomial is  $1 - \phi L$ , with root  $1/\phi$ . Hence the  $AR(1)$  process is covariance stationary if  $|\phi| < 1$ .

Let's investigate the moment structure of the  $AR(1)$  process. If we begin with the  $AR(1)$  process,

$$y_t = \phi y_{t-1} + \varepsilon_t,$$

and substitute backward for lagged  $y$ 's on the right side, we obtain the so-

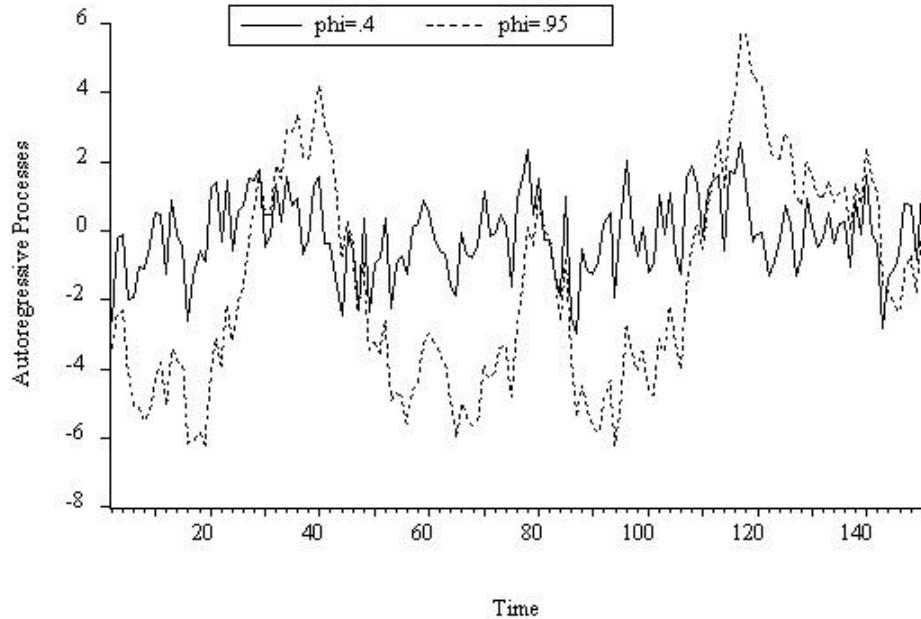


Figure 6.11: Realizations of Two AR(1) Processes

called “**moving-average representation**”

$$y_t = \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \dots$$

The existence of a moving-average representation is very intuitive. Ultimately the  $\varepsilon$ 's are the only things that move  $y$ , so it is natural that we should be able to express  $y$  in terms of the history of  $\varepsilon$ . We will have much more to say about that in Chapter 7. The existence of a moving-average representation is also very useful, because it facilitates some important calculations, to which we now turn.

From the moving average representation of the covariance stationary  $AR(1)$

process, we can compute the unconditional mean and variance,

$$E(y_t) = E(\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \dots)$$

$$= E(\varepsilon_t) + \phi E(\varepsilon_{t-1}) + \phi^2 E(\varepsilon_{t-2}) + \dots$$

$$= 0$$

and

$$var(y_t) = var(\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \dots)$$

$$= \sigma^2 + \phi^2\sigma^2 + \phi^4\sigma^2 + \dots$$

$$= \sigma^2 \sum_{i=0}^{\infty} \phi^{2i}$$

$$= \frac{\sigma^2}{1-\phi^2}.$$

The conditional moments, in contrast, are

$$E(y_t|y_{t-1}) = E(\phi y_{t-1} + \varepsilon_t|y_{t-1})$$

$$= \phi E(y_{t-1}|y_{t-1}) + E(\varepsilon_t|y_{t-1})$$

$$= \phi y_{t-1} + 0$$

$$= \phi y_{t-1}$$

and

$$\begin{aligned}
 var(y_t|y_{t-1}) &= var((\phi y_{t-1} + \varepsilon_t)|y_{t-1}) \\
 &= \phi^2 var(y_{t-1}|y_{t-1}) + var(\varepsilon_t|y_{t-1}) \\
 &= 0 + \sigma^2 \\
 &= \sigma^2.
 \end{aligned}$$

Note in particular that the simple way that the conditional mean adapts to the changing information set as the process evolves.

To find the autocovariances, we proceed as follows. The process is

$$y_t = \phi y_{t-1} + \varepsilon_t,$$

so that multiplying both sides of the equation by  $y_{t-\tau}$  we obtain

$$y_t y_{t-\tau} = \phi y_{t-1} y_{t-\tau} + \varepsilon_t y_{t-\tau}.$$

For  $\tau \geq 1$ , taking expectations of both sides gives

$$\gamma(\tau) = \phi \gamma(\tau - 1).$$

This is called the **Yule-Walker equation**. It is a recursive equation; that is, given  $\gamma(\tau)$ , for any  $\tau$ , the Yule-Walker equation immediately tells us how to get  $\gamma(\tau + 1)$ . If we knew  $\gamma(0)$  to start things off (an “initial condition”), we could use the Yule-Walker equation to determine the entire autocovariance sequence. And we *do* know  $\gamma(0)$ ; it’s just the variance of the process, which we already showed to be

$$\gamma(0) = \frac{\sigma^2}{1 - \phi^2}.$$

Thus we have

$$\begin{aligned}\gamma(0) &= \frac{\sigma^2}{1 - \phi^2} \\ \gamma(1) &= \phi \frac{\sigma^2}{1 - \phi^2} \\ \gamma(2) &= \phi^2 \frac{\sigma^2}{1 - \phi^2},\end{aligned}$$

and so on. In general, then,

$$\gamma(\tau) = \phi^\tau \frac{\sigma^2}{1 - \phi^2}, \tau = 0, 1, 2, \dots$$

Dividing through by  $\gamma(0)$  gives the autocorrelations,

$$\rho(\tau) = \phi^\tau, \tau = 0, 1, 2, \dots$$

Note the gradual autocorrelation decay, which is typical of autoregressive processes. The autocorrelations approach zero in the limit as the displacement approaches infinity. If  $\phi$  is positive, the autocorrelation decay is one-sided. If  $\phi$  is negative, the decay involves back-and-forth oscillations. The relevant case in business and economics is  $\phi > 0$ , but either way, the autocorrelations damp gradually. In Figure 6.12 and 6.13 we show the autocorrelation functions for  $AR(1)$  processes with parameters  $\phi = .4$  and  $\phi = .95$ . The persistence is much stronger when  $\phi = .95$ .

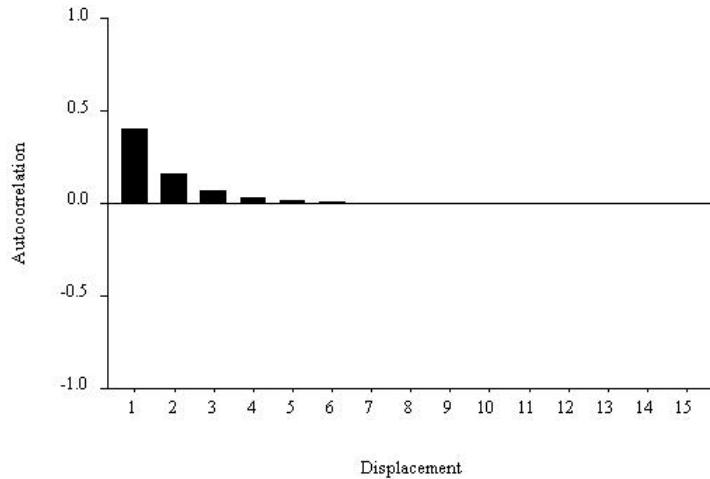
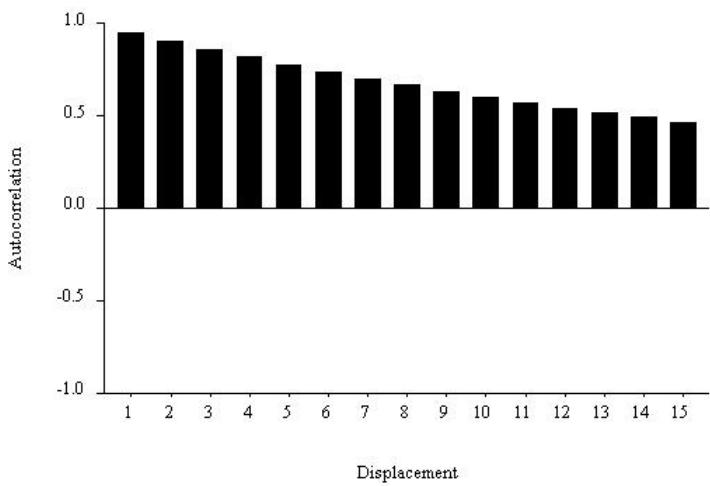
Finally, the partial autocorrelation function for the  $AR(1)$  process cuts off abruptly; specifically,

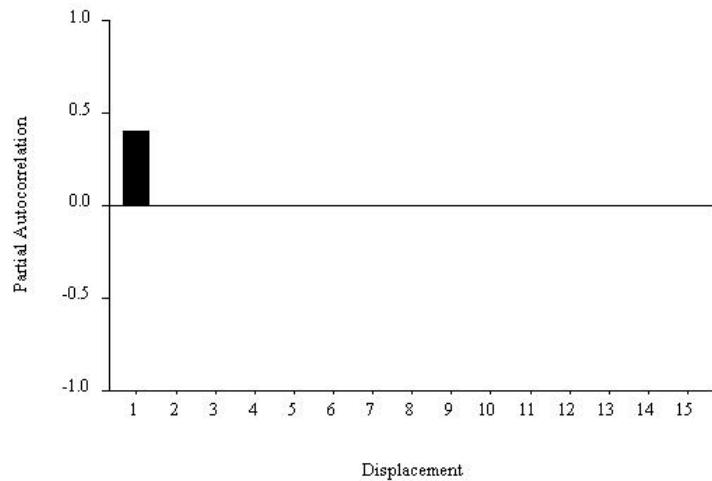
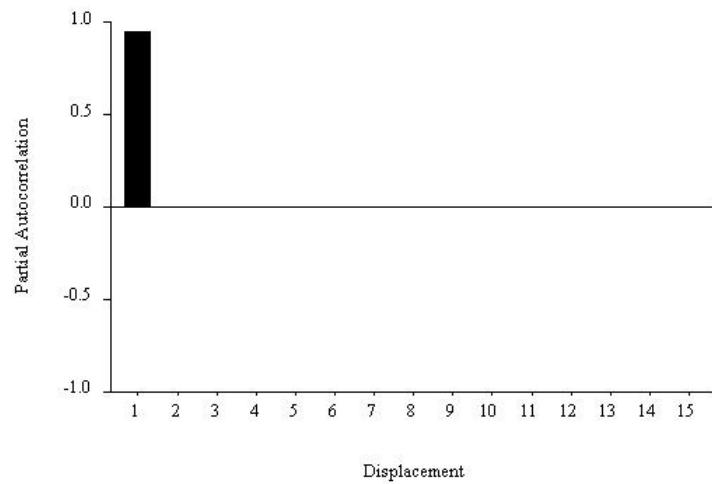
$$p(\tau) = \begin{cases} \phi, & \tau = 1 \\ 0, & \tau > 1. \end{cases}.$$

It's easy to see why. The partial autocorrelations are just the last coefficients in a sequence of successively longer population autoregressions. If the true process is in fact an  $AR(1)$ , the first partial autocorrelation is just the

autoregressive coefficient, and coefficients on all longer lags are zero.

In Figures 6.14 and 6.15 we show the partial autocorrelation functions for our two  $AR(1)$  processes. At displacement 1, the partial autocorrelations are simply the parameters of the process (.4 and .95, respectively), and at longer displacements, the partial autocorrelations are zero.

Figure 6.12: Population Autocorrelation Function:  $\rho = .4$ Figure 6.13: Population Autocorrelation Function:  $\rho = .95$

Figure 6.14: Partial Autocorrelation Function:  $\rho = .4$ Figure 6.15: Partial Autocorrelation Function:  $\rho = .95$

### 6.5.4 The $AR(p)$ Process

The general  $p$ -th order autoregressive process, or  $AR(p)$  for short, is

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

In lag operator form we write

$$\Phi(L)y_t = (1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p)y_t = \varepsilon_t.$$

In our discussion of the  $AR(p)$  process we dispense with mathematical derivations and instead rely on parallels with the  $AR(1)$  case to establish intuition for its key properties.

An  $AR(p)$  process is covariance stationary if and only if all roots of the autoregressive lag operator polynomial  $\Phi(L)$  are outside the unit circle.<sup>19</sup>

The autocorrelation function for the general  $AR(p)$  process, as with that of the  $AR(1)$  process, decays gradually with displacement. Finally, the  $AR(p)$  partial autocorrelation function has a sharp cutoff at displacement  $p$ , for the same reason that the  $AR(1)$  partial autocorrelation function has a sharp cutoff at displacement 1.

Let's discuss the  $AR(p)$  autocorrelation function in a bit greater depth. The key insight is that, in spite of the fact that its qualitative behavior (gradual damping) matches that of the  $AR(1)$  autocorrelation function, it can nevertheless display a richer variety of patterns, depending on the order and parameters of the process. It can, for example, have damped monotonic decay, as in the  $AR(1)$  case with a positive coefficient, but it can also have damped oscillation in ways that  $AR(1)$  can't have. In the  $AR(1)$  case, the only possible oscillation occurs when the coefficient is negative, in which case

---

<sup>19</sup>A necessary condition for covariance stationarity, which is often useful as a quick check, is  $\sum_{i=1}^p \phi_i < 1$ . If the condition is satisfied, the process may or may not be stationary, but if the condition is violated, the process can't be stationary.

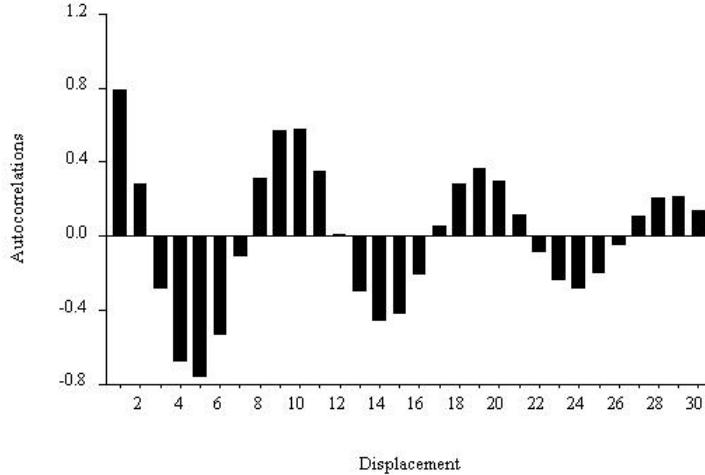


Figure 6.16: Autocorrelation Function of AR(2) with Complex Roots

the autocorrelations switch signs at each successively longer displacement. In higher-order autoregressive models, however, the autocorrelations can oscillate with much richer patterns reminiscent of cycles in the more traditional sense. This occurs when some roots of the autoregressive lag operator polynomial are complex.<sup>20</sup> Consider, for example, the  $AR(2)$  process,

$$y_t = 1.5y_{t-1} - .9y_{t-2} + \varepsilon_t.$$

The corresponding lag operator polynomial is  $1 - 1.5L + .9L^2$ , with two complex conjugate roots,  $.83 \pm .65i$ . The inverse roots are  $.75 \pm .58i$ , both of which are close to, but inside, the unit circle; thus the process is covariance stationary. It can be shown that the autocorrelation function for an  $AR(2)$  process is

$$\rho(0) = 1$$

$$\rho(\tau) = \phi_1\rho(\tau - 1) + \phi_2\rho(\tau - 2), \tau = 2, 3, \dots$$

$$\rho(1) = \frac{\phi_1}{1 - \phi_2}$$

Using this formula, we can evaluate the autocorrelation function for the

---

<sup>20</sup>Note that complex roots can't occur in the  $AR(1)$  case.

process at hand; we plot it in Figure 6.16. Because the roots are complex, the autocorrelation function oscillates, and because the roots are close to the unit circle, the oscillation damps slowly.

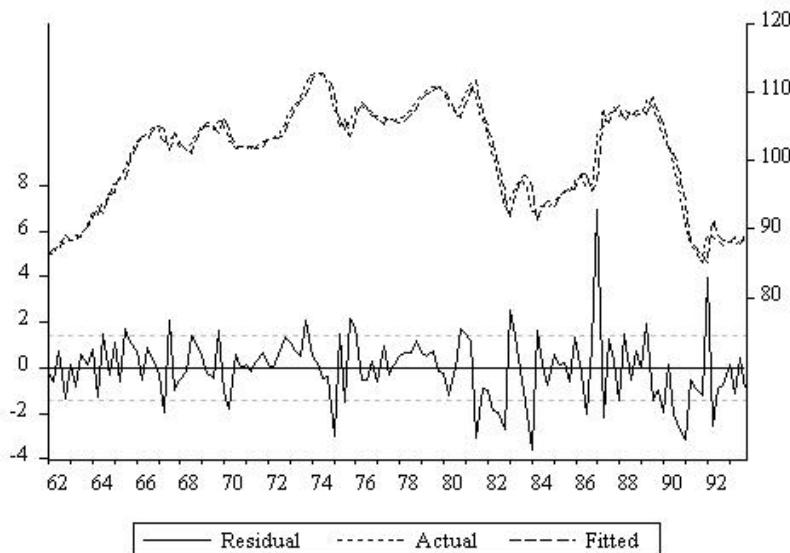
## 6.6 Canadian Employment II: Modeling Cycles

The sum of squares function for autoregressive processes is linear in the parameters, so that estimation is particularly stable and easy – just standard OLS regressions. In the  $AR(1)$  case, we simply run an ordinary least squares regression of  $y$  on one lag of  $y$ ; in the  $AR(p)$  case, we regress  $y$  on  $p$  lags of  $y$ .

We estimate  $AR(p)$  models,  $p = 1, 2, 3, 4$ . Both the  $AIC$  and the  $SIC$  suggest that the  $AR(2)$  is best. To save space, we report only the results of  $AR(2)$  estimation in Table 6.17a. The estimation results look good, and the residuals (Figure 6.17b) look like white noise. The residual correlogram (Table 6.18, Figure 6.19) supports that conclusion.

LS // Dependent Variable is CANEMP				
Sample: 1962:1 1993:4				
Included observations: 128				
Convergence achieved after 3 iterations				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	101.2413	3.399620	29.78017	0.0000
AR(1)	1.438810	0.078487	18.33188	0.0000
AR(2)	-0.476451	0.077902	-6.116042	0.0000
R-squared	0.963372	Mean dependent var	101.0176	
Adjusted R-squared	0.962786	S.D. dependent var	7.499163	
S.E. of regression	1.446663	Akaike info criterion	0.761677	
Sum squared resid	261.6041	Schwarz criterion	0.828522	
Log likelihood	-227.3715	F-statistic	1643.837	
Durbin-Watson stat	2.067024	Prob(F-statistic)	0.000000	
Inverted AR Roots	.92	.52		

(a) Employment: AR(2) Model



(b) Employment: AR(2) Model, Residual Plot

Figure 6.17: Employment: AR(2) Model

Sample: 1962:1 1993:4  
 Included observations: 128  
 Q-statistic probabilities adjusted for 2 ARMA term(s)

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	-0.035	-0.035	.088	0.1606	
2	0.044	0.042	.088	0.4115	
3	0.011	0.014	.088	0.4291 0.512	
4	0.051	0.050	.088	0.7786 0.678	
5	0.002	0.004	.088	0.7790 0.854	
6	0.019	0.015	.088	0.8272 0.935	
7	-0.024	-0.024	.088	0.9036 0.970	
8	0.078	0.072	.088	1.7382 0.942	
9	0.080	0.087	.088	2.6236 0.918	
10	0.050	0.050	.088	2.9727 0.936	
11	-0.023	-0.027	.088	3.0504 0.962	
12	-0.129	-0.148	.088	5.4385 0.860	

Figure 6.18: Employment: AR(2) Model, Residual Correlogram

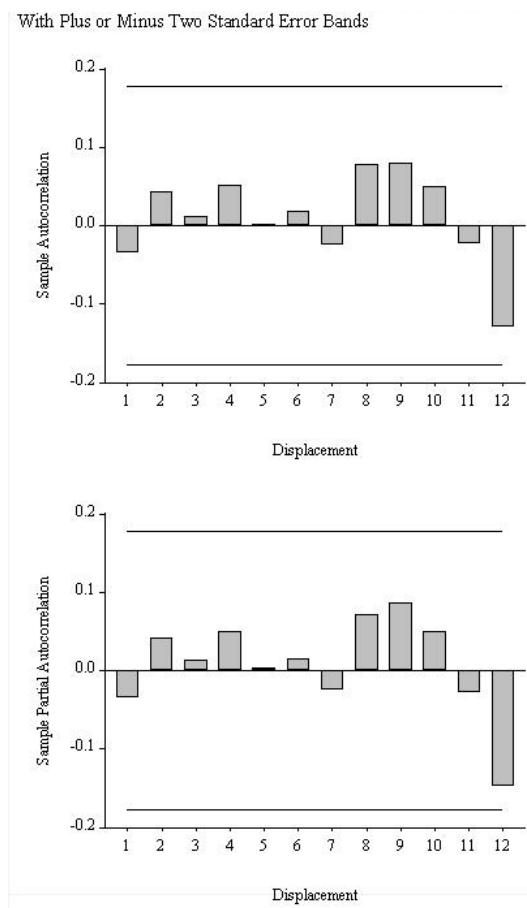


Figure 6.19: Employment: AR(2) Model, Residual Sample Autocorrelation and Partial Autocorrelation

## 6.7 Forecasting Cycles with Autoregressions

### 6.7.1 On the FRV Problem

We have seen that the FRV problem arises in general, but not in cross sections, and not in deterministic-trend time-series environments, and not in deterministic-seasonal time-series environments. The same is true in certain other time-series environments.

In particular, forget about trends and seasonals for the moment. Still the FRV problem does not arise if the RHS variables are *lagged* sufficiently relative the the forecast horizon of interest. Suppose, for example, that an acceptable model is

$$y_t = \beta_1 + \beta_2 x_{t-1} + \varepsilon_t. \quad (6.3)$$

The RHS variable is lagged by one period, so model 6.3 is immediately usable for 1-step-ahead forecasting without the FRV problem. More lags of  $x$  can of course be included; the key for 1-step-ahead forecasting is that all variables on the right be lagged by at least one period.

Forecasting more than one step ahead in model 6.3, however, would appear to lead again to the FRV problem: If we want to forecast  $h$  steps ahead, then all variables on the right must be lagged by at least  $h$  periods, not just by 1 period. Perhaps surprisingly, it actually remains *easy* to circumvent the FRV problem in autoregressive models. For example, models with  $y_t \rightarrow y_{t-1}$  or  $y_t \rightarrow y_{t-1}, x_{t-1}$  can effectively be transformed to models with  $y_t \rightarrow y_{t-h}$  or  $y_t \rightarrow y_{t-h}, x_{t-h}$ , as we will see in this section.

### 6.7.2 Information Sets, Conditional Expectations, and Linear Projections

By now you've gotten comfortable with the idea of an **information set**. Here we'll use that idea extensively. We denote the time- $T$  information set

by  $\Omega_T$ . Think of the information set as containing the available past history of the series,

$$\Omega_T = \{y_T, y_{T-1}, y_{T-2}, \dots\},$$

where for theoretical purposes we imagine history as having begun in the infinite past.

Based upon that information set, we want to find the **optimal forecast** of  $y$  at some future time  $T + h$ . The optimal forecast is the one with the smallest loss on average, that is, the forecast that minimizes **expected loss**. It turns out that under reasonably weak conditions the optimal forecast is the **conditional mean**,

$$E(y_{T+h}|\Omega_T),$$

the expected value of the future value of the series being forecast, conditional upon available information.

In general, the conditional mean need not be a linear function of the elements of the information set. Because linear functions are particularly tractable, we prefer to work with **linear forecasts** – forecasts that are linear in the elements of the information set – by finding the best linear approximation to the conditional mean, called the **linear projection**, denoted

$$P(y_{T+h}|\Omega_T).$$

This explains the common term “**linear least squares forecast**.” The linear projection is often very useful and accurate, because the conditional mean is often close to linear. In fact, in the Gaussian case the conditional expectation is exactly linear, so that

$$E(y_{T+h}|\Omega_T) = P(y_{T+h}|\Omega_T).$$

### 6.7.3 Point Forecasts for Autoregressions: Wold's Chain Rule

A very simple recursive method for computing optimal  $h$ -step-ahead point forecasts, for any desired  $h$ , is available for autoregressions.

The recursive method, called the **chain rule of forecasting**, is best learned by example. Consider the  $AR(1)$  process,

$$y_t = \phi y_{t-1} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

First we construct the optimal 1-step-ahead forecast, and then we construct the optimal 2-step-ahead forecast, which depends on the optimal 1-step-ahead forecast, which we've already constructed. Then we construct the optimal 3-step-ahead forecast, which depends on the already-computed 2-step-ahead forecast, which we've already constructed, and so on.

To construct the 1-step-ahead forecast, we write out the process for time  $T + 1$ ,

$$y_{T+1} = \phi y_T + \varepsilon_{T+1}.$$

Then, projecting the right-hand side on the time- $T$  information set, we obtain

$$y_{T+1,T} = \phi y_T.$$

Now let's construct the 2-step-ahead forecast. Write out the process for time  $T + 2$ ,

$$y_{T+2} = \phi y_{T+1} + \varepsilon_{T+2}.$$

Then project directly on the time- $T$  information set to get

$$y_{T+2,T} = \phi y_{T+1,T}.$$

Note that the future innovation is replaced by 0, as always, and that we have directly replaced the time  $T + 1$  value of  $y$  with its earlier-constructed optimal

forecast. Now let's construct the 3-step-ahead forecast. Write out the process for time  $T + 3$ ,

$$y_{T+3} = \phi y_{T+2} + \varepsilon_{T+3}.$$

Then project directly on the time- $T$  information set,

$$y_{T+3,T} = \phi y_{T+2,T}.$$

The required 2-step-ahead forecast was already constructed.

Continuing in this way, we can recursively build up forecasts for any and all future periods. Hence the name “chain rule of forecasting.” Note that, for the  $AR(1)$  process, only the most recent value of  $y$  is needed to construct optimal forecasts, for any horizon, and for the general  $AR(p)$  process only the  $p$  most recent values of  $y$  are needed. In particular, for our  $AR(1)$  case,

$$y_{T+h,T} = \phi^h y_T.$$

As usual, in truth the parameters are unknown and so must be estimated, so we turn infeasible forecasts into feasible (“operational”) forecasts by inserting the usual estimates where unknown parameters appear.

It is worth noting that thanks to Wold’s chain rule we have now solved the FRV problem for autoregressions, as we did earlier for cross sections, trends, and seasonals! We have of course worked through the calculations in detail only for the  $AR(1)$  case, but the approach is identical for the general  $AR(p)$  case.

#### 6.7.4 Density Forecasts

The chain rule is a device for simplifying the computation of point forecasts. Density forecasts require a bit more work. Let us again work through the  $AR(1)$  case in detail, assuming normality and ignoring parameter estimation uncertainty.

We know that

$$y_{T+h} \sim N(y_{T+h,T}, \sigma_h^2),$$

where  $\sigma_h^2 = \text{var}(y_{T+h} | \Omega_T)$  and  $\Omega_T = \{y_T, y_{T-1}, \dots\}$ . Using Wold's chain rule we already derived the formula for  $y_{T+h,T}$ , so all we need is the  $h$ -step-ahead forecast error variance,  $\sigma_h^2$ .

First let us simply assert the general result. It is

$$\sigma_h^2 = \sigma^2 \sum_{i=0}^{h-1} \phi^{2i}.$$

Now let us derive the general result. First recall that the optimal forecasts are

$$y_{T+1,T} = \phi y_T$$

$$y_{T+2,T} = \phi^2 y_T$$

$$y_{T+h,T} = \phi^h y_T.$$

Second, note that the corresponding forecast errors are

$$e_{T+1,T} = (y_{T+1} - y_{T+1,T}) = \varepsilon_{T+1}$$

$$e_{T+2,T} = (y_{T+2} - y_{T+2,T}) = \phi \varepsilon_{T+1} + \varepsilon_{T+2}$$

$$e_{T+h,T} = (y_{T+h} - y_{T+h,T}) = \varepsilon_{T+h} + \phi \varepsilon_{T+h-1} + \dots + \phi^{h-1} \varepsilon_{T+1}.$$

Third, note that the corresponding forecast error variances are

$$\sigma_1^2 = \sigma^2$$

$$\sigma_2^2 = \sigma^2(1 + \phi^2)$$

$$\sigma_h^2 = \sigma^2 \sum_{i=0}^{h-1} \phi^{2i}.$$

QED

Note that the limiting  $h$ -step-ahead forecast error variance is

$$\lim_{h \rightarrow \infty} \sigma_h^2 = \frac{\sigma^2}{1 - \phi^2},$$

the unconditional variance of the  $AR(1)$  process. (The conditioning information becomes progressively less valuable as  $h \rightarrow \infty$  in covariance stationary environments, so the conditional variance converges to the unconditional variance.)

As usual, in truth the parameters are unknown and so must be estimated, so we turn infeasible forecasts into feasible (“operational”) forecasts by inserting the usual estimates where unknown parameters appear. In addition, and also as usual, we can account for non-normality and parameter-estimation uncertainty using simulation methods. (Of course simulation could be used even under normality).

Density forecasts for higher-ordered autoregressions proceed in similar fashion. Point forecasts at any horizon come from Wold’s chain rule. Under normality we still need the corresponding  $h$ -step forecast-error variances, we infer from the moving-average representation. Dropping normality and using simulation methods does not even require the variance calculation.

## 6.8 Canadian Employment III: Forecasting

Now we put our forecasting technology to work to produce autoregressive point and interval forecasts for Canadian employment. Recall that the best autoregressive model was an  $AR(2)$ . In Figure 6.20 we show the 4-quarter-ahead extrapolation forecast, which reverts to the unconditional mean much less quickly, as seems natural given the high persistence of employment. The 4-quarter-ahead point forecast, in fact, is still well below the mean. Sim-

ilarly, the 95% error bands grow gradually and haven't approached their long-horizon values by four quarters out.

Figures 6.20 and 6.21 make clear the nature of the autoregressive forecasts. In Figure 6.21 we show the employment history, 4-quarter-ahead  $AR(2)$  extrapolation forecast, and the realization. The  $AR(2)$  forecast appears quite accurate; the mean squared forecast error is 1.3.

Figure 6.22 presents the 12-step-ahead extrapolation forecast, and Figure 6.23 presents a much longer-horizon extrapolation forecast. Eventually the unconditional mean *is* approached, and eventually the error bands do go flat, but only for very long-horizon forecasts, due to the high persistence in employment, which the  $AR(2)$  model captures.

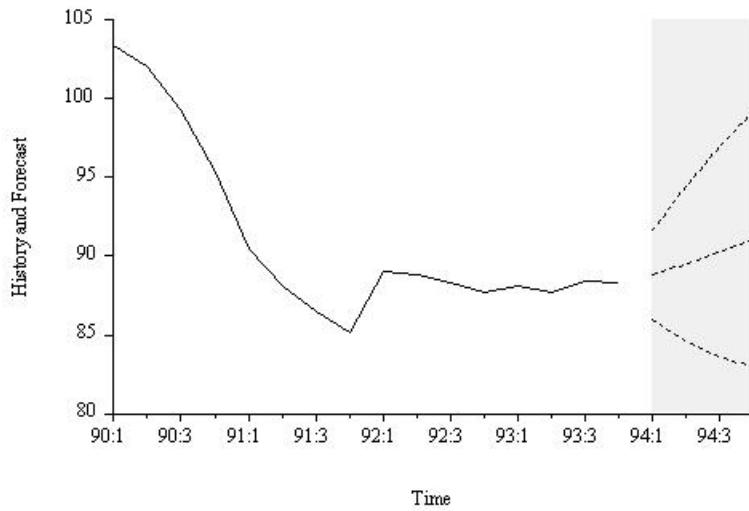


Figure 6.20: Employment History and Forecast: AR(2)

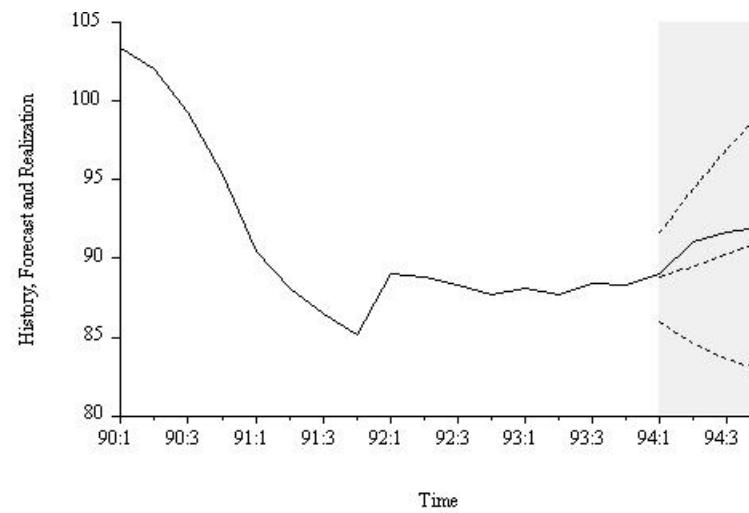


Figure 6.21: Employment History, Forecast, and Realization: AR(2)

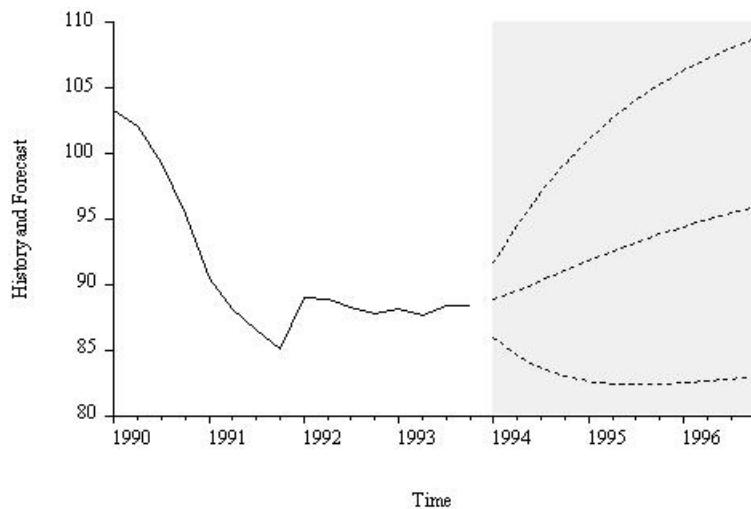


Figure 6.22: Employment History and Long-Horizon Forecast: AR(2)

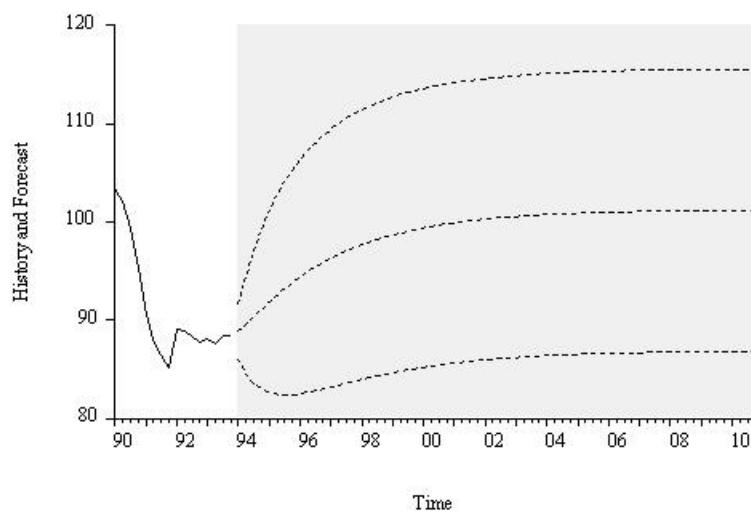


Figure 6.23: Employment History and Very Long-Horizon Forecast: AR(2)

## 6.9 Exercises, Problems and Complements

1. From FRED get Industrial Production, Total Index, 2012=100, Quarterly, Not Seasonally Adjusted, 1919:Q1-latest. First, hold out 2014.1-latest. Select and estimate your preferred model (deterministic trend + deterministic seasonal + autoregressive cyclical dynamics) using 1919:Q1-2013:Q4, and use your estimated model to generate a path forecast 2014:Q1-latest. Second, hold out nothing. Re-select and re-estimate using 1919:Q1-latest, and use your estimated model to generate a path forecast for the next eight quarters.
2. More on the stability condition for  $AR(1)$  processes.

The key stability condition is  $|\phi| < 1$ . Recall  $y_t = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$ . This implies that  $\text{var}(y_t) = \sum_{j=0}^{\infty} \phi^{2j} \sigma^2$ , which is the sum of a geometric series. Hence:

$$\text{var}(y_t) = \frac{\sigma^2}{1 - \phi^2} \text{ if } |\phi| < 1$$

$$\text{var}(y_t) = \infty \text{ otherwise}$$

3. A more complete picture of  $AR(1)$  stability.

The following are all aspects in which covariance stationarity corresponds to a nice, stable environment.

- (a) Series  $y_t$  is persistent but eventually reverts to a fixed mean.
- (b) Shocks  $\varepsilon_t$  have persistent effects but eventually die out. Hint: Consider  $y_t = \mu + \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$ ,  $|\phi| < 1$ .
- (c) Autocorrelations  $\rho(\tau)$  nonzero but decay to zero.
- (d) Autocorrelations  $\rho(\tau)$  depend on  $\tau$  (of course) but not on time. Hint: Use back substitution to relate  $y_t$  and  $y_{t-2}$ . How does it compare to the relation between  $y_t$  and  $y_{t-1}$  when  $|\phi| < 1$ ?

- (e) Series  $y_t$  varies but not too extremely. Hint: Consider  $\text{var}(y_t) = \frac{\sigma^2}{1-\phi^2}$ ,  $|\phi| < 1$ .

4. Autocorrelation functions of covariance stationary series.

While interviewing at a top investment bank, your interviewer is impressed by the fact that you have taken a course on forecasting. She decides to test your knowledge of the autocovariance structure of covariance stationary series and lists five autocovariance functions:

- a.  $\gamma(t, \tau) = \alpha$
- b.  $\gamma(t, \tau) = e^{-\alpha\tau}$
- c.  $\gamma(t, \tau) = \alpha\tau$
- d.  $\gamma(t, \tau) = \frac{\alpha}{\tau}$ , where  $\alpha$  is a positive constant. Which autocovariance function(s) are consistent with covariance stationarity, and which are not? Why?

5. Autocorrelation vs. partial autocorrelation.

Describe the difference between autocorrelations and partial autocorrelations. How can autocorrelations at certain displacements be positive while the partial autocorrelations at those same displacements are negative?

6. Sample autocorrelation functions of trending series.

A tell-tale sign of the slowly-evolving nonstationarity associated with trend is a sample autocorrelation function that damps extremely slowly.

- a. Find three trending series, compute their sample autocorrelation functions, and report your results. Discuss.
- b. Fit appropriate trend models, obtain the model residuals, compute their sample autocorrelation functions, and report your results. Discuss.

7. Sample autocorrelation functions of seasonal series.

A tell-tale sign of seasonality is a sample autocorrelation function with sharp peaks at the seasonal displacements (4, 8, 12, etc. for quarterly data, 12, 24, 36, etc. for monthly data, and so on).

- a. Find a series with both trend and seasonal variation. Compute its sample autocorrelation function. Discuss.
- b. Detrend the series. Discuss.
- c. Compute the sample autocorrelation function of the detrended series. Discuss.
- d. Seasonally adjust the detrended series. Discuss.
- e. Compute the sample autocorrelation function of the detrended, seasonally-adjusted series. Discuss.

8. Lag operator expressions, I.

Rewrite the following expressions without using the lag operator.

- a.  $(L^\tau)y_t = \varepsilon_t$
- b.  $y_t = \left( \frac{2+5L+.8L^2}{L-.6L^3} \right) \varepsilon_t$
- c.  $y_t = 2 \left( 1 + \frac{L^3}{L} \right) \varepsilon_t.$

9. Lag operator expressions, II.

Rewrite the following expressions in lag operator form.

- a.  $y_t + y_{t-1} + \dots + y_{t-N} = \alpha + \varepsilon_t + \varepsilon_{t-1} + \dots + \varepsilon_{t-N}$ , where  $\alpha$  is a constant
- b.  $y_t = \varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t.$

10. Simulating time series processes.

Many cutting-edge estimation and forecasting techniques involve simulation. Moreover, simulation is often a good way to get a feel for a model and its behavior. White noise can be simulated on a computer using **random number generators**, which are available in most statistics, econometrics and forecasting packages.

- (a) Simulate a Gaussian white noise realization of length 200. Call the white noise  $\varepsilon_t$ . Compute the correlogram. Discuss.
- (b) Form the distributed lag  $y_t = \varepsilon_t + .9\varepsilon_{t-1}$ ,  $t = 2, 3, \dots, 200$ . Compute the sample autocorrelations and partial autocorrelations. Discuss.
- (c) Let  $y_1 = 1$  and  $y_t = .9y_{t-1} + \varepsilon_t$ ,  $t = 2, 3, \dots, 200$ . Compute the sample autocorrelations and partial autocorrelations. Discuss.

## 11. Diagnostic checking of model residuals.

If a forecasting model has extracted all the systematic information from the data, then what's left – the residual – should be white noise. More precisely, the true innovations are white noise, and if a model is a good approximation to the DGP then its 1-step-ahead forecast errors should be approximately white noise. The model residuals are the in-sample analog of out-of-sample 1-step-ahead forecast errors. Hence the usefulness of various tests of the hypothesis that residuals are white noise.

The Durbin-Watson test is the most popular. Recall the Durbin-Watson test statistic, discussed in Chapter 2,

$$DW = \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T e_t^2}.$$

Note that

$$\sum_{t=2}^T (e_t - e_{t-1})^2 \approx 2 \sum_{t=2}^T e_t^2 - 2 \sum_{t=2}^T e_t e_{t-1}.$$

Thus

$$DW \approx 2(1 - \hat{\rho}(1)),$$

so that the Durbin-Watson test is effectively based only on the first sample autocorrelation and really only tests whether the first autocorrelation is zero. We say therefore that the Durbin-Watson is a test for **first-order serial correlation**. In addition, the Durbin-Watson test is not valid in the presence of lagged dependent variables.<sup>21</sup> On both counts, we'd like a more general and flexible framework for diagnosing serial correlation. The residual correlogram, comprised of the residual sample autocorrelations, the sample partial autocorrelations, and the associated  $Q$  statistics, delivers the goods.

- (a) When we discussed the correlogram in the text, we focused on the case of an observed time series, in which case we showed that the  $Q$  statistics are distributed as  $\chi_m^2$ . Now, however, we want to assess whether unobserved model disturbances are white noise. To do so, we use the model residuals, which are estimates of the unobserved disturbances. Because we fit a model to get the residuals, we need to account for the degrees of freedom used. The upshot is that the distribution of the  $Q$  statistics under the white noise hypothesis is better approximated by a  $\chi_{m-K}^2$  random variable, where  $K$  is the number of parameters estimated. That's why, for example, we don't report (and in fact the software doesn't compute) the  $p$ -values for the  $Q$  statistics associated with the residual correlogram of our employment forecasting model until  $m > K$ .
- (b) **Durbin's  $h$  test** is an alternative to the Durbin-Watson test. As

---

<sup>21</sup>Following standard, if not strictly appropriate, practice, in this book we often report and examine the Durbin-Watson statistic even when lagged dependent variables are included. We always supplement the Durbin-Watson statistic, however, with other diagnostics such as the residual correlogram, which remain valid in the presence of lagged dependent variables, and which almost always produce the same inference as the Durbin-Watson statistic.

with the Durbin-Watson test, it's designed to detect first-order serial correlation, but it's valid in the presence of lagged dependent variables. Do some background reading as well on Durbin's  $h$  test and report what you learned.

- (c) The **Breusch-Godfrey test** is another alternative to the Durbin-Watson test. It's designed to detect  $p^{th}$ -order serial correlation, where  $p$  is selected by the user, and is also valid in the presence of lagged dependent variables. Do some background reading on the Breusch-Godfrey procedure and report what you learned.
- (d) Which do you think is likely to be most useful to you in assessing the properties of residuals from forecasting models: the residual correlogram, Durbin's  $h$  test, or the Breusch-Godfrey test? Why?

## 12. Forecast accuracy across horizons.

You are a consultant to MedTrax, a large pharmaceutical company, which released a new ulcer drug three months ago and is concerned about recovering research and development costs. Accordingly, MedTrax has approached you for drug sales projections at 1- through 12-month-ahead horizons, which it will use to guide potential sales force realignments. In briefing you, MedTrax indicated that it expects your long-horizon forecasts (e.g., 12-month-ahead) to be just as accurate as your short-horizon forecasts (e.g., 1-month-ahead). Explain to MedTrax why that is unlikely, even if you do the best forecasting job possible.

## 13. Forecasting an $AR(1)$ process with known and unknown parameters.

Use the chain rule to forecast the  $AR(1)$  process,

$$y_t = \phi y_{t-1} + \varepsilon_t.$$

For now, assume that all parameters are known.

- a. Show that the optimal forecasts are

$$y_{T+1,T} = \phi y_T$$

$$y_{T+2,T} = \phi^2 y_T$$

$$y_{T+h,T} = \phi^h y_T.$$

- b. Show that the corresponding forecast errors are

$$e_{T+1,T} = (y_{T+1} - y_{T+1,T}) = \varepsilon_{T+1}$$

$$e_{T+2,T} = (y_{T+2} - y_{T+2,T}) = \phi \varepsilon_{T+1} + \varepsilon_{T+2}$$

$$e_{T+h,T} = (y_{T+h} - y_{T+h,T}) = \varepsilon_{T+h} + \phi \varepsilon_{T+h-1} + \dots + \phi^{h-1} \varepsilon_{T+1}.$$

- c. Show that the forecast error variances are

$$\sigma_1^2 = \sigma^2$$

$$\sigma_2^2 = \sigma^2(1 + \phi^2)$$

$$\sigma_h^2 = \sigma^2 \sum_{i=0}^{h-1} \phi^{2i}.$$

- d. Show that the limiting forecast error variance is

$$\lim_{h \rightarrow \infty} \sigma_h^2 = \frac{\sigma^2}{1 - \phi^2},$$

the unconditional variance of the  $AR(1)$  process.

- e. Now assume that the parameters are unknown and so must be estimated. Make your expressions for both the forecasts and the forecast error variances operational, by inserting least squares estimates where unknown parameters appear, and use them to produce an operational

point forecast and an operational 90% interval forecast for  $y_{T+2:T}$ .

14. Forecast error variances in models with estimated parameters.

As we've seen, computing forecast error variances that acknowledge parameter estimation uncertainty is very difficult; that's one reason why we've ignored it. We've learned a number of lessons about optimal forecasts while ignoring parameter estimation uncertainty, such as:

- a. Forecast error variance grows as the forecast horizon lengthens.
- b. In covariance stationary environments, the forecast error variance approaches the (finite) unconditional variance as the horizon grows.

Such lessons provide valuable insight and intuition regarding the workings of forecasting models and provide a useful benchmark for assessing actual forecasts. They sometimes need modification, however, when parameter estimation uncertainty is acknowledged. For example, in models with estimated parameters:

- a. Forecast error variance needn't grow monotonically with horizon. Typically we *expect* forecast error variance to increase monotonically with horizon, but it doesn't *have* to.
- b. Even in covariance stationary environments, the forecast error variance needn't converge to the unconditional variance as the forecast horizon lengthens; instead, it may grow without bound. Consider, for example, forecasting a series that's just a stationary  $AR(1)$  process around a linear trend. With known parameters, the point forecast will converge to the trend as the horizon grows, and the forecast error variance will converge to the unconditional variance of the  $AR(1)$  process. With estimated parameters, however, if the estimated trend parameters are even the slightest bit different from the true values (as they almost surely will be, due to sampling variation), that error

will be magnified as the horizon grows, so the forecast error variance will grow.

Thus, results derived under the assumption of known parameters should be viewed as a benchmark to guide our intuition, rather than as precise rules.

15. Direct vs. indirect autoregressive forecasts.

## **6.10 Notes**

# Chapter 7

## Cycles II: The Wold Representation and Its Approximation

This Chapter is a bit more abstract than most, but don't be put off. On the contrary, you may want to read it several times. The material in it is crucially important for time series modeling and forecasting and is therefore central to our concerns. In some parts (finite-ordered autoregressive models) it largely repeats Chapter 6, but that's intentional. It treats much more, including the Wold representation and its approximation and prediction using finite-ordered autoregressions, finite-ordered moving averages, and finite-ordered ARMA processes. Hence even the overlapping material is presented and integrated from a significantly more sophisticated perspective.

### 7.1 The Wold Representation and the General Linear Process

#### 7.1.1 The Wold Representation

Many different dynamic patterns are consistent with covariance stationarity. Thus, if we know only that a series is covariance stationary, it's not at all clear what sort of model we might fit to describe its evolution. The trend and seasonal models that we've studied aren't of use; they're models of specific

nonstationary components. Effectively, what we need now is an appropriate model for what's left after fitting the trend and seasonal components – a model for a covariance stationary residual. **Wold's representation theorem** points to the appropriate model.

Theorem:

Let  $\{y_t\}$  be any zero-mean covariance-stationary process.<sup>1</sup> Then we can write it as

$$y_t = B(L)\varepsilon_t = \sum_{i=0}^{\infty} b_i \varepsilon_{t-i}$$

$$\varepsilon_t \sim WN(0, \sigma^2),$$

where

$$b_0 = 1$$

$$\sum_{i=0}^{\infty} b_i^2 < \infty.$$

In short, the correct “model” for any covariance stationary series is some infinite distributed lag of white noise, called the **Wold representation**. The  $\varepsilon'_t$ s are often called **innovations**, because (as we'll see) they correspond to the 1-step-ahead forecast errors that we'd make if we were to use a particularly good forecast. That is, the  $\varepsilon'_t$ s represent that part of the evolution of  $y$  that's linearly unpredictable on the basis of the past of  $y$ . Note also that the  $\varepsilon'_t$ s, although uncorrelated, are not necessarily independent. Again, it's only for Gaussian random variables that lack of correlation implies independence, and the innovations are not necessarily Gaussian.

In our statement of Wold's theorem we assumed a zero mean. That may seem restrictive, but it's not. Rather, whenever you see  $y_t$ , just read  $(y_t - \mu)$ , so that the process is expressed in deviations from its mean. The deviation from the mean has a zero mean, by construction. Working with zero-mean

---

<sup>1</sup>Moreover, we require that the covariance stationary processes not contain any deterministic components.

processes therefore involves no loss of generality while facilitating notational economy. We'll use this device frequently.

### 7.1.2 The General Linear Process

Wold's theorem tells us that when formulating forecasting models for covariance stationary time series we need only consider models of the form

$$y_t = B(L)\varepsilon_t = \sum_{i=0}^{\infty} b_i \varepsilon_{t-i}$$

$$\varepsilon_t \sim WN(0, \sigma^2),$$

where the  $b_i$  are coefficients with  $b_0 = 1$  and  $\sum_{i=0}^{\infty} b_i^2 < \infty$ .

We call this the **general linear process**, “general” because any covariance stationary series can be written that way, and “linear” because the Wold representation expresses the series as a linear function of its innovations.

The general linear process is so important that it's worth examining its unconditional and conditional moment structure in some detail. Taking means and variances, we obtain the unconditional moments

$$E(y_t) = E\left(\sum_{i=0}^{\infty} b_i \varepsilon_{t-i}\right) = \sum_{i=0}^{\infty} b_i E\varepsilon_{t-i} = \sum_{i=0}^{\infty} b_i \cdot 0 = 0$$

and

$$var(y_t) = var\left(\sum_{i=0}^{\infty} b_i \varepsilon_{t-i}\right) = \sum_{i=0}^{\infty} b_i^2 var(\varepsilon_{t-i}) = \sum_{i=0}^{\infty} b_i^2 \sigma^2 = \sigma^2 \sum_{i=0}^{\infty} b_i^2.$$

At this point, in parallel to our discussion of white noise, we could compute and examine the autocovariance and autocorrelation functions of the general linear process. Those calculations, however, are rather involved, and not particularly revealing, so we'll proceed instead to examine the conditional mean and variance, where the information set  $\Omega_{t-1}$  upon which we condition

contains past innovations; that is,

$$\Omega_{t-1} = \varepsilon_{t-1}, \varepsilon_{t-2}, \dots$$

In this manner we can see how dynamics are modeled via conditional moments.<sup>2</sup> The conditional mean is

$$\begin{aligned} E(y_t|\Omega_{t-1}) &= E(\varepsilon_t|\Omega_{t-1}) + b_1 E(\varepsilon_{t-1}|\Omega_{t-1}) + b_2 E(\varepsilon_{t-2}|\Omega_{t-1}) + \dots \\ &= 0 + b_1 \varepsilon_{t-1} + b_2 \varepsilon_{t-2} + \dots = \sum_{i=1}^{\infty} b_i \varepsilon_{t-i}, \end{aligned}$$

and the conditional variance is

$$\text{var}(y_t|\Omega_{t-1}) = E[y_t - E(y_t|\Omega_{t-1})]^2|\Omega_{t-1}] = E(\varepsilon_t^2|\Omega_{t-1}) = E(\varepsilon_t^2) = \sigma^2.$$

The key insight is that the conditional mean *moves* over time in response to the evolving information set. The model captures the dynamics of the process, and the evolving conditional mean is one crucial way of summarizing them. An important goal of time series modeling, especially for forecasters, is capturing such conditional mean dynamics – the unconditional mean is constant (a requirement of stationarity), but the conditional mean varies in response to the evolving information set.<sup>3</sup>

## 7.2 Approximating the Wold Representation

When building forecasting models, we don't want to pretend that the model we fit is true. Instead, we want to be aware that we're *approximating* a

---

<sup>2</sup>Although Wold's theorem guarantees only serially uncorrelated white noise innovations, we shall sometimes make a stronger assumption of independent white noise innovations in order to focus the discussion. We do so, for example, in the following characterization of the conditional moment structure of the general linear process.

<sup>3</sup>Note, however, an embarrassing asymmetry: the conditional variance, like the unconditional variance, is a fixed constant. However, models that allow the conditional variance to change with the information set have been developed recently, as discussed in detail in Chapter ??.

more complex reality. That's the modern view, and it has important implications for forecasting. In particular, we've seen that the key to successful time series modeling and forecasting is parsimonious, yet accurate, approximation of the Wold representation. Here we consider three approximations: **moving average (MA) models**, **autoregressive (AR) models**, and **autoregressive moving average (ARMA) models**. The three models differ in their specifics and have different strengths in capturing different sorts of autocorrelation behavior.

We begin by characterizing the autocorrelation functions and related quantities associated with each model, under the assumption that the model is “true.” We do this separately for autoregressive, moving average, and ARMA models.<sup>4</sup> These characterizations have nothing to do with data or estimation, but they're crucial for developing a basic understanding of the properties of the models, which is necessary to perform intelligent modeling and forecasting. They enable us to make statements such as “If the data were really generated by an autoregressive process, then we'd expect its autocorrelation function to have property x.” Armed with that knowledge, we use the *sample* autocorrelations and partial autocorrelations, in conjunction with the AIC and the SIC, to suggest candidate forecasting models, which we then estimate.

### 7.2.1 Rational Distributed Lags

As we've seen, the Wold representation points to the crucial importance of models with infinite distributed lags. Infinite distributed lag models, in turn, are stated in terms of infinite polynomials in the lag operator, which are therefore very important as well. Infinite distributed lag models are not of immediate practical use, however, because they contain infinitely many pa-

---

<sup>4</sup>Sometimes, especially when characterizing population properties under the assumption that the models are correct, we refer to them as processes, which is short for **stochastic processes**. Hence the terms moving average process, autoregressive process, and ARMA process.

rameters, which certainly inhibits practical application! Fortunately, infinite polynomials in the lag operator needn't contain infinitely many free parameters. The infinite polynomial  $B(L)$  may for example be a ratio of finite-order (and perhaps very low-order) polynomials. Such polynomials are called rational polynomials, and distributed lags constructed from them are called **rational distributed lags**.

Suppose, for example, that

$$B(L) = \frac{\Theta(L)}{\Phi(L)},$$

where the numerator polynomial is of degree  $q$ ,

$$\Theta(L) = \sum_{i=0}^q \theta_i L^i,$$

and the denominator polynomial is of degree  $p$ ,

$$\Phi(L) = \sum_{i=0}^p \phi_i L^i.$$

There are *not* infinitely many free parameters in the  $B(L)$  polynomial; instead, there are only  $p + q$  parameters (the  $\theta$ 's and the  $\phi$ 's). If  $p$  and  $q$  are small, say 0, 1 or 2, then what seems like a hopeless task – estimation of  $B(L)$  – may actually be easy.

More realistically, suppose that  $B(L)$  is not exactly rational, but is approximately rational,

$$B(L) \approx \frac{\Theta(L)}{\Phi(L)},$$

Then we can **approximate the Wold representation** using a rational distributed lag. Rational distributed lags produce models of cycles that economize on parameters (they're parsimonious), while nevertheless providing accurate approximations to the Wold representation. The popular ARMA

and ARIMA forecasting models, which we'll introduce shortly, are simply rational approximations to the Wold representation.

### 7.2.2 Moving Average (*MA*) Models

The finite-order moving average processes is a natural and obvious approximation to the Wold representation, which is an infinite-order moving average process. Finite-order moving average processes also have direct motivation: the fact that all variation in time series, one way or another, is driven by shocks of various sorts suggests the possibility of modeling time series directly as distributed lags of current and past shocks, that is, as moving average processes.<sup>5</sup>

#### The *MA*(1) Process

The first-order moving average, or *MA*(1), process is

$$\begin{aligned} y_t &= \varepsilon_t + \theta \varepsilon_{t-1} = (1 + \theta L) \varepsilon_t \\ \varepsilon_t &\sim WN(0, \sigma^2). \end{aligned}$$

The defining characteristic of the MA process in general, and the *MA*(1) in particular, is that the current value of the observed series is expressed as a function of current and lagged unobservable shocks – think of it as a regression model with nothing but current and lagged disturbances on the right-hand side.

To help develop a feel for the behavior of the *MA*(1) process, we show two simulated realizations of length 150 in Figure 7.1. The processes are

$$y_t = \varepsilon_t + .4\varepsilon_{t-1}$$

---

<sup>5</sup>Economic equilibria, for example, may be disturbed by shocks that take some time to be fully assimilated.

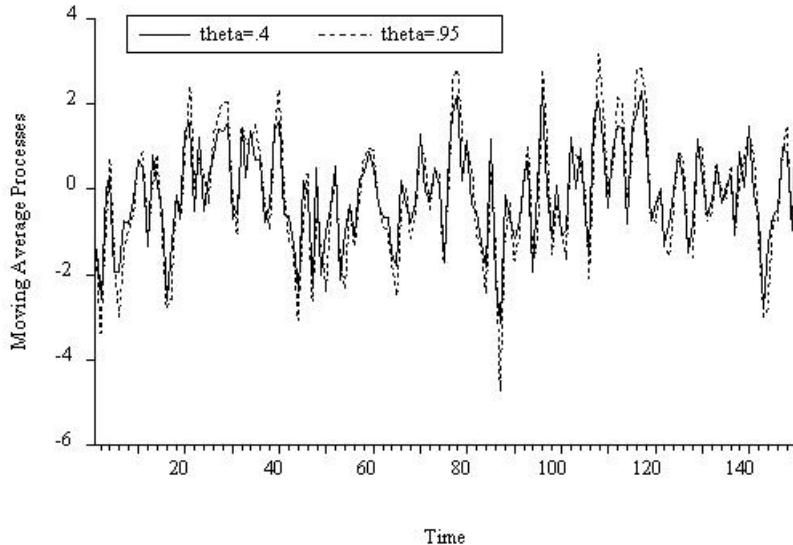


Figure 7.1: Realizations of Two MA(1) Processes

and

$$y_t = \varepsilon_t + .95\varepsilon_{t-1},$$

where in each case

$$\varepsilon_t \sim iid N(0, 1).$$

To construct the realizations, we used the same series of underlying white noise shocks; the only difference in the realizations comes from the different coefficients. Past shocks feed *positively* into the current value of the series, with a small weight of  $\theta=.4$  in one case and a large weight of  $\theta=.95$  in the other. You might think that  $\theta=.95$  would induce much more persistence than  $\theta=.4$ , but it doesn't. The structure of the *MA(1)* process, in which only the first lag of the shock appears on the right, forces it to have a very short memory, and hence weak dynamics, regardless of the parameter value.

The unconditional mean and variance are

$$Ey_t = E(\varepsilon_t) + \theta E(\varepsilon_{t-1}) = 0$$

and

$$\text{var}(y_t) = \text{var}(\varepsilon_t) + \theta^2 \text{var}(\varepsilon_{t-1}) = \sigma^2 + \theta^2 \sigma^2 = \sigma^2(1 + \theta^2).$$

Note that for a fixed value of  $\sigma$ , as  $\theta$  increases in absolute value so too does the unconditional variance. That's why the  $MA(1)$  process with parameter  $\theta=.95$  varies a bit more than the process with a parameter of  $\theta=.4$ .

The conditional mean and variance of an  $MA(1)$ , where the conditioning information set is

$$\Omega_{t-1} = \varepsilon_{t-1}, \varepsilon_{t-2}, \dots,$$

are

$$E(y_t|\Omega_{t-1}) = E(\varepsilon_t + \theta\varepsilon_{t-1}|\Omega_{t-1}) = E(\varepsilon_t|\Omega_{t-1}) + \theta E(\varepsilon_{t-1}|\Omega_{t-1}) = \theta\varepsilon_{t-1}$$

and

$$\text{var}(y_t|\Omega_{t-1}) = E[y_t - E(y_t|\Omega_{t-1})]^2|\Omega_{t-1}] = E(\varepsilon_t^2|\Omega_{t-1}) = E(\varepsilon_t^2) = \sigma^2.$$

The conditional mean explicitly adapts to the information set, in contrast to the unconditional mean, which is constant. Note, however, that only the first lag of the shock enters the conditional mean – more distant shocks have no effect on the current conditional expectation. This is indicative of the one-period memory of  $MA(1)$  processes, which we'll now characterize in terms of the autocorrelation function.

To compute the autocorrelation function for the  $MA(1)$  process, we must first compute the autocovariance function. We have

$$\begin{aligned} \gamma(\tau) &= E(y_t y_{t-\tau}) = E((\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-\tau} + \theta\varepsilon_{t-\tau-1})) = \\ &\quad 0, \text{ otherwise}. \end{aligned}$$

(The proof is left as a problem.) The autocorrelation function is just the

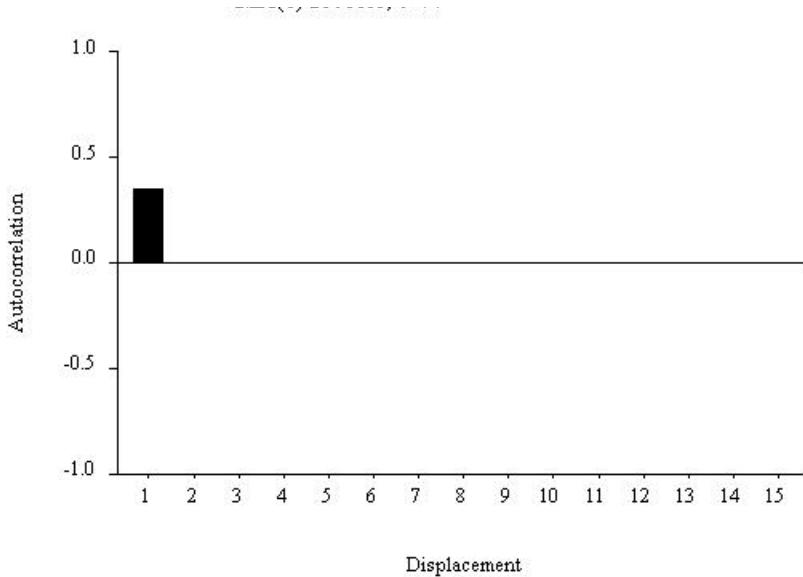


Figure 7.2: MA(1) Population Autocorrelation Function -  $\theta = .4$

autocovariance function scaled by the variance,

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)} = \begin{cases} \frac{\theta}{1+\theta^2}, & \tau = 1 \\ 0, & \text{otherwise.} \end{cases}.$$

The key feature here is the sharp *cutoff in the autocorrelations*. All autocorrelations are zero beyond displacement 1, the order of the *MA* process. In Figures 7.2 and 7.3, we show the autocorrelation functions for our two *MA*(1) processes with parameters  $\theta=.4$  and  $\theta=.95$ . At displacement 1, the process with parameter  $\theta=.4$  has a smaller autocorrelation (.34) than the process with parameter  $\theta=.95$ , (.50) but both drop to zero beyond displacement 1.

Note that the requirements of covariance stationarity (constant unconditional mean, constant and finite unconditional variance, autocorrelation depends only on displacement) are met for any *MA*(1) process, *regardless* of the values of its parameters. If, moreover,  $|\theta| < 1$ , then we say that the *MA*(1) process is **invertible**. In that case, we can “invert” the *MA*(1) process and express the current value of the series not in terms of a current shock and a

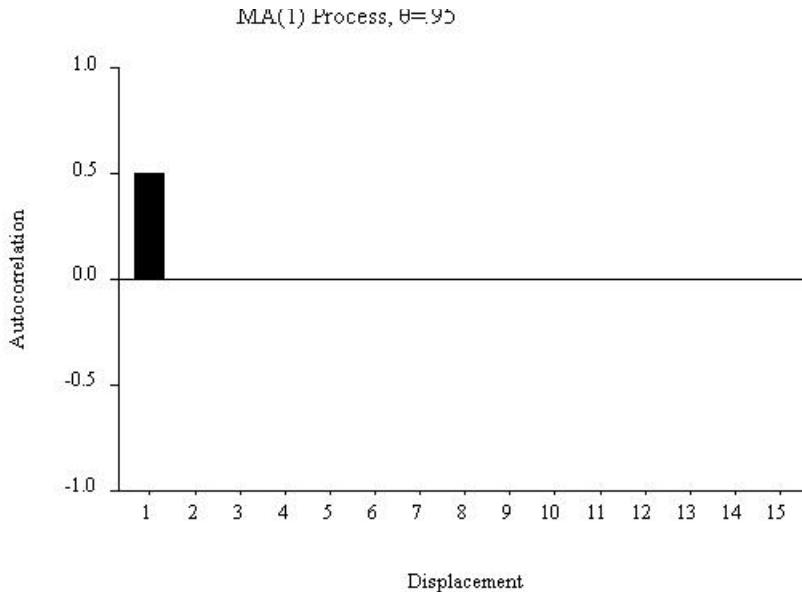


Figure 7.3: MA(1) Population Autocorrelation Function -  $\theta = .95$

lagged shock, but rather in terms of a current shock *and lagged values of the series*. That's called an **autoregressive representation**. An autoregressive representation has a current shock and lagged observable values of the series on the right, whereas a moving average representation has a current shock and lagged unobservable shocks on the right.

Let's compute the autoregressive representation. The process is

$$y_t = \varepsilon_t + \theta \varepsilon_{t-1}$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

Thus we can solve for the innovation as

$$\varepsilon_t = y_t - \theta \varepsilon_{t-1}.$$

Lagging by successively more periods gives expressions for the innovations at various dates,

$$\varepsilon_{t-1} = y_{t-1} - \theta \varepsilon_{t-2}$$

$$\varepsilon_{t-2} = y_{t-2} - \theta\varepsilon_{t-3}$$

$$\varepsilon_{t-3} = y_{t-3} - \theta\varepsilon_{t-4},$$

and so forth. Making use of these expressions for lagged innovations we can substitute backward in the  $MA(1)$  process, yielding

$$y_t = \varepsilon_t + \theta y_{t-1} - \theta^2 y_{t-2} + \theta^3 y_{t-3} - \dots$$

In lag-operator notation, we write the infinite autoregressive representation as

$$\frac{1}{1 + \theta L} y_t = \varepsilon_t.$$

Note that the back substitution used to obtain the autoregressive representation only makes sense, and in fact a convergent autoregressive representation only exists, if  $|\theta| < 1$ , because in the back substitution we raise  $\theta$  to progressively higher powers.

We can restate the invertibility condition in another way: the inverse of the root of the moving average lag operator polynomial  $(1 + \theta L)$  must be less than one in absolute value. Recall that a polynomial of degree  $m$  has  $m$  roots. Thus the  $MA(1)$  lag operator polynomial has one root, which is the solution to

$$1 + \theta L = 0.$$

The root is  $L = -1/\theta$ , so its inverse will be less than one in absolute value if  $|\theta| < 1$ , and the two invertibility conditions are equivalent. The “inverse root” way of stating invertibility conditions seems tedious, but it turns out to be of greater applicability than the  $|\theta| < 1$  condition, as we’ll see shortly.

Autoregressive representations are appealing to forecasters, because one way or another, if a model is to be used for real-world forecasting, it’s got to link the present observables to the past history of observables, so that we can extrapolate to form a forecast of future observables based on present

and past observables. Superficially, moving average models don't seem to meet that requirement, because the current value of a series is expressed in terms of current and lagged unobservable shocks, not observable variables. But under the invertibility conditions that we've described, moving average processes have equivalent autoregressive representations. Thus, although we want autoregressive representations for forecasting, we don't have to start with an autoregressive model. However, we typically restrict ourselves to invertible processes, because for forecasting purposes we want to be able to express current observables as functions of past observables.

Finally, let's consider the partial autocorrelation function for the  $MA(1)$  process. From the infinite autoregressive representation of the  $MA(1)$  process, we see that the partial autocorrelation function will decay gradually to zero. As we discussed earlier, the partial autocorrelations are just the coefficients on the last included lag in a sequence of progressively higher-order autoregressive approximations. If  $\theta > 0$ , then the pattern of decay will be one of damped oscillation; otherwise, the decay will be one-sided.

In Figures 7.4 and 7.5 we show the partial autocorrelation functions for our example  $MA(1)$  processes. For each process,  $|\theta| < 1$ , so that an autoregressive representation exists, and  $\theta > 0$ , so that the coefficients in the autoregressive representations alternate in sign. Specifically, we showed the general autoregressive representation to be

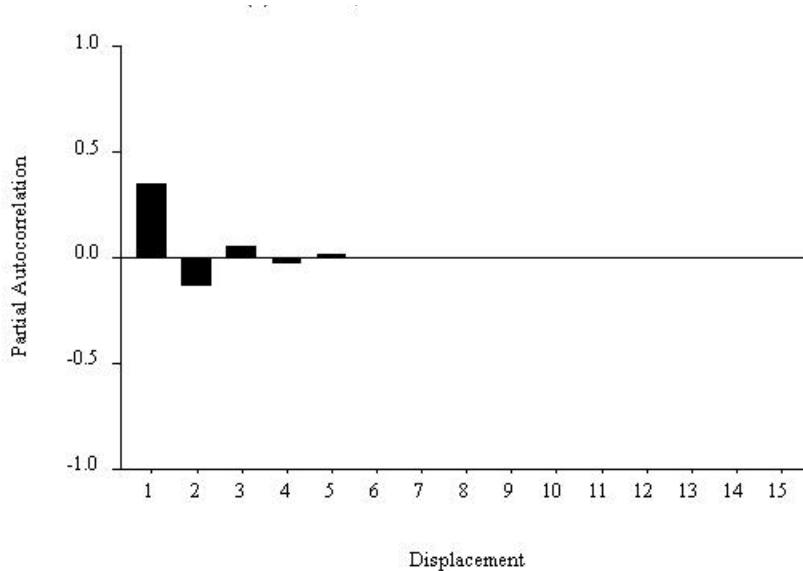
$$y_t = \varepsilon_t + \theta y_{t-1} - \theta^2 y_{t-2} + \theta^3 y_{t-3} - \dots,$$

so the autoregressive representation for the process with  $\theta=.4$  is

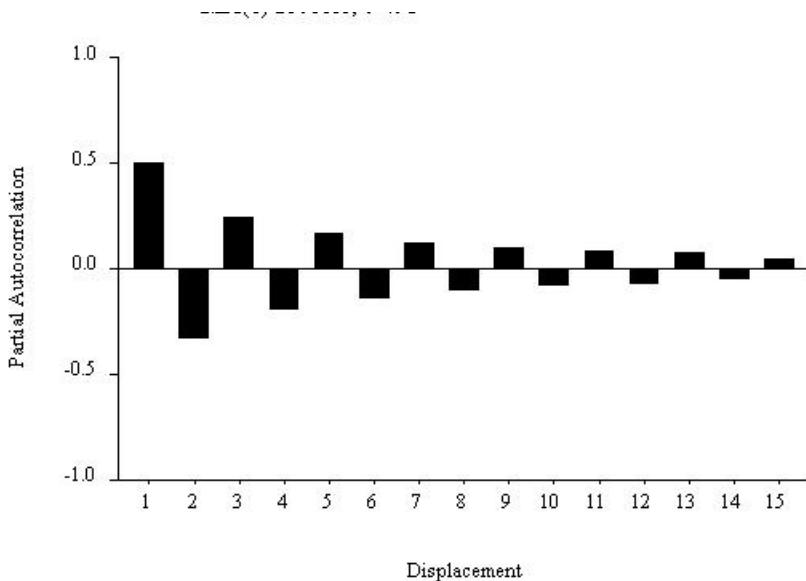
$$y_t = \varepsilon_t + .4y_{t-1} - .4^2 y_{t-2} + \dots = \varepsilon_t + .4y_{t-1} - .16y_{t-2} + \dots,$$

and the autoregressive representation for the process with  $\theta=.95$  is

$$y_t = \varepsilon_t + .95y_{t-1} - .95^2 y_{t-2} + \dots = \varepsilon_t + .95y_{t-1} - .9025y_{t-2} + \dots$$

Figure 7.4: MA(1) Population Partial Autocorrelation Function -  $\theta = .4$ 

The partial autocorrelations display a similar damped oscillation.<sup>6</sup> The decay, however, is slower for the  $\theta=.95$  case.

Figure 7.5: MA(1) Population Partial Autocorrelation Function -  $\theta = .95$ 


---

<sup>6</sup>Note, however, that the partial autocorrelations are *not* the successive coefficients in the infinite autoregressive representation. Rather, they are the coefficients on the last included lag in sequence of progressively longer autoregressions. The two are related but distinct.

### The $MA(q)$ Process

Now consider the general finite-order moving average process of order  $q$ , or  $MA(q)$  for short,

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} = \Theta(L) \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2),$$

where

$$\Theta(L) = 1 + \theta_1 L + \dots + \theta_q L^q$$

is a  $q$ th-order lag operator polynomial. The  $MA(q)$  process is a natural generalization of the  $MA(1)$ . By allowing for more lags of the shock on the right side of the equation, the  $MA(q)$  process can capture richer dynamic patterns, which we can potentially exploit for improved forecasting. The  $MA(1)$  process is of course a special case of the  $MA(q)$ , corresponding to  $q = 1$ .

The properties of the  $MA(q)$  processes parallel those of the  $MA(1)$  process in all respects, so in what follows we'll refrain from grinding through the mathematical derivations. Instead we'll focus on the key features of practical importance. Just as the  $MA(1)$  process was covariance stationary for any value of its parameters, so too is the finite-order  $MA(q)$  process. As with the  $MA(1)$  process, the  $MA(q)$  process is *invertible* only if a root condition is satisfied. The  $MA(q)$  lag operator polynomial has  $q$  roots; when  $q > 1$  the possibility of complex roots arises. The condition for invertibility of the  $MA(q)$  process is that the inverses of all of the roots must be inside the unit circle, in which case we have the convergent autoregressive representation,

$$\frac{1}{\Theta(L)} y_t = \varepsilon_t.$$

The conditional mean of the  $MA(q)$  process evolves with the information

set, in contrast to the unconditional moments, which are fixed. In contrast to the  $MA(1)$  case, in which the conditional mean depends on only the first lag of the innovation, in the  $MA(q)$  case the conditional mean depends on  $q$  lags of the innovation. Thus the  $MA(q)$  process has the potential for longer memory.

The potentially longer memory of the  $MA(q)$  process emerges clearly in its autocorrelation function. In the  $MA(1)$  case, all autocorrelations beyond displacement 1 are zero; in the  $MA(q)$  case all autocorrelations beyond displacement  $q$  are zero. This autocorrelation cutoff is a distinctive property of moving average processes. The partial autocorrelation function of the  $MA(q)$  process, in contrast, decays gradually, in accord with the infinite autoregressive representation, in either an oscillating or one-sided fashion, depending on the parameters of the process.

In closing this section, let's step back for a moment and consider in greater detail the precise way in which finite-order moving average processes approximate the Wold representation. The Wold representation is

$$y_t = B(L)\varepsilon_t,$$

where  $B(L)$  is of infinite order. The  $MA(1)$ , in contrast, is simply a first-order moving average, in which a series is expressed as a one-period moving average of current and past innovations. Thus when we fit an  $MA(1)$  model we're using the first-order polynomial  $1 + \theta L$  to approximate the infinite-order polynomial  $B(L)$ . Note that  $1 + \theta L$  is a rational polynomial with numerator polynomial of degree one and degenerate denominator polynomial (degree zero).

$MA(q)$  process have the potential to deliver better approximations to the Wold representation, at the cost of more parameters to be estimated. The Wold representation involves an infinite moving average; the  $MA(q)$  process

approximates the infinite moving average with a *finite-order* moving average,

$$y_t = \Theta(L)\varepsilon_t,$$

whereas the *MA*(1) process approximates the infinite moving average with only a *first-order* moving average, which can sometimes be very restrictive.

Soon we shall see that *MA* processes are absolutely central for understanding forecasting and properties of forecast errors, even if they usually not used directly as forecasting models. Other approximations to the Wold representation are typically more useful for producing forecasts, in particular autoregressive (*AR*) and mixed autoregressive moving-average (*ARMA*) models, to which we now turn.

### 7.2.3 Autoregressive (*AR*) Models

The autoregressive process is also a natural approximation to the Wold representation. We've seen, in fact, that under certain conditions a moving average process has an autoregressive representation, so an autoregressive process is in a sense the same as a moving average process. Like the moving average process, the autoregressive process has direct motivation; it's simply a stochastic difference equation, a simple mathematical model in which the current value of a series is linearly related to its past values, plus an additive stochastic shock. Stochastic difference equations are a natural vehicle for discrete-time stochastic dynamic modeling.

#### The *AR*(1) Process

The first-order autoregressive process, *AR*(1) for short, is

$$y_t = \phi y_{t-1} + \varepsilon_t$$

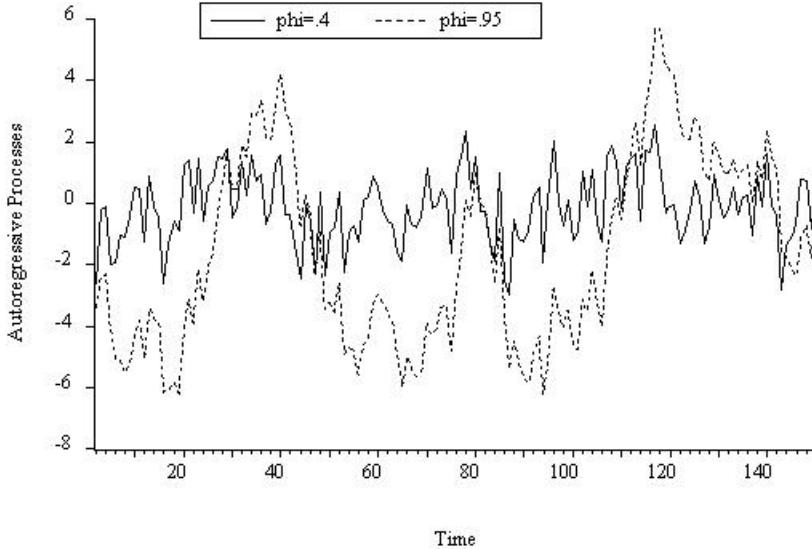


Figure 7.6: Realization of Two AR(1) Processes

$$\varepsilon_t \sim WN(0, \sigma^2).$$

In lag operator form, we write

$$(1 - \phi L)y_t = \varepsilon_t.$$

In Figure 7.6 we show simulated realizations of length 150 of two AR(1) processes; the first is

$$y_t = .4y_{t-1} + \varepsilon_t,$$

and the second is

$$y_t = .95y_{t-1} + \varepsilon_t,$$

where in each case

$$\varepsilon_t \text{ iid } N(0, 1),$$

and the same innovation sequence underlies each realization.

The fluctuations in the  $AR(1)$  with parameter  $\phi = .95$  appear much more persistent than those of the  $AR(1)$  with parameter  $\phi = .4$ . This contrasts sharply with the  $MA(1)$  process, which has a very short memory regardless

of parameter value. Thus the  $AR(1)$  model is capable of capturing much more persistent dynamics than is the  $MA(1)$ .

Recall that a finite-order moving average process is always covariance stationary, but that certain conditions must be satisfied for invertibility, in which case an autoregressive representation exists. For autoregressive processes, the situation is precisely the reverse. Autoregressive processes are always invertible – in fact invertibility isn’t even an issue, as finite-order autoregressive processes *already are* in autoregressive form – but certain conditions must be satisfied for an autoregressive process to be covariance stationary.

If we begin with the  $AR(1)$  process,

$$y_t = \phi y_{t-1} + \varepsilon_t,$$

and substitute backward for lagged  $y$ ’s on the right side, we obtain

$$y_t = \varepsilon_t + \phi \varepsilon_{t-1} + \phi^2 \varepsilon_{t-2} + \dots$$

In lag operator form we write

$$y_t = \frac{1}{1 - \phi L} \varepsilon_t.$$

This moving average representation for  $y$  is convergent if and only if  $|\phi| < 1$ ; thus,  $|\phi| < 1$  is the condition for covariance stationarity in the  $AR(1)$  case. Equivalently, the condition for covariance stationarity is that the inverse of the root of the autoregressive lag operator polynomial be less than one in absolute value.

From the moving average representation of the covariance stationary  $AR(1)$

process, we can compute the unconditional mean and variance,

$$E(y_t) = E(\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \dots)$$

$$= E(\varepsilon_t) + \phi E(\varepsilon_{t-1}) + \phi^2 E(\varepsilon_{t-2}) + \dots$$

$$= 0$$

and

$$\text{var}(y_t) = \text{var}(\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \dots)$$

$$= \sigma^2 + \phi^2\sigma^2 + \phi^4\sigma^2 + \dots$$

$$= \sigma^2 \sum_{i=0}^{\infty} \phi^{2i}$$

$$= \frac{\sigma^2}{1-\phi^2}.$$

The conditional moments, in contrast, are

$$E(y_t|y_{t-1}) = E(\phi y_{t-1} + \varepsilon_t|y_{t-1})$$

$$= \phi E(y_{t-1}|y_{t-1}) + E(\varepsilon_t|y_{t-1})$$

$$= \phi y_{t-1} + 0$$

$$= \phi y_{t-1}$$

and

$$\begin{aligned} \text{var}(y_t|y_{t-1}) &= \text{var}((\phi y_{t-1} + \varepsilon_t) \mid y_{t-1}) \\ &= \phi^2 \text{var}(y_{t-1}|y_{t-1}) + \text{var}(\varepsilon_t|y_{t-1}) \\ &= 0 + \sigma^2 \\ &= \sigma^2. \end{aligned}$$

Note in particular that the simple way that the conditional mean adapts to the changing information set as the process evolves.

To find the autocovariances, we proceed as follows. The process is

$$y_t = \phi y_{t-1} + \varepsilon_t,$$

so that multiplying both sides of the equation by  $y_{t-\tau}$  we obtain

$$y_t y_{t-\tau} = \phi y_{t-1} y_{t-\tau} + \varepsilon_t y_{t-\tau}.$$

For  $\tau \geq 1$ , taking expectations of both sides gives

$$\gamma(\tau) = \phi \gamma(\tau - 1).$$

This is called the **Yule-Walker equation**. It is a recursive equation; that is, given  $\gamma(\tau)$ , for any  $\tau$ , the Yule-Walker equation immediately tells us how to get  $\gamma(\tau + 1)$ . If we knew  $\gamma(0)$  to start things off (an “initial condition”), we could use the Yule-Walker equation to determine the entire autocovariance sequence. And we *do* know  $\gamma(0)$ ; it’s just the variance of the process, which we already showed to be

$$\gamma(0) = \frac{\sigma^2}{1 - \phi^2}.$$

Thus we have

$$\begin{aligned}\gamma(0) &= \frac{\sigma^2}{1 - \phi^2} \\ \gamma(1) &= \phi \frac{\sigma^2}{1 - \phi^2} \\ \gamma(2) &= \phi^2 \frac{\sigma^2}{1 - \phi^2},\end{aligned}$$

and so on. In general, then,

$$\gamma(\tau) = \phi^\tau \frac{\sigma^2}{1 - \phi^2}, \quad \tau = 0, 1, 2, \dots$$

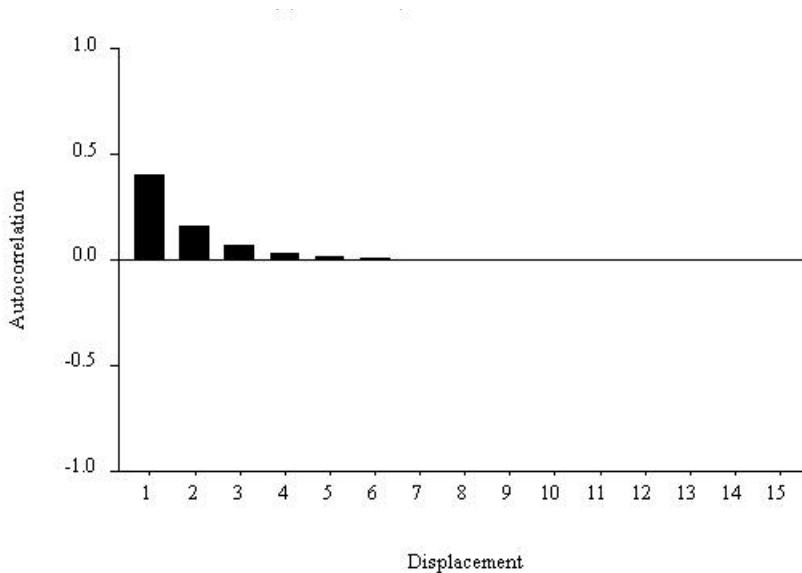
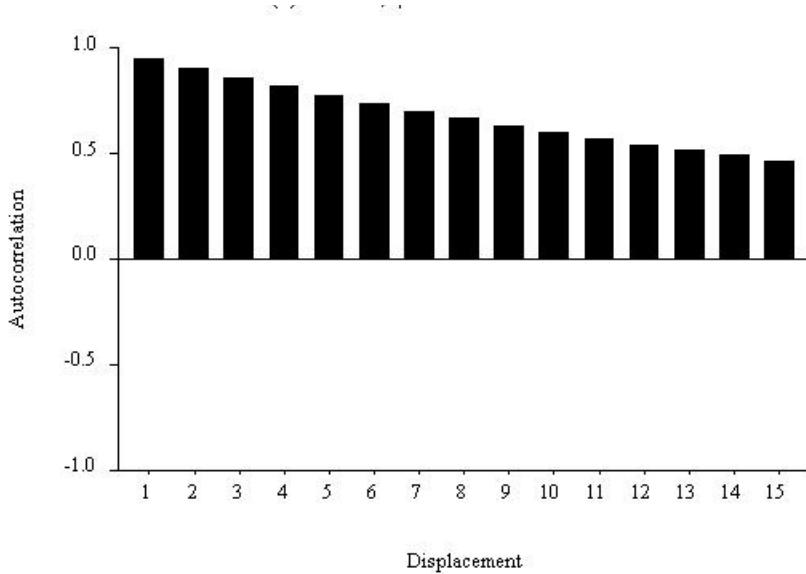
Dividing through by  $\gamma(0)$  gives the autocorrelations,

$$\rho(\tau) = \phi^\tau, \quad \tau = 0, 1, 2, \dots$$

Note the gradual autocorrelation decay, which is typical of autoregressive processes. The autocorrelations approach zero, but only in the limit as the displacement approaches infinity. In particular, they don't cut off to zero, as is the case for moving average processes. If  $\phi$  is positive, the autocorrelation decay is one-sided. If  $\phi$  is negative, the decay involves back-and-forth oscillations. The relevant case in business and economics is  $\phi > 0$ , but either way, the autocorrelations damp gradually, not abruptly. In Figure 7.7 and 7.8 we show the autocorrelation functions for  $AR(1)$  processes with parameters  $\phi = .4$  and  $\phi = .95$ . The persistence is much stronger when  $\phi = .95$ , in contrast to the  $MA(1)$  case, in which the persistence was weak regardless of the parameter.

Finally, the partial autocorrelation function for the  $AR(1)$  process cuts off abruptly; specifically,

$$p(\tau) = \begin{cases} \phi, & \tau = 1 \\ 0, & \tau > 1. \end{cases}.$$

Figure 7.7: AR(1) Population Autocorrelation Function -  $\rho = .4$ Figure 7.8: AR(1) Population Autocorrelation Function -  $\rho = .95$ 

It's easy to see why. The partial autocorrelations are just the last coefficients in a sequence of successively longer population autoregressions. If the true process is in fact an  $AR(1)$ , the first partial autocorrelation is just the autoregressive coefficient, and coefficients on all longer lags are zero.

In Figures 7.9 and 7.10 we show the partial autocorrelation functions for

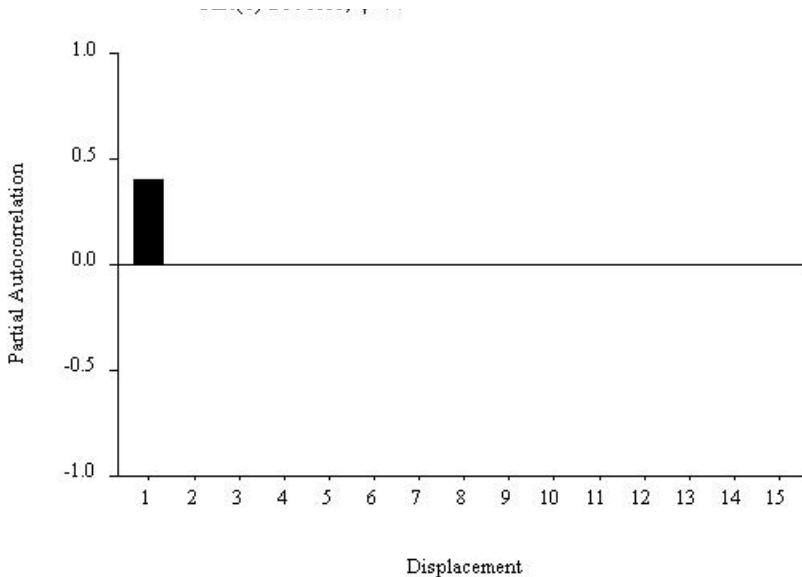


Figure 7.9: AR(1) Population Partial Autocorrelation Function -  $\rho = .4$

our two  $AR(1)$  processes. At displacement 1, the partial autocorrelations are simply the parameters of the process (.4 and .95, respectively), and at longer displacements, the partial autocorrelations are zero.

### The $AR(p)$ Process

The general  $p$ -th order autoregressive process, or  $AR(p)$  for short, is

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

In lag operator form we write

$$\Phi(L)y_t = (1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p) y_t = \varepsilon_t.$$

As with our discussion of the  $MA(q)$  process, in our discussion of the  $AR(p)$  process we dispense here with mathematical derivations and instead rely on parallels with the  $AR(1)$  case to establish intuition for its key properties.

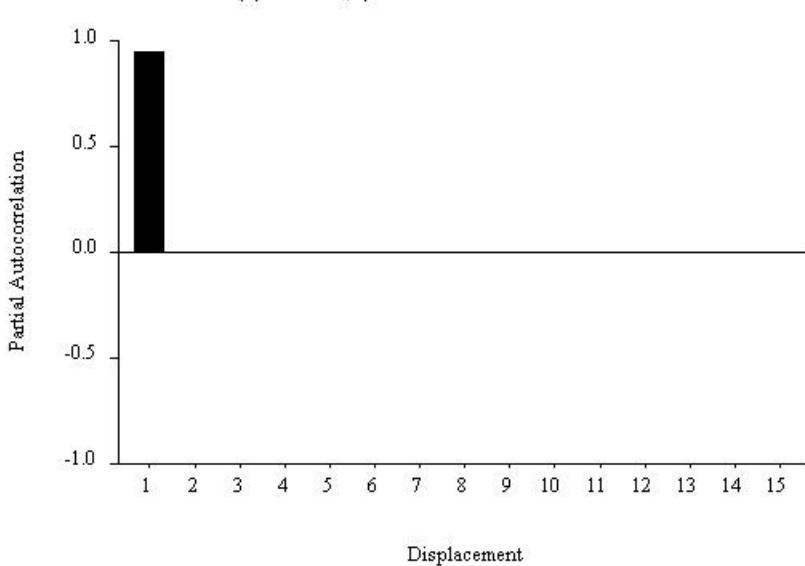


Figure 7.10: AR1) Population Partial Autocorrelation Function -  $\rho = .95$

An  $AR(p)$  process is covariance stationary if and only if the inverses of all roots of the autoregressive lag operator polynomial  $\Phi(L)$  are inside the unit circle.<sup>7</sup> In the covariance stationary case we can write the process in the convergent infinite moving average form

$$y_t = \frac{1}{\Phi(L)} \varepsilon_t.$$

The autocorrelation function for the general  $AR(p)$  process, as with that of the  $AR(1)$  process, decays gradually with displacement. Finally, the  $AR(p)$  partial autocorrelation function has a sharp cutoff at displacement  $p$ , for the same reason that the  $AR(1)$  partial autocorrelation function has a sharp cutoff at displacement 1.

Let's discuss the  $AR(p)$  autocorrelation function in a bit greater depth.

---

<sup>7</sup>A necessary condition for covariance stationarity, which is often useful as a quick check, is

$$\sum_{i=1}^p \phi_i < 1.$$

If the condition is satisfied, the process may or may not be stationary, but if the condition is violated, the process can't be stationary.

The key insight is that, in spite of the fact that its qualitative behavior (gradual damping) matches that of the  $AR(1)$  autocorrelation function, it can nevertheless display a richer variety of patterns, depending on the order and parameters of the process. It can, for example, have damped monotonic decay, as in the  $AR(1)$  case with a positive coefficient, but it can also have damped oscillation in ways that  $AR(1)$  can't have. In the  $AR(1)$  case, the only possible oscillation occurs when the coefficient is negative, in which case the autocorrelations switch signs at each successively longer displacement. In higher-order autoregressive models, however, the autocorrelations can oscillate with much richer patterns reminiscent of cycles in the more traditional sense. This occurs when some roots of the autoregressive lag operator polynomial are complex.<sup>8</sup>

Consider, for example, the  $AR(2)$  process,

$$y_t = 1.5y_{t-1} - .9y_{t-2} + \varepsilon_t.$$

The corresponding lag operator polynomial is  $1 - 1.5L + .9L^2$ , with two complex conjugate roots,  $.83 \pm .65i$ . The inverse roots are  $.75 \pm .58i$ , both of which are close to, but inside, the unit circle; thus the process is covariance stationary. It can be shown that the autocorrelation function for an  $AR(2)$  process is

$$\rho(0) = 1$$

$$\rho(\tau) = \phi_1\rho(\tau - 1) + \phi_2\rho(\tau - 2), \quad \tau = 2, 3, \dots$$

$$\rho(1) = \frac{\phi_1}{1 - \phi_2}$$

Using this formula, we can evaluate the autocorrelation function for the process at hand; we plot it in Figure 7.11. Because the roots are complex, the autocorrelation function oscillates, and because the roots are close to the unit circle, the oscillation damps slowly.

---

<sup>8</sup>Note that complex roots can't occur in the  $AR(1)$  case.

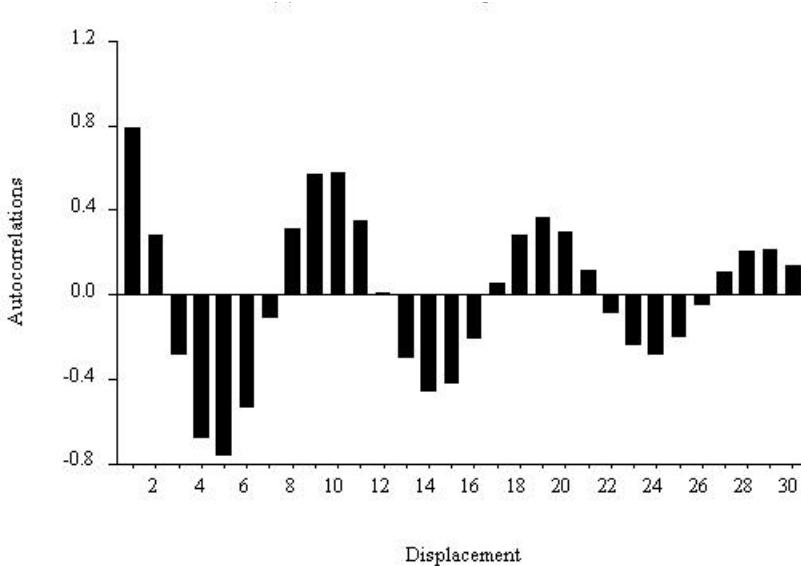


Figure 7.11: Population Autocorrelation Function - AR(2) with Complex Roots

Finally, let's step back once again to consider in greater detail the precise way that finite-order autoregressive processes approximate the Wold representation. As always, the Wold representation is  $y_t = B(L)\varepsilon_t$ , where  $B(L)$  is of infinite order. The  $AR(1)$ , as compared to the  $MA(1)$ , is simply a different approximation to the Wold representation. The moving average representation associated with the  $AR(1)$  process is  $y_t = 1/1 - \phi L\varepsilon_t$ . Thus, when we fit an  $AR(1)$  model, we're using  $1/1 - \phi L$ , a rational polynomial with degenerate numerator polynomial (degree zero) and denominator polynomial of degree one, to approximate  $B(L)$ . The moving average representation associated with the  $AR(1)$  process is of infinite order, as is the Wold representation, but it does not have infinitely many free coefficients. In fact, only one parameter,  $\phi$ , underlies it.

The  $AR(p)$  is an obvious generalization of the  $AR(1)$  strategy for approximating the Wold representation. The moving average representation associated with the  $AR(p)$  process is  $y_t = 1/\Phi(L)\varepsilon_t$ . When we fit an  $AR(p)$  model to approximate the Wold representation we're still using a rational polynomial with degenerate numerator polynomial (degree zero), but the de-

nominator polynomial is of higher degree.

### 7.2.4 Autoregressive Moving Average (ARMA) Models

Autoregressive and moving average models are often combined in attempts to obtain better and more parsimonious approximations to the Wold representation, yielding the autoregressive moving average process, **ARMA(p,q)** for short. As with moving average and autoregressive processes, ARMA processes also have direct motivation.<sup>9</sup> First, if the random shock that drives an autoregressive process is itself a moving average process, then it can be shown that we obtain an ARMA process. Second, ARMA processes can arise from aggregation. For example, sums of AR processes, or sums of AR and MA processes, can be shown to be ARMA processes. Finally, AR processes observed subject to measurement error also turn out to be ARMA processes.

The simplest ARMA process that's not a pure autoregression or pure moving average is the ARMA(1,1), given by

$$y_t = \phi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$$

$$\varepsilon_t \sim WN(0, \sigma^2),$$

or in lag operator form,

$$(1 - \phi L) y_t = (1 + \theta L) \varepsilon_t,$$

where  $|\phi| < 1$  is required for stationarity and  $|\theta| < 1$  is required for invertibility.<sup>10</sup> If the covariance stationarity condition is satisfied, then we have the moving average representation

$$y_t = \frac{(1 + \theta L)}{(1 - \phi L)} \varepsilon_t,$$

---

<sup>9</sup>For more extensive discussion, see Granger and Newbold (1986).

<sup>10</sup>Both stationarity and invertibility need to be checked in the ARMA case, because both autoregressive and moving average components are present.

which is an infinite distributed lag of current and past innovations. Similarly, if the invertibility condition is satisfied, then we have the infinite autoregressive representation,

$$\frac{(1 - \phi L)}{(1 + \theta L)} y_t = \varepsilon_t.$$

The ARMA(p,q) process is a natural generalization of the ARMA(1,1) that allows for multiple moving average and autoregressive lags. We write

$$y_t = \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}$$

$$\varepsilon_t \sim WN(0, \sigma^2),$$

or

$$\Phi(L)y_t = \Theta(L)\varepsilon_t,$$

where

$$\Phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p$$

and

$$\Theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q.$$

If the inverses of all roots of  $\Phi(L)$  are inside the unit circle, then the process is covariance stationary and has convergent infinite moving average representation

$$y_t = \frac{\Theta(L)}{\Phi(L)} \varepsilon_t.$$

If the inverses of all roots of  $\Theta(L)$  are inside the unit circle, then the process is invertible and has convergent infinite autoregressive representation

$$\frac{\Phi(L)}{\Theta(L)} y_t = \varepsilon_t.$$

As with autoregressions and moving averages, ARMA processes have a fixed unconditional mean but a time-varying conditional mean. In contrast to pure moving average or pure autoregressive processes, however, neither the

autocorrelation nor partial autocorrelation functions of ARMA processes cut off at any particular displacement. Instead, each damps gradually, with the precise pattern depending on the process.

ARMA models approximate the Wold representation by a ratio of two finite-order lag-operator polynomials, neither of which is degenerate. Thus ARMA models use ratios of full-fledged polynomials in the lag operator to approximate the Wold representation,

$$y_t = \frac{\Theta(L)}{\Phi(L)} \varepsilon_t.$$

ARMA models, by allowing for both moving average and autoregressive components, often provide accurate approximations to the Wold representation that nevertheless have just a few parameters. That is, ARMA models are often both highly accurate and highly parsimonious. In a particular situation, for example, it might take an AR(5) to get the same approximation accuracy as could be obtained with an ARMA(2,1), but the AR(5) has five parameters to be estimated, whereas the ARMA(2,1) has only three.

### 7.3 Forecasting Cycles From a Moving-Average Perspective: Wiener-Kolmogorov

By now you've gotten comfortable with the idea of an **information set**. Here we'll use that idea extensively. We denote the time- $T$  information set by  $\Omega_T$ . As first pass it seems most natural to think of the information set as containing the available past history of the series,

$$\Omega_T = \{y_T, y_{T-1}, y_{T-2}, \dots\},$$

where for theoretical purposes we imagine history as having begun in the infinite past.

### 7.3. FORECASTING CYCLES FROM A MOVING-AVERAGE PERSPECTIVE: WIENER-KOLMOGOV

So long as  $y$  is covariance stationary, however, we can just as easily express the information available at time  $T$  in terms of current and past shocks,

$$\Omega_T = \{\varepsilon_T, \varepsilon_{T-1}, \varepsilon_{T-2}, \dots\}.$$

Suppose, for example, that the process to be forecast is a covariance stationary  $AR(1)$ ,

$$y_t = \phi y_{t-1} + \varepsilon_t.$$

Then immediately,

$$\varepsilon_T = y_T - \phi y_{T-1}$$

$$\varepsilon_{T-1} = y_{T-1} - \phi y_{T-2}$$

$$\varepsilon_{T-2} = y_{T-2} - \phi y_{T-3},$$

and so on. In other words, we can figure out the current and lagged  $\varepsilon$ 's from the current and lagged  $y$ 's. More generally, for any covariance stationary and invertible series, we can infer the history of  $\varepsilon$  from the history of  $y$ , and the history of  $y$  from the history of  $\varepsilon$ .

Assembling the discussion thus far, we can view the time- $T$  information set as containing the current and past values of either (or both)  $y$  and  $\varepsilon$ ,

$$\Omega_T = y_T, y_{T-1}, y_{T-2}, \dots, \varepsilon_T, \varepsilon_{T-1}, \varepsilon_{T-2}, \dots$$

Based upon that information set, we want to find the **optimal forecast** of  $y$  at some future time  $T + h$ . The optimal forecast is the one with the smallest loss on average, that is, the forecast that minimizes **expected loss**. It turns out that under reasonably weak conditions the optimal forecast is the **conditional mean**,

$$E(y_{T+h}|\Omega_T),$$

the expected value of the future value of the series being forecast, conditional upon available information.

In general, the conditional mean need not be a linear function of the elements of the information set. Because linear functions are particularly tractable, we prefer to work with linear forecasts – forecasts that are linear in the elements of the information set – by finding the best linear approximation to the conditional mean, called the **linear projection**, denoted

$$P(y_{T+h}|\Omega_T).$$

This explains the common term “**linear least squares forecast**.” The linear projection is often very useful and accurate, because the conditional mean is often close to linear. In fact, in the Gaussian case the conditional expectation is exactly linear, so that

$$E(y_{T+h}|\Omega_T) = P(y_{T+h}|\Omega_T).$$

### 7.3.1 Optimal Point Forecasts for Finite-Order Moving Averages

Our forecasting method is always the same: we write out the process for the future time period of interest,  $T + h$ , and project it on what’s known at time  $T$ , when the forecast is made. This process is best learned by example. Consider an  $MA(2)$  process,

$$y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

Suppose we’re standing at time  $T$  and we want to forecast  $y_{T+1}$ . First we write out the process for  $T + 1$ ,

$$y_{T+1} = \varepsilon_{T+1} + \theta_1\varepsilon_T + \theta_2\varepsilon_{T-1}.$$

Then we project on the time- $T$  information set, which simply means that all future innovations are replaced by zeros. Thus

$$y_{T+1,T} = P(y_{T+1}|\Omega_T) = \theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1}.$$

To forecast 2 steps ahead, we note that

$$y_{T+2} = \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1} + \theta_2 \varepsilon_T,$$

and we project on the time- $T$  information set to get

$$y_{T+2,T} = \theta_2 \varepsilon_T.$$

Continuing in this fashion, we see that

$$y_{T+h,T} = 0,$$

for all  $h > 2$ .

Now let's compute the corresponding **forecast errors**.<sup>11</sup> We have:

$$e_{T+1,T} = \varepsilon_{T+1} - W N$$

$$e_{T+2,T} = \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1} \quad (MA(1))$$

$$e_{T+h,T} = \varepsilon_{T+h} + \theta_1 \varepsilon_{T+h-1} + \theta_2 \varepsilon_{T+h-2} \quad (MA(2)),$$

for all  $h > 2$ .

Finally, the **forecast error variances** are:

$$\sigma_1^2 = \sigma^2$$

$$\sigma_2^2 = \sigma^2(1 + \theta_1^2)$$

$$\sigma_h^2 = \sigma^2(1 + \theta_1^2 + \theta_2^2),$$

---

<sup>11</sup>Recall that the forecast error is simply the difference between the actual and forecasted values. That is,  $e_{T+h,T} = y_{T+h} - y_{T+h,T}$ .

for all  $h > 2$ . Moreover, the forecast error variance for  $h>2$  is just the unconditional variance of  $y_t$ .

Now consider the general  $MA(q)$  case. The model is

$$y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \dots + \theta_q\varepsilon_{t-q}.$$

First, consider the forecasts. If  $h \leq q$ , the forecast has the form

$$y_{T+h,T} = 0 + \text{“adjustment,”}$$

whereas if  $h > q$  the forecast is

$$y_{T+h,T} = 0.$$

Thus, an  $MA(q)$  process is not forecastable (apart from the unconditional mean) more than  $q$  steps ahead. All the dynamics in the  $MA(q)$  process, which we exploit for forecasting, “wash out” by the time we get to horizon  $q$ , which reflects the autocorrelation structure of the  $MA(q)$  process. (Recall that, as we showed earlier, it cuts off at displacement  $q$ .) Second, consider the corresponding forecast errors. They are

$$e_{T+h,T} = MA(h - 1)$$

for  $h \leq q$  and

$$e_{T+h,T} = MA(q)$$

for  $h > q$ . The  $h$ -step-ahead forecast error for  $h > q$  is just the process itself, minus its mean.

Finally, consider the forecast error variances. For  $h \leq q$ ,

$$\sigma_h^2 \leq \text{var}(y_t),$$

whereas for  $h > q$ ,

$$\sigma_h^2 = \text{var}(y_t).$$

In summary, we've thus far studied the  $MA(2)$ , and then the general  $MA(q)$ , process, computing the optimal h-step-ahead forecast, the corresponding forecast error, and the forecast error variance. As we'll now see, the emerging patterns that we cataloged turn out to be quite general.

### 7.3.2 Optimal Point Forecasts for Infinite-Order Moving Averages

By now you're getting the hang of it, so let's consider the general case of an infinite-order  $MA$  process. The infinite-order moving average process may seem like a theoretical curiosity, but precisely the opposite is true. Any covariance stationary process can be written as a (potentially infinite-order) moving average process, and moving average processes are easy to understand and manipulate, because they are written in terms of white noise shocks, which have very simple statistical properties. Thus, if you take the time to understand the mechanics of constructing optimal forecasts for infinite moving-average processes, you'll understand everything, and you'll have some powerful technical tools and intuition at your command.

Recall that the general linear process is

$$y_t = \sum_{i=0}^{\infty} b_i \varepsilon_{t-i},$$

where

$$\varepsilon_t \sim WN(0, \sigma^2)$$

$$\begin{aligned} b_0 &= 1 \\ \sigma^2 \sum_{i=0}^{\infty} b_i^2 &< \infty. \end{aligned}$$

We proceed in the usual way. We first write out the process at the future

time of interest:

$$y_{T+h} = \varepsilon_{T+h} + b_1\varepsilon_{T+h-1} + \dots + b_h\varepsilon_T + b_{h+1}\varepsilon_{T-1} + \dots$$

Then we project  $y_{T+h}$  on the time- $T$  information set. The projection yields zeros for all of the future  $\varepsilon$ 's (because they are white noise and hence unforecastable), leaving

$$y_{T+h,T} = b_h\varepsilon_T + b_{h+1}\varepsilon_{T-1} + \dots$$

It follows that the  $h$ -step ahead forecast error is serially correlated; it follows an  $MA(h-1)$  process,

$$e_{T+h,T} = (y_{T+h} - y_{T+h,T}) = \sum_{i=0}^{h-1} b_i\varepsilon_{T+h-i},$$

with mean 0 and variance

$$\sigma_h^2 = \sigma^2 \sum_{i=0}^{h-1} b_i^2.$$

A number of remarks are in order concerning the optimal forecasts of the general linear process, and the corresponding forecast errors and forecast error variances. First, the 1-step-ahead forecast error is simply  $\varepsilon_{T+1}$ .  $\varepsilon_{T+1}$  is that part of  $y_{T+1}$  that can't be linearly forecast on the basis of  $\Omega_t$  (which, again, is why it is called the innovation). Second, although it might at first seem strange that an *optimal* forecast error would be serially correlated, as is the case when  $h > 1$ , nothing is awry. The serial correlation can't be used to improve forecasting performance, because the autocorrelations of the  $MA(h-1)$  process cut off just before the beginning of the time- $T$  information set  $\varepsilon_T, \varepsilon_{T-1}, \dots$ . This is a general and tremendously important property of the errors associated with optimal forecasts: *errors from optimal forecasts can't be forecast using information available when the forecast was made*. If you can forecast the forecast error, then you can improve the forecast, which means that it couldn't have been optimal. Finally, note that as  $h$  approaches

infinity  $y_{T+h,T}$  approaches zero, the unconditional mean of the process, and  $\sigma_h^2$  approaches  $\sigma^2 \sum_{i=0}^{\infty} b_i^2$ , the unconditional variance of the process, which reflects the fact that as  $h$  approaches infinity the conditioning information on which the forecast is based becomes progressively less useful. In other words, the distant future is harder to forecast than the near future!

### 7.3.3 Interval and Density Forecasts

Now we construct interval and density forecasts. Regardless of whether the moving average is finite or infinite, we proceed in the same way, as follows. The definition of the  $h$ -step-ahead forecast error is

$$e_{T+h,T} = y_{T+h} - y_{T+h,T}.$$

Equivalently, the  $h$ -step-ahead realized value,  $y_{T+h}$ , equals the forecast plus the error,

$$y_{T+h} = y_{T+h,T} + e_{T+h,T}.$$

If the innovations are normally distributed, then the future value of the series of interest is also normally distributed, conditional upon the information set available at the time the forecast was made, and so we have the 95%  $h$ -step-ahead interval forecast  $y_{T+h,T} \pm 1.96\sigma_h$ .<sup>12</sup> In similar fashion, we construct the  $h$ -step-ahead density forecast as

$$N(y_{T+h,T}, \sigma_h^2).$$

The mean of the conditional distribution of  $y_{T+h}$  is  $y_{T+h,T}$ , which of course must be the case because we constructed the point forecast as the conditional mean, and the variance of the conditional distribution is  $\sigma_h^2$ , the variance of

---

<sup>12</sup>Confidence intervals at any other desired confidence level may be constructed in similar fashion, by using a different critical point of the standard normal distribution. A 90% interval forecast, for example, is  $y_{T+h,T} \pm 1.64\sigma_h$ . In general, for a Gaussian process, a  $(1 - \alpha) \cdot 100\%$  confidence interval is  $y_{T+h,T} \pm z_{\alpha/2}\sigma_h$ , where  $z_{\alpha/2}$  is that point on the  $N(0, 1)$  distribution such that  $\text{prob}(z > z_{\alpha/2}) = \alpha/2$ .

the forecast error.

As an example of interval and density forecasting, consider again the  $MA(2)$  process,

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

Assuming normality, the 1-step-ahead 95% interval forecast is

$$y_{T+1,T} = (\theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1}) \pm 1.96\sigma,$$

and the 1-step-ahead density forecast is

$$N(\theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1}, \sigma^2).$$

### 7.3.4 Making the Forecasts Operational

So far we've assumed that the parameters of the process being forecast are known. In practice, of course, they must be estimated. To make our forecasting procedures operational, we simply replace the unknown parameters in our formulas with estimates, and the unobservable innovations with residuals.

Consider, for example, the  $MA(2)$  process,

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}.$$

As you can readily verify using the methods we've introduced, the 2-step ahead optimal forecast, assuming known parameters, is

$$y_{T+2,T} = \theta_2 \varepsilon_T,$$

with corresponding forecast error

$$e_{T+2,T} = \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1},$$

and forecast-error variance

$$\sigma_2^2 = \sigma^2(1 + \theta_1^2).$$

To make the forecast operational, we replace unknown parameters with estimates and the time- $T$  innovation with the time- $T$  residual, yielding

$$\hat{y}_{T+2,T} = \hat{\theta}_2 \hat{\varepsilon}_T$$

and forecast error variance

$$\hat{\sigma}_2^2 = \hat{\sigma}^2(1 + \hat{\theta}_1^2).$$

Then, if desired, we can construct operational 2-step-ahead interval and density forecasts, as

$$\hat{y}_{T+2,T} \pm z_{\alpha/2} \hat{\sigma}_2$$

and

$$N(\hat{y}_{T+2,T}, \hat{\sigma}_2^2).$$

The strategy of taking a forecast formula derived under the assumption of known parameters, and replacing unknown parameters with estimates, is a natural way to operationalize the construction of point forecasts. However, using the same strategy to produce operational interval or density forecasts involves a subtlety that merits additional discussion. The forecast error variance estimate so obtained can be interpreted as one that ignores parameter estimation uncertainty, as follows. Recall once again that the actual future value of the series is

$$y_{T+2} = \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1} + \theta_2 \varepsilon_T,$$

and that the operational forecast is

$$\hat{y}_{T+2,T} = \hat{\theta}_2 \varepsilon_T.$$

Thus the exact forecast error is

$$\hat{e}_{T+2,T} = y_{T+2} - \hat{y}_{T+2,T} = \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1} + (\theta_2 - \hat{\theta}_2) \varepsilon_T,$$

the variance of which is very difficult to evaluate. So we make a convenient approximation: we ignore parameter estimation uncertainty by assuming that estimated parameters equal true parameters. We therefore set

$$(\theta_2 - \hat{\theta}_2)$$

to zero, which yields

$$\hat{e}_{T+2,T} = \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1},$$

with variance

$$\sigma_2^2 = \sigma^2(1 + \theta_1^2),$$

which we make operational as

$$\hat{\sigma}_2^2 = \hat{\sigma}^2(1 + \hat{\theta}_1^2).$$

## 7.4 Forecasting Cycles From an Autoregressive Perspective: Wold's Chain Rule

### 7.4.1 Point Forecasts of Autoregressive Processes

Because any covariance stationary  $AR(p)$  process can be written as an infinite moving average, there's no need for specialized forecasting techniques for autoregressions. Instead, we can simply transform the autoregression into a moving average, and then use the techniques we developed for forecasting

moving averages. It turns out, however, that a very simple recursive method for computing the optimal forecast is available in the autoregressive case.

The recursive method, called the **chain rule of forecasting**, is best learned by example. Consider the  $AR(1)$  process,

$$y_t = \phi y_{t-1} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

First we construct the optimal 1-step-ahead forecast, and then we construct the optimal 2-step-ahead forecast, which depends on the optimal 1-step-ahead forecast, which we've already constructed. Then we construct the optimal 3-step-ahead forecast, which depends on the already-computed 2-step-ahead forecast, which we've already constructed, and so on.

To construct the 1-step-ahead forecast, we write out the process for time  $T + 1$ ,

$$y_{T+1} = \phi y_T + \varepsilon_{T+1}.$$

Then, projecting the right-hand side on the time- $T$  information set, we obtain

$$y_{T+1,T} = \phi y_T.$$

Now let's construct the 2-step-ahead forecast. Write out the process for time  $T + 2$ ,

$$y_{T+2} = \phi y_{T+1} + \varepsilon_{T+2}.$$

Then project directly on the time- $T$  information set to get

$$y_{T+2,T} = \phi y_{T+1,T}.$$

Note that the future innovation is replaced by 0, as always, and that we have directly replaced the time  $T + 1$  value of  $y$  with its earlier-constructed optimal forecast. Now let's construct the 3-step-ahead forecast. Write out the process

for time  $T + 3$ ,

$$y_{T+3} = \phi y_{T+2} + \varepsilon_{T+3}.$$

Then project directly on the time- $T$  information set,

$$y_{T+3,T} = \phi y_{T+2,T}.$$

The required 2-step-ahead forecast was already constructed.

Continuing in this way, we can recursively build up forecasts for any and all future periods. Hence the name “chain rule of forecasting.” Note that, for the  $AR(1)$  process, only the most recent value of  $y$  is needed to construct optimal forecasts, for any horizon, and for the general  $AR(p)$  process only the  $p$  most recent values of  $y$  are needed.

#### 7.4.2 Point Forecasts of ARMA processes

Now we consider forecasting covariance stationary ARMA processes. Just as with autoregressive processes, we could always convert an ARMA process to an infinite moving average, and then use our earlier-developed methods for forecasting moving averages. But also as with autoregressive processes, a simpler method is available for forecasting ARMA processes directly, by combining our earlier approaches to moving average and autoregressive forecasting.

As always, we write out the  $ARMA(p, q)$  process for the future period of interest,

$$y_{T+h} = \phi_1 y_{T+h-1} + \dots + \phi_p y_{T+h-p} + \varepsilon_{T+h} + \theta_1 \varepsilon_{T+h-1} + \dots + \theta_q \varepsilon_{T+h-q}.$$

On the right side we have various future values of  $y$  and  $\varepsilon$ , and perhaps also past values, depending on the forecast horizon. We replace everything on the right-hand side with its projection on the time- $T$  information set. That is, we replace all future values of  $y$  with optimal forecasts (built up recursively

using the chain rule) and all future values of  $\varepsilon$  with optimal forecasts (0), yielding

$$y_{T+h,T} = \phi_1 y_{T+h-1,T} + \dots + \phi_p y_{T+h-p,T} + \varepsilon_{T+h,T} + \theta_1 \varepsilon_{T+h-1,T} + \dots + \theta_q \varepsilon_{T+h-q,T}.$$

When evaluating this formula, note that the optimal time- $T$  “forecast” of any value of  $y$  or  $\varepsilon$  dated time  $T$  or earlier is just  $y$  or  $\varepsilon$  itself.

As an example, consider forecasting the  $ARMA(1, 1)$  process,

$$y_t = \phi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

Let's find  $y_{T+1,T}$ . The process at time  $T + 1$  is

$$y_{T+1} = \phi y_T + \varepsilon_{T+1} + \theta \varepsilon_T.$$

Projecting the right-hand side on  $\Omega_T$  yields

$$y_{T+1,T} = \phi y_T + \theta \varepsilon_T.$$

Now let's find  $y_{T+2,T}$ . The process at time  $T + 2$  is

$$y_{T+2} = \phi y_{T+1} + \varepsilon_{T+2} + \theta \varepsilon_{T+1}.$$

Projecting the right-hand side on  $\Omega_T$  yields

$$y_{T+2,T} = \phi y_{T+1,T}.$$

Substituting our earlier-computed 1-step-ahead forecast yields

$$y_{T+2,T} = \phi (\phi y_T + \theta \varepsilon_T) \tag{7.1}$$

$$= \phi^2 y_T + \phi \theta \varepsilon_T. \tag{7.2}$$

Continuing, it is clear that

$$y_{T+h,T} = \phi y_{T+h-1,T},$$

for all  $h > 1$ .

### 7.4.3 Interval and Density Forecasts

The chain rule, whether applied to pure autoregressive models or to ARMA models, is a device for simplifying the computation of point forecasts. Interval and density forecasts require the  $h$ -step-ahead forecast error variance, which we get from the moving average representation, as discussed earlier. It is

$$\sigma_h^2 = \sigma^2 \sum_{i=0}^{h-1} b_i^2,$$

which we operationalize as

$$\hat{\sigma}_h^2 = \hat{\sigma}^2 \sum_{i=0}^{h-1} \hat{b}_i^2.$$

Note that we don't actually estimate the moving average representation; rather, we solve backward for as many  $b$ 's as we need, in terms of the original model parameters, which we then replace with estimates.

Let's illustrate by constructing a 2-step-ahead 95% interval forecast for the  $ARMA(1, 1)$  process. We already constructed the 2-step-ahead point forecast,  $y_{T+2,T}$ ; we need only compute the 2-step-ahead forecast error variance. The process is

$$y_t = \phi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$$

Substitute backward for  $y_{t-1}$  to get

$$y_t = \phi(\phi y_{t-2} + \varepsilon_{t-1} + \theta \varepsilon_{t-2}) + \varepsilon_t + \theta \varepsilon_{t-1} \quad (7.3)$$

$$= \varepsilon_t + (\phi + \theta) \varepsilon_{t-1} + \dots \quad (7.4)$$

We need not substitute back any farther, because the 2-step-ahead forecast error variance is

$$\sigma_2^2 = \sigma^2(1 + b_1^2),$$

where  $b_1$  is the coefficient on  $\varepsilon_{t-1}$  in the moving average representation of the ARMA(1,1) process, which we just calculated to be  $(\phi + \theta)$ . Thus the 2-step-ahead interval forecast is  $y_{T+2,T} \pm 1.96\sigma_2$ , or  $(\phi^2 y_T + \phi\theta\varepsilon_T) \pm 1.96\sigma\sqrt{1 + (\phi + \theta)^2}$ . We make this operational as  $(\hat{\phi}^2 y_T + \hat{\phi}\hat{\theta}\varepsilon_T) \pm 1.96\hat{\sigma}\sqrt{1 + (\hat{\phi} + \hat{\theta})^2}$ .

## 7.5 Canadian Employment

We earlier examined the correlogram for the Canadian employment series, and we saw that the sample autocorrelations damp slowly and the sample partial autocorrelations cut off, just the opposite of what's expected for a moving average. Thus the correlogram indicates that a finite-order moving average process would not provide a good approximation to employment dynamics. Nevertheless, nothing stops us from fitting moving average models, so let's fit them and use the AIC and the SIC to guide model selection.

Moving average models are nonlinear in the parameters; thus, estimation proceeds by nonlinear least squares (numerical minimization). The idea is the same as when we encountered nonlinear least squares in our study of nonlinear trends – pick the parameters to minimize the sum of squared residuals – but finding an expression for the residual is a little bit trickier. To understand why moving average models are nonlinear in the parameters, and to get a feel for how they're estimated, consider an invertible MA(1) model, with a

nonzero mean explicitly included for added realism,

$$y_t = \mu + \varepsilon_t + \theta \varepsilon_{t-1}.$$

Substitute backward  $m$  times to obtain the autoregressive approximation

$$y_t \approx \frac{\mu}{1+\theta} + \theta y_{t-1} - \theta^2 y_{t-2} + \dots + (-1)^{m+1} \theta^m y_{t-m} + \varepsilon_t.$$

Thus an invertible moving average can be approximated as a finite-order autoregression. The larger is  $m$ , the better the approximation. This lets us (approximately) express the residual in terms of observed data, after which we can use a computer to solve for the parameters that minimize the sum of squared residuals,

$$\hat{\mu}, \hat{\theta} = \underset{\mu, \theta}{\operatorname{argmin}} \sum_{t=1}^T \left[ y_t - \left( \frac{\mu}{1+\theta} + \theta y_{t-1} - \theta^2 y_{t-2} + \dots + (-1)^{m+1} \theta^m y_{t-m} \right) \right]^2$$

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T \left[ y_t - \left( \frac{\hat{\mu}}{1+\hat{\theta}} + \hat{\theta} y_{t-1} - \hat{\theta}^2 y_{t-2} + \dots + (-1)^{m+1} \hat{\theta}^m y_{t-m} \right) \right]^2.$$

The parameter estimates must be found using numerical optimization methods, because the parameters of the autoregressive approximation are restricted. The coefficient of the second lag of  $y$  is the square of the coefficient on the first lag of  $y$ , and so on. The parameter restrictions must be imposed in estimation, which is why we can't simply run an ordinary least squares regression of  $y$  on lags of itself.

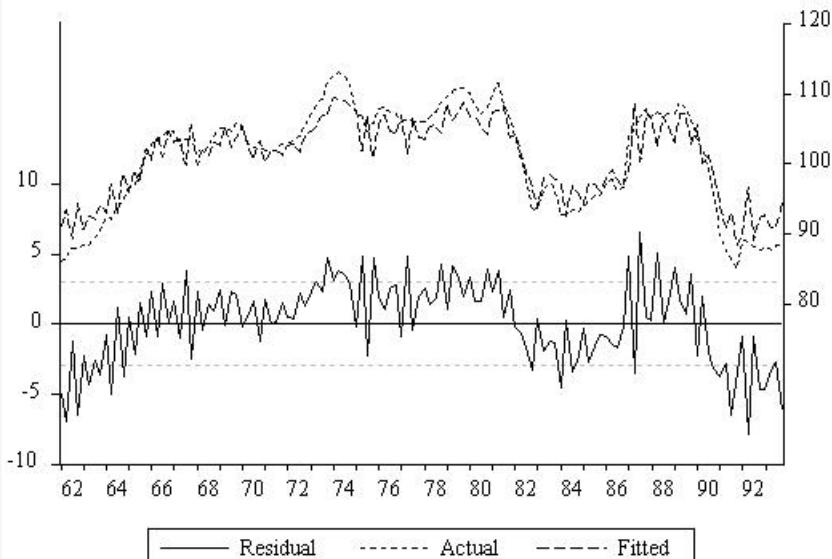
The next step would be to estimate  $MA(q)$  models,  $q = 1, 2, 3, 4$ . Both the  $AIC$  and the  $SIC$  suggest that the  $MA(4)$  is best. To save space, we

report only the results of  $MA(4)$  estimation in Table 7.12a. The results of the  $MA(4)$  estimation, although better than lower-order  $MAs$ , are nevertheless poor. The  $R^2$  of .84 is rather low, for example, and the Durbin-Watson statistic indicates that the  $MA(4)$  model fails to account for all the serial correlation in employment. The residual plot, which we show in Figure 7.12b, clearly indicates a neglected cycle, an impression confirmed by the residual correlogram (Table 7.13, Figure 7.14).

LS // Dependent Variable is CANEMP				
Sample: 1962:1 1993:4				
Included observations: 128				
Convergence achieved after 49 iterations				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	100.5438	0.843322	119.2234	0.0000
MA(1)	1.587641	0.063908	24.84246	0.0000
MA(2)	0.994369	0.089995	11.04917	0.0000
MA(3)	-0.020305	0.046550	-0.436189	0.6635
MA(4)	-0.298387	0.020489	-14.56311	0.0000
R-squared	0.849951	Mean dependent var	101.0176	
Adjusted R-squared	0.845071	S.D. dependent var	7.499163	
S.E. of regression	2.951747	Akaike info criterion	2.203073	
Sum squared resid	1071.676	Schwarz criterion	2.314481	
Log likelihood	-317.6208	F-statistic	174.1826	
Durbin-Watson stat	1.246600	Prob(F-statistic)	0.000000	
Inverted MA Roots	.41	-.56+.72i	-.56-.72i	-.87

(a) Employment MA(4) Regression

Residual Plot



(b) Employment MA(4) Residual Plot

Figure 7.12: Employment: MA(4) Model

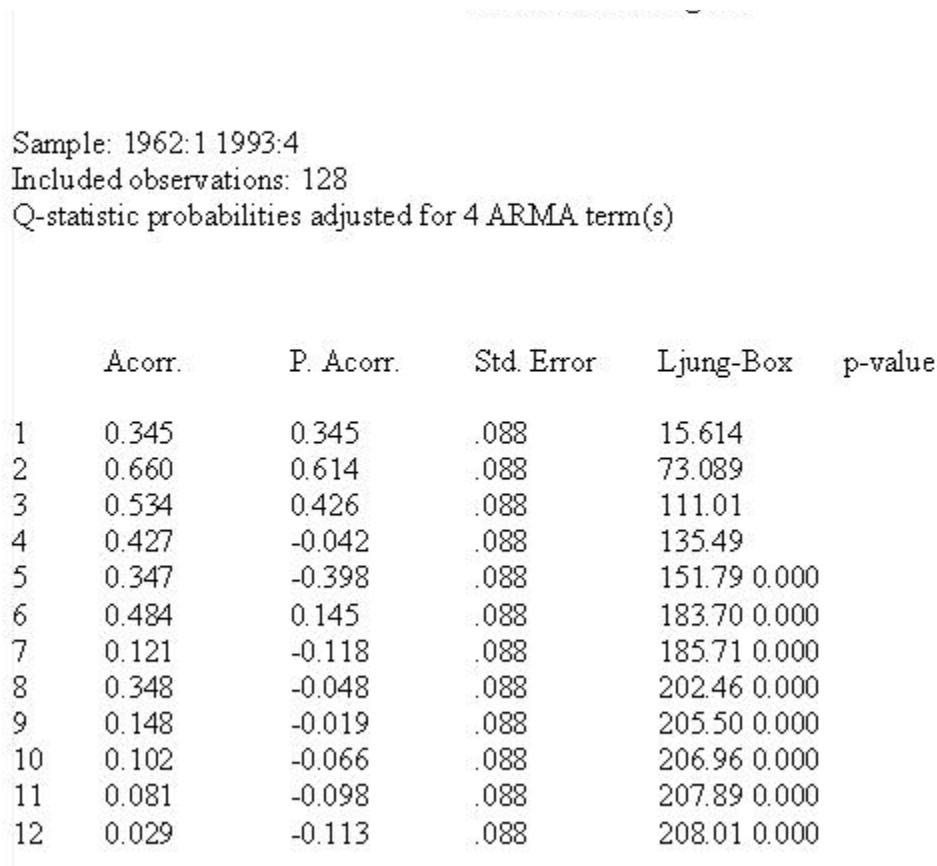


Figure 7.13: Employment MA(4) Residual Correlogram

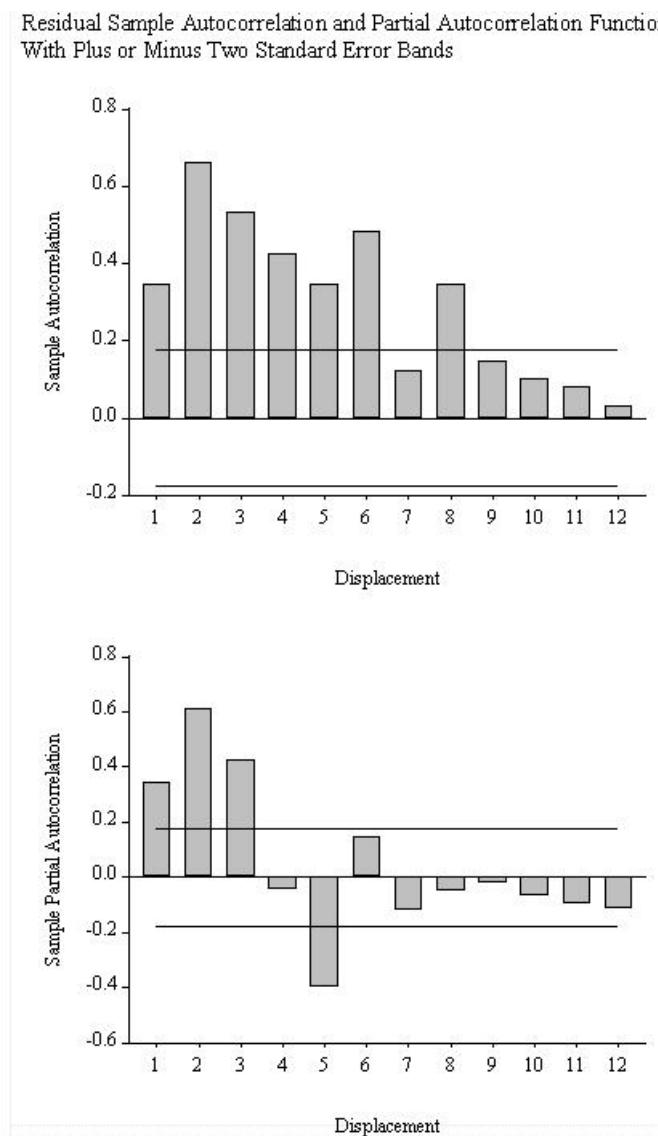


Figure 7.14: Employment MA(4) Residual Sample Autocorrelation and Partial Autocorrelation

If we insist on using a moving average model, we'd want to explore orders greater than four, but all the results thus far indicate that moving average processes don't provide good approximations to employment dynamics. Thus let's consider alternative approximations, such as autoregressions. Autoregressions can be conveniently estimated by ordinary least squares regression. Consider, for example, the  $AR(1)$  model,

$$\begin{aligned} (y_t - \mu) &= \phi(y_{t-1} - \mu) + \varepsilon_t \\ \varepsilon_t &\sim (0, \sigma^2) \end{aligned}$$

We can write it as

$$y_t = c + \phi y_{t-1} + \varepsilon_t$$

where  $c = \mu(1 - \phi)$ . The least squares estimators are

$$\begin{aligned} \hat{c}, \hat{\phi} &= \underset{c, \phi}{\operatorname{argmin}} \sum_{t=1}^T [y_t - c - \phi y_{t-1}]^2 \\ \hat{\sigma}^2 &= \frac{1}{T} \sum_{t=1}^T [y_t - \hat{c} - \hat{\phi} y_{t-1}]^2. \end{aligned}$$

The implied estimate of  $\mu$  is

$$\hat{\mu} = \hat{c}/(1 - \hat{\phi}).$$

Unlike the moving average case, for which the sum of squares function is nonlinear in the parameters, requiring the use of numerical minimization methods, the sum of squares function for autoregressive processes is linear in the parameters, so that estimation is particularly stable and easy. In the  $AR(1)$  case, we simply run an ordinary least squares regression of  $y$  on one

lag of  $y$ ; in the  $AR(p)$  case, we regress  $y$  on  $p$  lags of  $y$ .

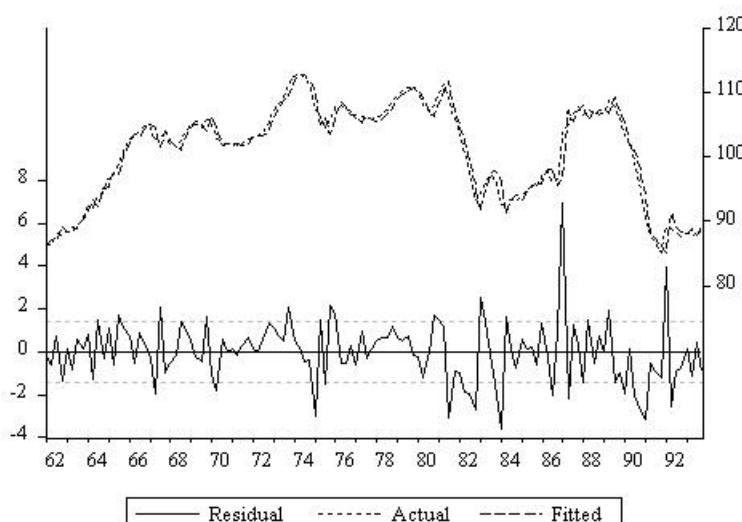
We estimate  $AR(p)$  models,  $p = 1, 2, 3, 4$ . Both the  $AIC$  and the  $SIC$  suggest that the  $AR(2)$  is best. To save space, we report only the results of  $AR(2)$  estimation in Table 7.15a. The estimation results look good, and the residuals (Figure 7.15b) look like white noise. The residual correlogram (Table 7.16, Figure 7.17) supports that conclusion.

LS // Dependent Variable is CANEMP  
 Sample: 1962:1 1993:4  
 Included observations: 128  
 Convergence achieved after 3 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	101.2413	3.399620	29.78017	0.0000
AR(1)	1.438810	0.078487	18.33188	0.0000
AR(2)	-0.476451	0.077902	-6.116042	0.0000
R-squared	0.963372		Mean dependent var	101.0176
Adjusted R-squared	0.962786		S.D. dependent var	7.499163
S.E. of regression	1.446663		Akaike info criterion	0.761677
Sum squared resid	261.6041		Schwarz criterion	0.828522
Log likelihood	-227.3715		F-statistic	1643.837
Durbin-Watson stat	2.067024		Prob(F-statistic)	0.000000
Inverted AR Roots	.92	.52		

(a) Employment AR(2) Model

Residual Plot



(b) Employment AR(2) Residual Plot

Figure 7.15: Employment: MA(4) Model

Sample: 1962:1 1993:4  
 Included observations: 128  
 Q-statistic probabilities adjusted for 2 ARMA term(s)

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	-0.035	-0.035	.088	0.1606	
2	0.044	0.042	.088	0.4115	
3	0.011	0.014	.088	0.4291 0.512	
4	0.051	0.050	.088	0.7786 0.678	
5	0.002	0.004	.088	0.7790 0.854	
6	0.019	0.015	.088	0.8272 0.935	
7	-0.024	-0.024	.088	0.9036 0.970	
8	0.078	0.072	.088	1.7382 0.942	
9	0.080	0.087	.088	2.6236 0.918	
10	0.050	0.050	.088	2.9727 0.936	
11	-0.023	-0.027	.088	3.0504 0.962	
12	-0.129	-0.148	.088	5.4385 0.860	

Figure 7.16: Employment AR(2) Residual Correlogram

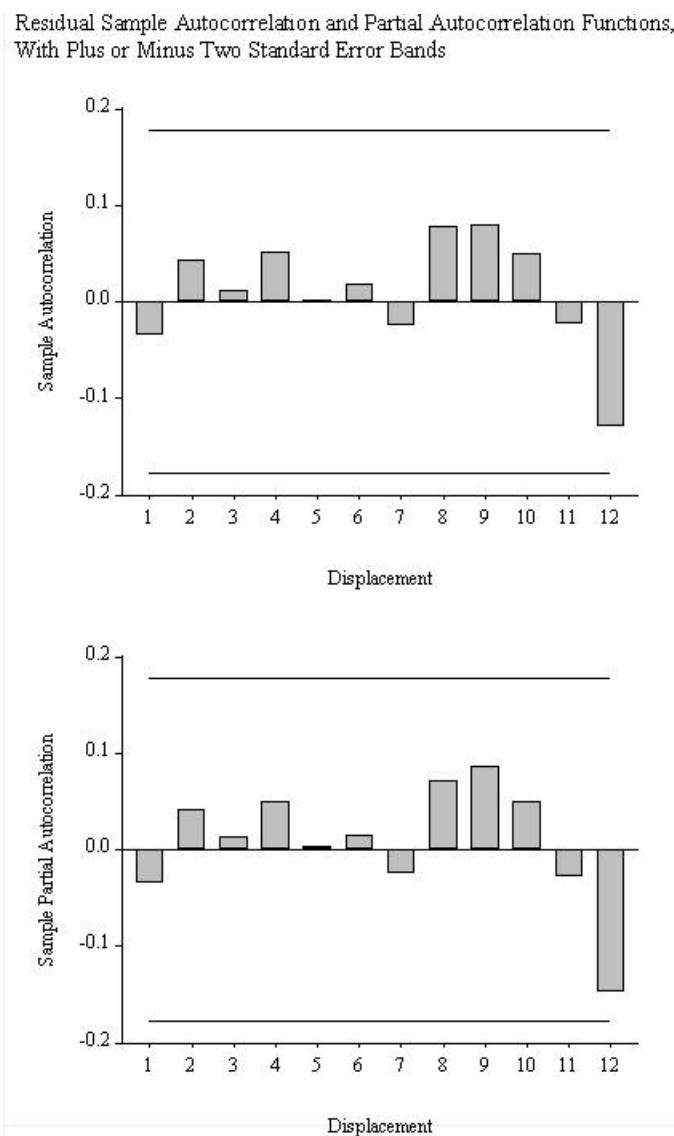


Figure 7.17: Employment AR(2) Residual Sample Autocorrelation and Partial Autocorrelation

				MA Order		
		0	1	2	3	4
	0		2.86	2.32	2.47	2.20
	1	1.01	.83	.79	.80	.81
AR Order	2	.762	.77	.78	.80	.80
	3	.77	.761	.77	.78	.79
	4	.79	.79	.77	.79	.80

(a) Employment AIC Values						
Various ARMA Models						
				MA Order		
		0	1	2	3	4
	0		2.91	2.38	2.56	2.31
	1	1.05	.90	.88	.91	.94
AR Order	2	.83	.86	.89	.92	.96
	3	.86	.87	.90	.94	.96
	4	.90	.92	.93	.97	1.00

(b) Employment SIC Values						
---------------------------	--	--	--	--	--	--

Figure 7.18: Employment - Information Criterion for ARMA Models

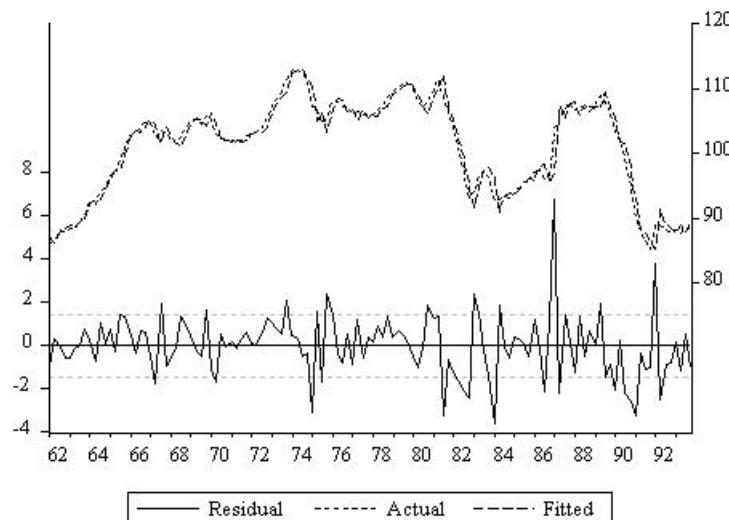
Finally, we consider  $ARMA(p, q)$  approximations to the Wold representation.  $ARMA$  models are estimated in a fashion similar to moving average models; they have autoregressive approximations with nonlinear restrictions on the parameters, which we impose when doing a numerical sum of squares minimization. We examine all  $ARMA(p, q)$  models with  $p$  and  $q$  less than or equal to four; the  $SIC$  and  $AIC$  values appear in Tables 7.18a and 7.18b. The  $SIC$  selects the  $AR(2)$  (an  $ARMA(2, 0)$ ), which we've already discussed. The  $AIC$ , which penalizes degrees of freedom less harshly, selects an  $ARMA(3, 1)$  model. The  $ARMA(3, 1)$  model looks good; the estimation results appear in Table 7.19a, the residual plot in Figure 7.19b, and the residual correlogram in Table 7.20 and Figure fig: employment arma(3,1) residual sample autocorrelation and partial autocorrelation.

LS // Dependent Variable is CANEMP  
 Sample: 1962:1 1993:4  
 Included observations: 128  
 Convergence achieved after 17 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	101.1378	3.538602	28.58130	0.0000
AR(1)	0.500493	0.087503	5.719732	0.0000
AR(2)	0.872194	0.067096	12.99917	0.0000
AR(3)	-0.443355	0.080970	-5.475560	0.0000
MA(1)	0.970952	0.035015	27.72924	0.0000
R-squared	0.964535	Mean dependent var	101.0176	
Adjusted R-squared	0.963381	S. D. dependent var	7.499163	
S.E. of regression	1.435043	Akaike info criterion	0.760668	
Sum squared resid	253.2997	Schwarz criterion	0.872076	
Log likelihood	-225.3069	F-statistic	836.2912	
Durbin-Watson stat	2.057302	Prob(F-statistic)	0.000000	
Inverted AR Roots	.93	.51	-.94	
Inverted MA Roots	-.97			

(a) Employment ARMA(3,1) Model

Residual Plot



(b) Employment ARMA(3,1) Residual Plot

Figure 7.19: Employment: MA(4) Model

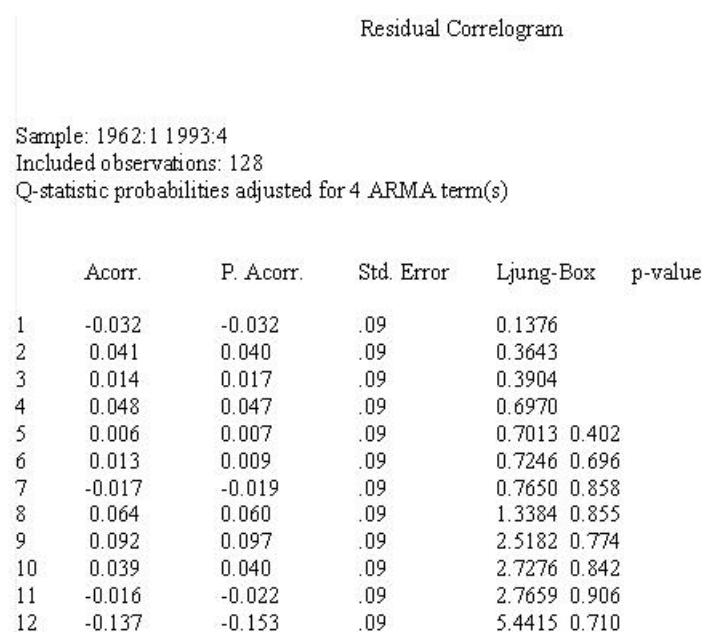


Figure 7.20: Employment ARMA(3,1) Correlogram

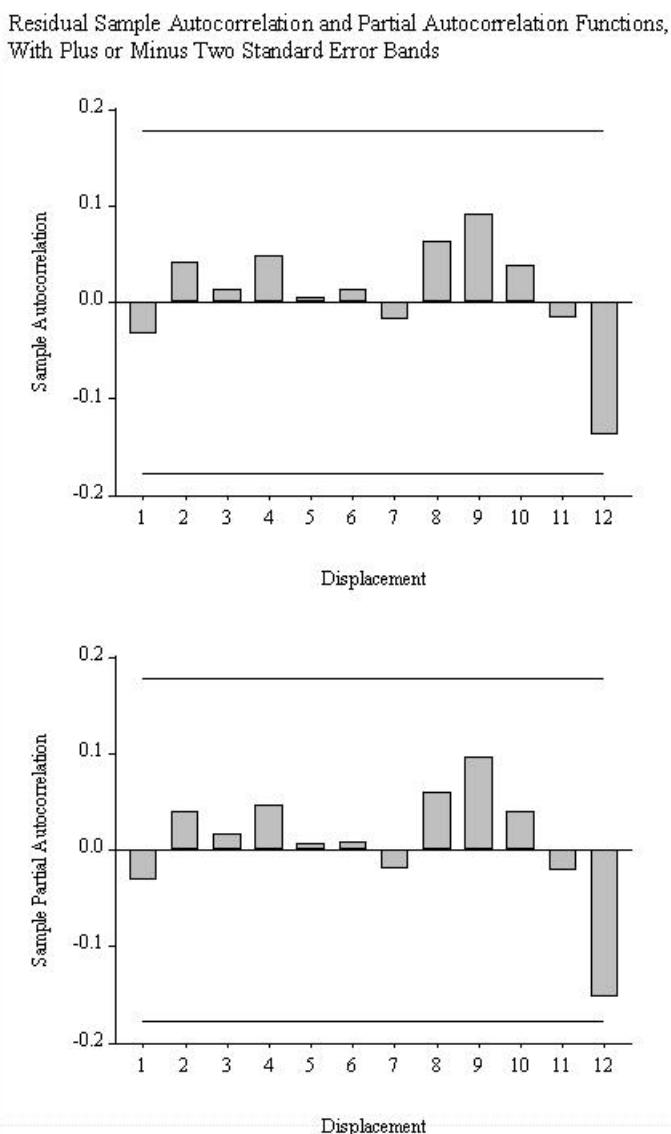


Figure 7.21: Employment ARMA(3,1) Residual Sample Autocorrelation and Partial Auto-correlation

Although the  $ARMA(3, 1)$  looks good, apart from its lower  $AIC$  it looks no better than the  $AR(2)$ , which basically seemed perfect. In fact, there are at least three reasons to prefer the  $AR(2)$ . First, for the reasons that we discussed in Chapter 15 , when the  $AIC$  and the  $SIC$  disagree we recommend using the more parsimonious model selected by the  $SIC$ . Second, if we consider a model selection strategy involving not just examination of the  $AIC$  and  $SIC$ , but also examination of autocorrelations and partial autocorrelations, which we advocate, we're led to the  $AR(2)$ . Finally, and importantly, the impression that the  $ARMA(3, 1)$  provides a richer approximation to employment dynamics is likely spurious in this case. The  $ARMA(3, 1)$  has a inverse autoregressive root of -.94 and an inverse moving average root of -.97. Those roots are of course just *estimates* and are likely to be statistically indistinguishable from one another, in which case we can *cancel* them, which brings us down to an  $ARMA(2, 0)$ , or  $AR(2)$ , model with roots virtually indistinguishable from those of our earlier-estimated  $AR(2)$  process! We refer to this situation as one of **common factors** in an  $ARMA$  model. Look out for such situations, which can lead to substantial model simplification.

Now we put our forecasting technology to work to produce point and interval forecasts for Canadian employment. Recall that the best moving average model was an  $MA(4)$ , while the best autoregressive model, as well as the best  $ARMA$  model and the best model overall, was an  $AR(2)$ .

Consider forecasting with the  $MA(4)$  model. Figure 7.22 shows employment history together with operational 4-quarter-ahead point and interval extrapolation forecasts. The 4-quarter-ahead extrapolation forecast reverts quickly to the mean of the employment index. In 1993.4, the last quarter of historical data, employment is well below its mean, but the forecast calls for a quick rise. The forecasted quick rise seems unnatural, because employment dynamics are historically very persistent. If employment is well below its mean in 1993.4, we'd expect it to stay below its mean for some time.

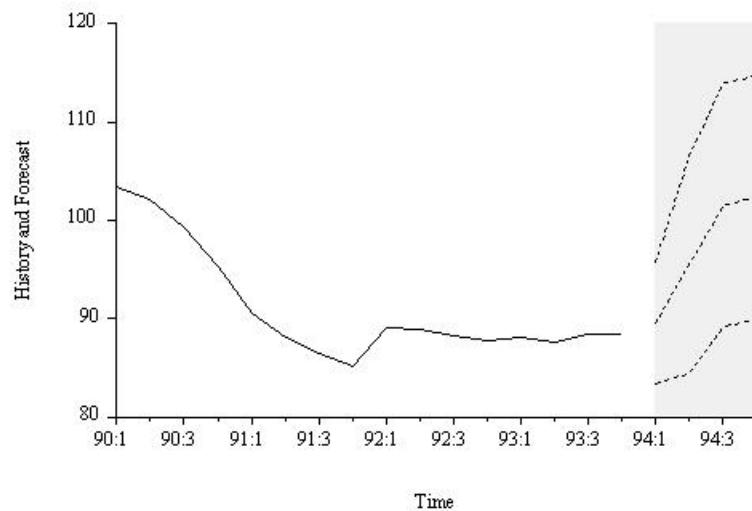


Figure 7.22: Employment History and Forecast - MA(4)

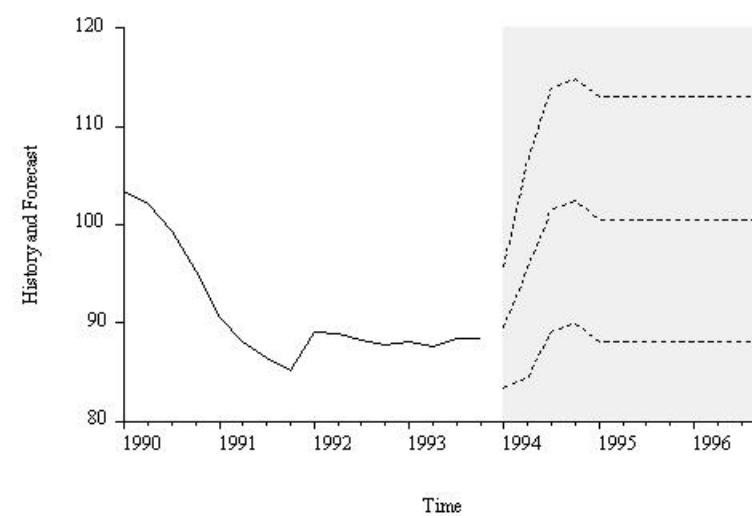


Figure 7.23: Employment History and Long-Horizon Forecast - MA(4)

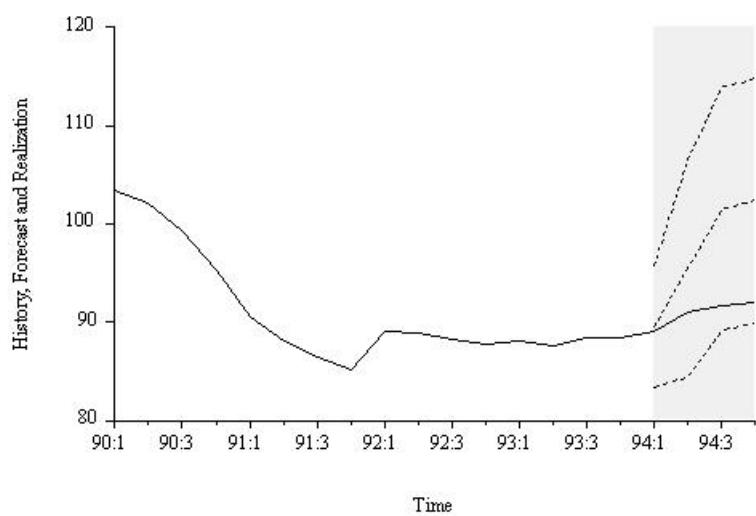


Figure 7.24: Employment History, Forecast, and Realization - MA(4)

The MA(4) model is unable to capture such persistence. The quick reversion of the MA(4) forecast to the mean is a manifestation of the short memory of moving average processes. Recall, in particular, that an MA(4) process has a 4-period memory – all autocorrelations are zero beyond displacement 4. Thus, all forecasts more than four steps ahead are simply equal to the unconditional mean (100.2), and all 95% interval forecasts more than four steps ahead are plus or minus 1.96 unconditional standard deviations. All of this is made clear in Figure 7.23, in which we show the employment history together with 12-step-ahead point and interval extrapolation forecasts.

In Figure 7.24 we show the 4-quarter-ahead forecast and realization. Our suspicions are confirmed. The actual employment series stays well below its mean over the forecast period, whereas the forecast rises quickly back to the mean. The mean squared forecast error is a large 55.9.

Now consider forecasting with the AR(2) model. In Figure 7.25 we show the 4-quarter-ahead extrapolation forecast, which reverts to the unconditional mean much less quickly, as seems natural given the high persistence of employment. The 4-quarter-ahead point forecast, in fact, is still well below the mean. Similarly, the 95% error bands grow gradually and haven't approached their long-horizon values by four quarters out.

Figures 7.26 and 7.28 make clear the very different nature of the autoregressive forecasts. Figure 7.26 presents the 12-step-ahead extrapolation forecast, and Figure 7.28 presents a much longer-horizon extrapolation forecast. Eventually the unconditional mean *is* approached, and eventually the error bands do go flat, but only for very long-horizon forecasts, due to the high persistence in employment, which the AR(2) model captures.

In Figure 7.27 we show the employment history, 4-quarter-ahead AR(2) extrapolation forecast, and the realization. The AR(2) forecast appears quite accurate; the mean squared forecast error is 1.3, drastically smaller than that of the MA(4) forecast.

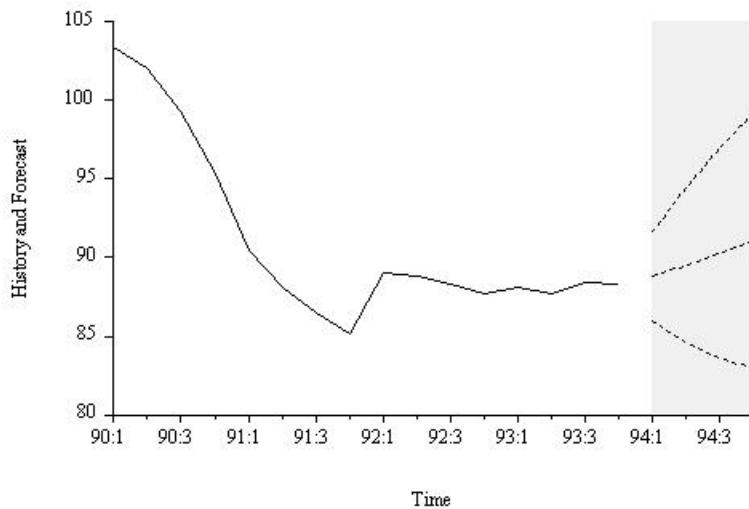


Figure 7.25: Employment History and Forecast - AR(2)

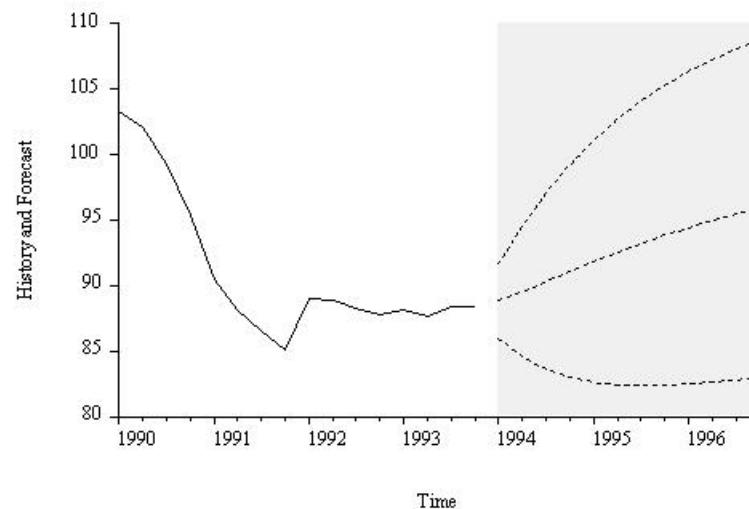


Figure 7.26: Employment History and Forecast, 12-step ahead - AR(2)

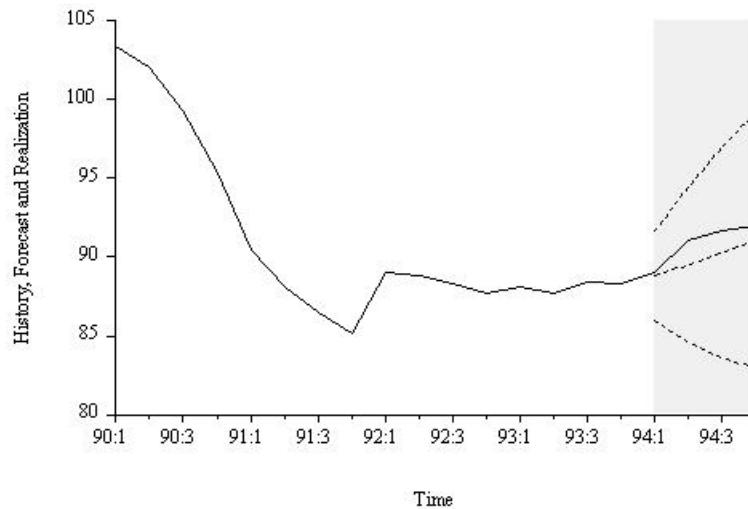


Figure 7.27: Employment History, Forecast, and Realization - AR(2)

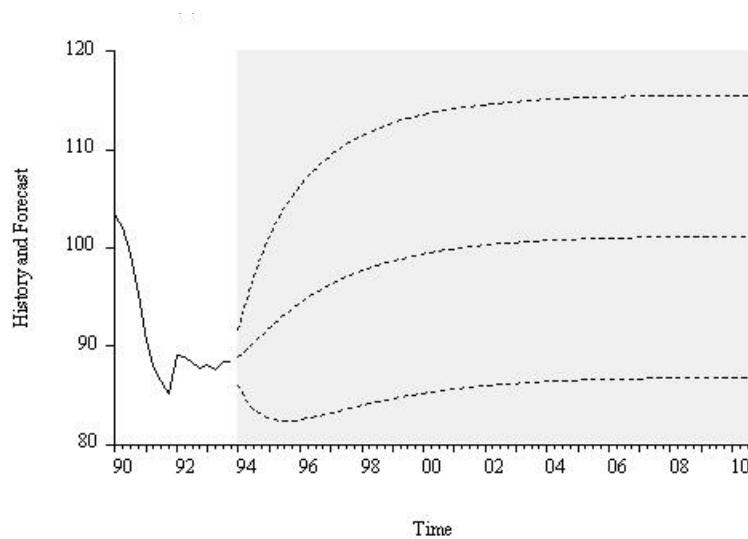


Figure 7.28: Employment History and Long-Horizon Forecast - AR(2)

## 7.6 Exercises, Problems and Complements

1. Shapes of correlograms.

Given the following ARMA processes, sketch the expected forms of the autocorrelation and partial autocorrelation functions. (Hint: examine the roots of the various autoregressive and moving average lag operator polynomials.)

$$(a) \quad y_t = \left( \frac{1}{1 - 1.05L - .09L^2} \right) \varepsilon_t$$

$$(b) \quad y_t = (1 - .4L)\varepsilon_t$$

$$(c) \quad y_t = \left( \frac{1}{1 - .7L} \right) \varepsilon_t.$$

2. The autocovariance function of the  $MA(1)$  process, revisited.

In the text we wrote

$$\gamma(\tau) = E(y_t y_{t-\tau}) = E((\varepsilon_t + \theta \varepsilon_{t-1})(\varepsilon_{t-\tau} + \theta \varepsilon_{t-\tau-1})) = \begin{cases} \theta \sigma^2, & \tau = 1 \\ 0, & \text{otherwise} \end{cases}.$$

Fill in the missing steps by evaluating explicitly the expectation

$$E((\varepsilon_t + \theta \varepsilon_{t-1})(\varepsilon_{t-\tau} + \theta \varepsilon_{t-\tau-1})).$$

3. ARMA algebra.

Derive expressions for the autocovariance function, autocorrelation function, conditional mean, unconditional mean, conditional variance and unconditional variance of the following processes:

$$(a) \quad y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}$$

$$(b) \quad y_t = \phi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}.$$

4. Mechanics of fitting ARMA models.

You have [data for daily transfers over BankWire](#), a financial wire transfer system in a country responsible for much of the world's finance, over a recent span of 200 business days.

- (a) Is trend or seasonality operative? Defend your answer.
- (b) Find a parsimonious  $ARMA(p, q)$  model that fits well, and defend its adequacy.
- (c) In item 4b above, you were asked to find a parsimonious ARMA(p,q) model that fits the transfer data well, and to defend its adequacy. Repeat the exercise, this time using only the first 175 days for model selection and fitting. Is it necessarily the case that the selected ARMA model will remain the same as when all 200 days are used? Does yours?
- (d) Use your estimated model to produce point and interval forecasts for days 176 through 200. Plot them and discuss the forecast pattern.
- (e) Compare your forecasts to the actual realizations. Do the forecasts perform well? Why or why not?

5. A different way to estimate autoregressive models.

We discussed estimation of autoregressive models using ordinary least squares. We could also write the model as a regression on an intercept, with a serially correlated disturbance. Thus the autoregressive model is

$$y_t = \mu + \varepsilon_t$$

$$\Phi(L)\varepsilon_t = v_t$$

$$v_t \sim WN(0, \sigma^2).$$

We can estimate the model using nonlinear least squares. Eviews and other forecasting packages proceed in precisely that way.<sup>13</sup>

This framework – regression on a constant with serially correlated disturbances – has a number of attractive features. First, the mean of the process is the regression constant term.<sup>14</sup> Second, it leads us naturally toward regression on more than just a constant, as other right-hand side variables can be added as desired.

## 6. Aggregation and disaggregation: top-down vs. bottom-up forecasting models.

Related to the issue of methods and complexity discussed in Chapter 2 is the question of aggregation. Often we want to forecast an aggregate, such as total sales of a manufacturing firm, but we can take either an aggregated or disaggregated approach.

Suppose, for example, that total sales is composed of sales of three products. The aggregated, or top-down, or macro, approach is simply to model and forecast total sales. The disaggregated, or bottom- up, or micro, approach is to model and forecast separately the sales of the individual products, and then to add them together.

- (a) Perhaps surprisingly, it's impossible to know in advance whether the aggregated or disaggregated approach is better. It all depends on the specifics of the situation; the only way to tell is to try both approaches and compare the forecasting results.
- (b) However, in real-world situations characterized by likely model misspecification and parameter estimation uncertainty, there are reasons to suspect that the aggregated approach may be preferable.

---

<sup>13</sup>That's why, for example, information on the number of iterations required for convergence is presented even for estimation of the autoregressive model.

<sup>14</sup>Hence the notation “ $\mu$ ” for the intercept.

First, standard (e.g., linear) models fit to aggregated series may be less prone to specification error, because aggregation can produce approximately linear relationships even when the underlying disaggregated relationships are not linear. Second, if the disaggregated series depend in part on a common factor (e.g., general business conditions) then it will emerge more clearly in the aggregate data. Finally, modeling and forecasting of one aggregated series, as opposed to many disaggregated series, relies on far fewer parameter estimates.

- (c) Of course, if our interest centers on the disaggregated components, then we have no choice but to take a disaggregated approach.
  - (d) Sometimes, even if interest centers on an aggregate, there may no data available for it, but there may be data for relevant components. Consider, for example, forecasting the number of pizzas eaten next year by Penn students. There's no annual series available for "pizzas eaten by Penn students," but there may be series of Penn enrollment, annual U.S. pizza consumption, U.S. population, etc. from which a forecast could be built. This is called "Fermi-izing" the problem, after the great Italian physicist Enrico Fermi. See [Tetlock and Gardner \(2015\)](#), chapter 5.
  - (e) It is possible that an aggregate forecast may be useful in forecasting disaggregated series. Why? (Hint: See Fildes and Stekler, 2000.)
7. Forecasting an  $ARMA(2, 2)$  process.

Consider the  $ARMA(2, 2)$  process:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}.$$

- a. Verify that the optimal 1-step ahead forecast made at time  $T$  is

$$y_{T+1,T} = \phi_1 y_T + \phi_2 y_{T-1} + \theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1}.$$

- b. Verify that the optimal 2-step ahead forecast made at time  $T$  is

$$y_{T+2,T} = \phi_1 y_{T+1,T} + \phi_2 y_T + \theta_2 \varepsilon_T,$$

and express it purely in terms of elements of the time- $T$  information set.

- c. Verify that the optimal 3-step ahead forecast made at time  $T$  is

$$y_{T+3,T} = \phi_1 y_{T+2,T} + \phi_2 y_{T+1,T},$$

and express it purely in terms of elements of the time- $T$  information set.

- d. Show that for any forecast horizon  $h$  greater than or equal to three,

$$y_{T+h,T} = \phi_1 y_{T+h-1,T} + \phi_2 y_{T+h-2,T}.$$

## 8. ARMA lag inclusion.

In our  $MA$  model fitting for employment, why did we leave the  $MA(3)$  term in the preferred  $MA(4)$  model, despite the insignificant  $p$ -value? Discuss costs and benefits of dropping the insignificant  $MA(3)$  term.

## 9. Modeling cyclical dynamics.

As a research analyst at the U.S. Department of Energy, you have been asked to model non-seasonally-adjusted U.S. imports of crude oil.

- (a) Find a suitable time series on the web.
- (b) Create a model that captures the trend in the series.

- (c) Adding to the model from part 9b, create a model with trend and a full set of seasonal dummy variables.
- (d) Observe the residuals of the model from part b and their correlogram. Is there evidence neglected dynamics? If so, what to do?

10. Applied *ARMA* modeling.

Nile.com, a successful on-line bookseller, monitors and forecasts the number of “hits” per day to its web page. You have daily hits data for 1/1/98 through 9/28/98.

- a. Fit and assess the standard linear, quadratic, and log linear trend models.
- b. For a few contiguous days roughly in late April and early May, hits were much higher than usual during a big sale. Do you find evidence of a corresponding group of outliers in the residuals from your trend models? Do they influence your trend estimates much? How should you treat them?
- c. Model and assess the significance of day-of-week effects in Nile.com web page hits.
- d. Select a final model, consisting only of trend and seasonal components, to use for forecasting.
- e. Use your model to forecast Nile.com hits through the end of 1998.
- f. Generalize your earlier trend + seasonal model to allow for cyclical dynamics, if present, via  $ARMA(p, q)$  disturbances. Write the full specification of your model in general notation (e.g., with  $p$  and  $q$  left unspecified).
- g. Estimate all models, corresponding to  $p = 0, 1, 2, 3$  and  $q = 0, 1, 2, 3$ , while leaving the original trend and seasonal specifications intact, and select the one that optimizes *SIC*.

- h. Using the model selected in part 10g, write theoretical expressions for the 1- and 2-day-ahead point forecasts and 95% interval forecasts, using estimated parameters.
- i. Calculate those point and interval forecasts for Nile.com for 9/29 and 9/30.

## 11. Mechanics of fitting ARMA models.

On the book's web page you will find data for daily transfers over BankWire, a financial wire transfer system in a country responsible for much of the world's finance, over a recent span of 200 business days.

- a. Is trend or seasonality operative? Defend your answer.
- b. Find a parsimonious  $ARMA(p, q)$  model that fits well, and defend its adequacy.
- c. Repeat the exercise 11b, this time using only the first 175 days for model selection and fitting. Is it necessarily the case that the selected ARMA model will remain the same as when all 200 days are used? Does yours?
- d. Use your estimated model to produce point and interval forecasts for days 176 through 200. Plot them and discuss the forecast pattern.
- e. Compare your forecasts to the actual realizations. Do the forecasts perform well? Why or why not?
- f. Discuss precisely how your software constructs point and interval forecasts. It should certainly match our discussion in spirit, but it may differ in some of the details. Are you uncomfortable with any of the assumptions made? How, if at all, could the forecasts be improved?

## 7.7 Notes

Our discussion of estimation was a bit fragmented; we discussed estimation of moving average and ARMA models using nonlinear least squares, whereas we discussed estimation of autoregressive models using ordinary least squares. A more unified approach proceeds by writing each model as a regression on an intercept, with a serially correlated disturbance. Thus the moving average model is

$$\begin{aligned} y_t &= \mu + \varepsilon_t \\ \varepsilon_t &= \Theta(L)v_t \\ v_t &\sim WN(0, \sigma^2), \end{aligned}$$

the autoregressive model is

$$\begin{aligned} y_t &= \mu + \varepsilon_t \\ \Phi(L)\varepsilon_t &= v_t \\ v_t &\sim WN(0, \sigma^2), \end{aligned}$$

and the ARMA model is

$$\begin{aligned} y_t &= \mu + \varepsilon_t \\ \Phi(L)\varepsilon_t &= \Theta(L)v_t \\ v_t &\sim WN(0, \sigma^2). \end{aligned}$$

We can estimate each model in identical fashion using nonlinear least squares. Eviews and other forecasting packages proceed in precisely that way.<sup>15</sup>

This framework – regression on a constant with serially correlated disturbances – has a number of attractive features. First, the mean of the process is the regression constant term.<sup>16</sup> Second, it leads us naturally toward re-

---

<sup>15</sup>That's why, for example, information on the number of iterations required for convergence is presented even for estimation of the autoregressive model.

<sup>16</sup>Hence the notation “ $\mu$ ” for the intercept.

gression on more than just a constant, as other right-hand side variables can be added as desired. Finally, it exploits the fact that because autoregressive and moving average models are special cases of the ARMA model, their estimation is also a special case of estimation of the ARMA model.

Our description of estimating ARMA models – compute the autoregressive representation, truncate it, and estimate the resulting approximate model by nonlinear least squares – is conceptually correct but intentionally simplified. The actual estimation methods implemented in modern software are more sophisticated, and the precise implementations vary across software packages. Beneath it all, however, all estimation methods are closely related to our discussion, whether implicitly or explicitly. You should consult your software manual for details.

# Chapter 8

## Noise: Conditional Variance Dynamics

The celebrated Wold decomposition makes clear that every covariance stationary series may be viewed as ultimately driven by underlying weak white noise innovations. Hence it is no surprise that every forecasting model discussed in this book is driven by underlying white noise. To take a simple example, if the series  $y_t$  follows an AR(1) process, then  $y_t = \phi y_{t-1} + \varepsilon_t$ , where  $\varepsilon_t$  is white noise. In some situations it is inconsequential whether  $\varepsilon_t$  is weak or strong white noise, that is, whether  $\varepsilon_t$  is independent, as opposed to merely serially uncorrelated. Hence, so to simplify matters we sometimes

*iid*

assume strong white noise,  $\varepsilon_t \sim (0, \sigma^2)$ . Throughout this book, we have

thus far taken that approach, sometimes explicitly and sometimes implicitly.

When  $\varepsilon_t$  is independent, there is no distinction between the unconditional distribution of  $\varepsilon_t$  and the distribution of  $\varepsilon_t$  conditional upon its past, by definition of independence. Hence  $\sigma^2$  is both the unconditional and conditional variance of  $\varepsilon_t$ . The Wold decomposition, however, does not require that  $\varepsilon_t$  be serially independent; rather it requires only that  $\varepsilon_t$  be serially uncorrelated. If  $\varepsilon_t$  is dependent, then its unconditional and conditional distributions will differ. We denote the unconditional innovation distribution by  $\varepsilon_t \sim (0, \sigma^2)$ .

We are particularly interested in conditional dynamics characterized by **heteroskedasticity**, or time-varying volatility. Hence we denote the conditional distribution by  $\varepsilon_t | \Omega_{t-1} \sim (0, \sigma_t^2)$ , where  $\Omega_{t-1} = \varepsilon_{t-1}, \varepsilon_{t-2}, \dots$ . The conditional variance  $\sigma_t^2$  will in general evolve as  $\Omega_{t-1}$  evolves, which focuses attention on the possibility of time-varying innovation volatility.<sup>1</sup>

Allowing for **time-varying volatility** is crucially important in certain economic and financial contexts. The volatility of financial asset returns, for example, is often time-varying. That is, markets are sometimes tranquil and sometimes turbulent, as can readily be seen by examining the time series of stock market returns in Figure 1, to which we shall return in detail. Time-varying volatility has important implications for financial risk management, asset allocation and asset pricing, and it has therefore become central part of the emerging field of **financial econometrics**. Quite apart from financial applications, however, time-varying volatility also has direct implications for interval and density forecasting in a wide variety of applications: correct confidence intervals and density forecasts in the presence of volatility fluctuations require time-varying confidence interval widths and time-varying density forecast spreads. The forecasting models that we have considered thus far, however, do not allow for that possibility. In this chapter we do so.

## 8.1 The Basic ARCH Process

Consider the general linear process,

$$\begin{aligned} y_t &= B(L)\varepsilon_t \\ B(L) &= \sum_{i=0}^{\infty} b_i L^i \end{aligned}$$

---

<sup>1</sup> In principle, aspects of the conditional distribution other than the variance, such as conditional skewness, could also fluctuate. Conditional variance fluctuations are by far the most important in practice, however, so we assume that fluctuations in the conditional distribution of  $\varepsilon$  are due exclusively to fluctuations in  $\sigma_t^2$ .

$$\sum_{i=0}^{\infty} b_i^2 < \infty$$

$$b_0 = 1$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

We will work with various cases of this process.

*iid*

Suppose first that  $\varepsilon_t$  is strong white noise,  $\varepsilon_t \sim WN(0, \sigma^2)$ . Let us

review some results already discussed for the general linear process, which will prove useful in what follows. The *unconditional* mean and variance of  $y$  are

$$E(y_t) = 0$$

and

$$E(y_t^2) = \sigma^2 \sum_{i=0}^{\infty} b_i^2,$$

which are both time-invariant, as must be the case under covariance stationarity. However, the *conditional* mean of  $y$  is time-varying:

$$E(y_t | \Omega_{t-1}) = \sum_{i=1}^{\infty} b_i \varepsilon_{t-i},$$

where the information set is

$$\Omega_{t-1} = \varepsilon_{t-1}, \varepsilon_{t-2}, \dots$$

The ability of the general linear process to capture covariance stationary conditional mean dynamics is the source of its power.

Because the volatility of many economic time series varies, one would hope that the general linear process could capture conditional variance dynamics

as well, but such is not the case for the model as presently specified: the conditional variance of  $y$  is constant at

$$E((y_t - E(y_t | \Omega_{t-1}))^2 | \Omega_{t-1}) = \sigma^2.$$

This potentially unfortunate restriction manifests itself in the properties of the  $h$ -step-ahead conditional prediction error variance. The minimum mean squared error forecast is the conditional mean,

$$E(y_{t+h} | \Omega_t) = \sum_{i=0}^{\infty} b_{h+i} \varepsilon_{t-i},$$

and so the associated prediction error is

$$y_{t+h} - E(y_{t+h} | \Omega_t) = \sum_{i=0}^{h-1} b_i \varepsilon_{t+h-i},$$

which has a conditional prediction error variance of

$$E((y_{t+h} - E(y_{t+h} | \Omega_t))^2 | \Omega_t) = \sigma^2 \sum_{i=0}^{h-1} b_i^2.$$

The conditional prediction error variance is different from the unconditional variance, but it is not time-varying: it depends only on  $h$ , not on the conditioning information  $\Omega_t$ . In the process as presently specified, the conditional variance is not allowed to adapt to readily available and potentially useful conditioning information.

So much for the general linear process with iid innovations. Now we extend it by allowing  $\varepsilon_t$  to be weak rather than strong white noise, *with a particular nonlinear dependence structure*. In particular, suppose that, as before,

$$y_t = B(L)\varepsilon_t$$

$$B(L) = \sum_{i=0}^{\infty} b_i L^i$$

$$\sum_{i=0}^{\infty} b_i^2 < \infty$$

$$b_0 = 1,$$

but now suppose as well that

$$\varepsilon_t | \Omega_{t-1} \sim N(0, \sigma_t^2)$$

$$\sigma_t^2 = \omega + \gamma(L)\varepsilon_t^2$$

$$\omega > 0 \quad \gamma(L) = \sum_{i=1}^p \gamma_i L^i \quad \gamma_i \geq 0 \text{ for all } i \quad \sum \gamma_i < 1.$$

Note that we parameterize the innovation process in terms of its conditional density,

$$\varepsilon_t | \Omega_{t-1},$$

which we assume to be normal with a zero conditional mean and a conditional variance that depends linearly on  $p$  past squared innovations.  $\varepsilon_t$  is serially uncorrelated but not serially independent, because the current conditional variance  $\sigma_t^2$  depends on the history of  $\varepsilon_t$ .<sup>2</sup> The stated regularity conditions are sufficient to ensure that the conditional and unconditional variances are positive and finite, and that  $y_t$  is covariance stationary.

The unconditional moments of  $\varepsilon_t$  are constant and are given by

$$E(\varepsilon_t) = 0$$

---

<sup>2</sup> In particular,  $\sigma_t^2$  depends on the previous  $p$  values of  $\varepsilon_t$  via the distributed lag

$$\gamma(L)\varepsilon_t^2.$$

and

$$E(\varepsilon_t - E(\varepsilon_t))^2 = \frac{\omega}{1 - \sum \gamma_i}.$$

The important result is not the particular formulae for the unconditional mean and variance, but the fact that they are fixed, as required for covariance stationarity. As for the conditional moments of  $\varepsilon_t$ , its conditional variance is time-varying,

$$E((\varepsilon_t - E(\varepsilon_t | \Omega_{t-1}))^2 | \Omega_{t-1}) = \omega + \gamma(L)\varepsilon_t^2,$$

and of course its conditional mean is zero by construction.

Assembling the results to move to the unconditional and conditional moments of  $y$  as opposed to  $\varepsilon_t$ , it is easy to see that both the unconditional mean and variance of  $y$  are constant (again, as required by covariance stationarity), but that both the conditional mean and variance are time-varying:

$$\begin{aligned} E(y_t | \Omega_{t-1}) &= \sum_{i=1}^{\infty} b_i \varepsilon_{t-i} \\ E((y_t - E(y_t | \Omega_{t-1}))^2 | \Omega_{t-1}) &= \omega + \gamma(L)\varepsilon_t^2. \end{aligned}$$

Thus, we now treat conditional mean and variance dynamics in a symmetric fashion by allowing for movement in each, as determined by the evolving information set  $\Omega_{t-1}$ . In the above development,  $\varepsilon_t$  is called an **ARCH(p)** process, and the full model sketched is an infinite-ordered moving average with ARCH(p) innovations, where ARCH stands for autoregressive conditional heteroskedasticity. Clearly  $\varepsilon_t$  is conditionally heteroskedastic, because its conditional variance fluctuates. There are many models of conditional heteroskedasticity, but most are designed for cross-sectional contexts, such as when the variance of a cross-sectional regression disturbance depends on

one or more of the regressors.<sup>3</sup> However, heteroskedasticity is often present as well in the time-series contexts relevant for forecasting, particularly in financial markets. The particular conditional variance function associated with the ARCH process,

$$\sigma_t^2 = \omega + \gamma(L)\varepsilon_t^2 ,$$

is tailor-made for time-series environments, in which one often sees **volatility clustering**, such that large changes tend to be followed by large changes, and small by small, *of either sign*. That is, one may see persistence, or serial correlation, in **volatility dynamics** (conditional variance dynamics), quite apart from persistence (or lack thereof) in conditional mean dynamics. The ARCH process approximates volatility dynamics in an autoregressive fashion; hence the name *autoregressive* conditional heteroskedasticity. To understand why, note that the ARCH conditional variance function links today's conditional variance positively to earlier lagged  $\varepsilon_t^2$ 's, so that large  $\varepsilon_t^2$ 's in the recent past produce a large conditional variance today, thereby increasing the likelihood of a large  $\varepsilon_t^2$  today. Hence ARCH processes are to conditional variance dynamics precisely as standard autoregressive processes are to conditional mean dynamics. The ARCH process may be viewed as a model for the disturbance in a broader model, as was the case when we introduced it above as a model for the innovation in a general linear process. Alternatively, if there are no conditional mean dynamics of interest, the ARCH process may be used for an observed series. It turns out that financial asset returns often have negligible conditional mean dynamics but strong conditional variance dynamics; hence in much of what follows we will view the ARCH process as a model for an observed series, which for convenience we will sometimes call a "return."

---

<sup>3</sup> The variance of the disturbance in a model of household expenditure, for example, may depend on income.

## 8.2 The GARCH Process

Thus far we have used an ARCH(p) process to model conditional variance dynamics. We now introduce the **GARCH(p,q)** process (GARCH stands for generalized ARCH), which we shall subsequently use almost exclusively. As we shall see, GARCH is to ARCH (for conditional variance dynamics) as ARMA is to AR (for conditional mean dynamics).

The pure GARCH(p,q) process is given by<sup>4</sup>

$$y_t = \varepsilon_t$$

$$\begin{aligned} \varepsilon_t | \Omega_{t-1} &\sim N(0, \sigma_t^2) \\ \sigma_t^2 &= \omega + \alpha(L)\varepsilon_t^2 + \beta(L)\sigma_t^2 \\ \alpha(L) &= \sum_{i=1}^p \alpha_i L^i, \quad \beta(L) = \sum_{i=1}^q \beta_i L^i \\ \omega > 0, \quad \alpha_i \geq 0, \quad \beta_i \geq 0, \quad \sum \alpha_i + \sum \beta_i &< 1. \end{aligned}$$

The stated conditions ensure that the conditional variance is positive and that  $y_t$  is covariance stationary.

Back substitution on  $\sigma_t^2$  reveals that the GARCH(p,q) process can be represented as a restricted infinite-ordered ARCH process,

$$\sigma_t^2 = \frac{\omega}{1 - \sum \beta_i} + \frac{\alpha(L)}{1 - \beta(L)} \varepsilon_t^2 = \frac{\omega}{1 - \sum \beta_i} + \sum_{i=1}^{\infty} \delta_i \varepsilon_{t-i}^2,$$

which precisely parallels writing an ARMA process as a restricted infinite-ordered AR. Hence the GARCH(p,q) process is a parsimonious approximation to what may truly be infinite-ordered ARCH volatility dynamics.

---

<sup>4</sup> By “pure” we mean that we have allowed only for conditional variance dynamics, by setting  $y_t = \varepsilon_t$ . We could of course also introduce conditional mean dynamics, but doing so would only clutter the discussion while adding nothing new.

It is important to note a number of special cases of the GARCH(p,q) process. First, of course, the ARCH(p) process emerges when

$$\beta(L) = 0.$$

Second, if *both*  $\alpha(L)$  and  $\beta(L)$  are zero, then the process is simply iid Gaussian noise with variance  $\omega$ . Hence, although ARCH and GARCH processes may at first appear unfamiliar and potentially ad hoc, they are in fact much more general than standard iid white noise, which emerges as a potentially highly-restrictive special case.

Here we highlight some important properties of GARCH processes. All of the discussion of course applies as well to ARCH processes, which are special cases of GARCH processes. First, consider the second-order moment structure of GARCH processes. The first two unconditional moments of the pure GARCH process are constant and given by

$$E(\varepsilon_t) = 0$$

and

$$E(\varepsilon_t - E(\varepsilon_t))^2 = \frac{\omega}{1 - \sum \alpha_i - \sum \beta_i},$$

while the conditional moments are

$$E(\varepsilon_t | \Omega_{t-1}) = 0$$

and of course

$$E((\varepsilon_t - E(\varepsilon_t | \Omega_{t-1}))^2 | \Omega_{t-1}) = \omega + \alpha(L)\varepsilon_t^2 + \beta(L)\sigma_t^2.$$

In particular, the unconditional variance is fixed, as must be the case under covariance stationarity, while the conditional variance is time-varying. It is no

*surprise* that the conditional variance is time-varying – the GARCH process was of course *designed* to allow for a time-varying conditional variance – but it is certainly worth emphasizing: the conditional variance is itself a serially correlated time series process.

Second, consider the unconditional higher-order (third and fourth) moment structure of GARCH processes. Real-world financial asset returns, which are often modeled as GARCH processes, are typically unconditionally symmetric but leptokurtic (that is, more peaked in the center and with fatter tails than a normal distribution). It turns out that the implied unconditional distribution of the conditionally Gaussian GARCH process introduced above is also symmetric and leptokurtic. The unconditional leptokurtosis of GARCH processes follows from the persistence in conditional variance, which produces clusters of “low volatility” and “high volatility” episodes associated with observations in the center and in the tails of the unconditional distribution, respectively. Both the unconditional symmetry and unconditional leptokurtosis agree nicely with a variety of financial market data.

Third, consider the conditional prediction error variance of a GARCH process, and its dependence on the conditioning information set. Because the conditional variance of a GARCH process is a serially correlated random variable, it is of interest to examine the optimal h-step-ahead prediction, prediction error, and conditional prediction error variance. Immediately, the h-step-ahead prediction is

$$E(\varepsilon_{t+h} \mid \Omega_t) = 0,$$

and the corresponding prediction error is

$$\varepsilon_{t+h} - E(\varepsilon_{t+h} \mid \Omega_t) = \varepsilon_{t+h}.$$

This implies that the conditional variance of the prediction error,

$$E((\varepsilon_{t+h} - E(\varepsilon_{t+h} | \Omega_t))^2 | \Omega_t) = E(\varepsilon_{t+h}^2 | \Omega_t),$$

depends on both  $h$  and

$$\Omega_t,$$

because of the dynamics in the conditional variance. Simple calculations reveal that the expression for the GARCH(p, q) process is given by

$$E(\varepsilon_{t+h}^2 | \Omega_t) = \omega \left( \sum_{i=0}^{h-2} (\alpha(1) + \beta(1))^i \right) + (\alpha(1) + \beta(1))^{h-1} \sigma_{t+1}^2.$$

In the limit, this conditional variance reduces to the unconditional variance of the process,

$$\lim_{h \rightarrow \infty} E(\varepsilon_{t+h}^2 | \Omega_t) = \frac{\omega}{1 - \alpha(1) - \beta(1)}.$$

For finite  $h$ , the dependence of the prediction error variance on the current information set  $\Omega_t$  can be exploited to improve interval and density forecasts.

Fourth, consider the relationship between  $\varepsilon_t^2$  and  $\sigma_t^2$ . The relationship is important: GARCH dynamics in  $\sigma_t^2$  turn out to introduce ARMA dynamics in  $\varepsilon_t^2$ .<sup>5</sup> More precisely, if  $\varepsilon_t$  is a GARCH(p,q) process, then

$$\varepsilon_t^2$$

has the ARMA representation

$$\varepsilon_t^2 = \omega + (\alpha(L) + \beta(L))\varepsilon_t^2 - \beta(L)\nu_t + \nu_t,$$

where

$$\nu_t = \varepsilon_t^2 - \sigma_t^2$$

---

<sup>5</sup> Put differently, the GARCH process approximates conditional variance dynamics in the same way that an ARMA process approximates conditional mean dynamics.

is the difference between the squared innovation and the conditional variance at time  $t$ . To see this, note that if  $\varepsilon_t$  is GARCH(p,q), then

$$\sigma_t^2 = \omega + \alpha(L)\varepsilon_t^2 + \beta(L)\sigma_t^2.$$

Adding and subtracting

$$\beta(L)\varepsilon_t^2$$

from the right side gives

$$\begin{aligned}\sigma_t^2 &= \omega + \alpha(L)\varepsilon_t^2 + \beta(L)\varepsilon_t^2 - \beta(L)\varepsilon_t^2 + \beta(L)\sigma_t^2 \\ &= \omega + (\alpha(L) + \beta(L))\varepsilon_t^2 - \beta(L)(\varepsilon_t^2 - \sigma_t^2).\end{aligned}$$

Adding

$$\varepsilon_t^2$$

to each side then gives

$$\sigma_t^2 + \varepsilon_t^2 = \omega + (\alpha(L) + \beta(L))\varepsilon_t^2 - \beta(L)(\varepsilon_t^2 - \sigma_t^2) + \varepsilon_t^2,$$

so that

$$\begin{aligned}\varepsilon_t^2 &= \omega + (\alpha(L) + \beta(L))\varepsilon_t^2 - \beta(L)(\varepsilon_t^2 - \sigma_t^2) + (\varepsilon_t^2 - \sigma_t^2), \\ &= \omega + (\alpha(L) + \beta(L))\varepsilon_t^2 - \beta(L)\nu_t + \nu_t.\end{aligned}$$

Thus,

$$\varepsilon_t^2$$

is an ARMA( $(\max(p,q), p)$ ) process with innovation  $\nu_t$ , where

$$\nu_t \in [-\sigma_t^2, \infty).$$

$\varepsilon_t^2$  is covariance stationary if the roots of  $\alpha(L)+\beta(L)=1$  are outside the

unit circle.

Fifth, consider in greater depth the similarities and differences between  $\sigma_t^2$  and

$$\varepsilon_t^2.$$

It is worth studying closely the key expression,

$$\nu_t = \varepsilon_t^2 - \sigma_t^2,$$

which makes clear that

$$\varepsilon_t^2$$

is effectively a “proxy” for  $\sigma_t^2$ , behaving similarly but not identically, with  $\nu_t$  being the difference, or error. In particular,  $\varepsilon_t^2$  is a *noisy* proxy:  $\varepsilon_t^2$  is an unbiased estimator of  $\sigma_t^2$ , but it is more volatile. It seems reasonable, then, that reconciling the noisy proxy  $\varepsilon_t^2$  and the true underlying  $\sigma_t^2$  should involve some sort of smoothing of  $\varepsilon_t^2$ . Indeed, in the GARCH(1,1) case  $\sigma_t^2$  is precisely obtained by exponentially smoothing  $\varepsilon_t^2$ . To see why, consider the exponential smoothing recursion, which gives the current smoothed value as a convex combination of the current unsmoothed value and the lagged smoothed value,

$$\bar{\varepsilon}_t^2 = \gamma \varepsilon_t^2 + (1 - \gamma) \bar{\varepsilon}_{t-1}^2.$$

Back substitution yields an expression for the current smoothed value as an exponentially weighted moving average of past actual values:

$$\bar{\varepsilon}_t^2 = \sum w_j \varepsilon_{t-j}^2,$$

where

$$w_j = \gamma(1 - \gamma)^j.$$

Now compare this result to the GARCH(1,1) model, which gives the current volatility as a linear combination of lagged volatility and the lagged

squared return,  $\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$ .

Back substitution yields  $\sigma_t^2 = \frac{\omega}{1-\beta} + \alpha \sum \beta^{j-1} \varepsilon_{t-j}^2$ , so that the GARCH(1,1) process gives current volatility as an exponentially weighted moving average of past squared returns.

Sixth, consider the temporal aggregation of GARCH processes. By temporal aggregation we mean aggregation over time, as for example when we convert a series of daily returns to weekly returns, and then to monthly returns, then quarterly, and so on. It turns out that convergence toward normality under temporal aggregation is a feature of real-world financial asset returns. That is, although high-frequency (e.g., daily) returns tend to be fat-tailed relative to the normal, the fat tails tend to get thinner under temporal aggregation, and normality is approached. Convergence to normality under temporal aggregation is also a property of covariance stationary GARCH processes. The key insight is that a low-frequency change is simply the sum of the corresponding high-frequency changes; for example, an annual change is the sum of the internal quarterly changes, each of which is the sum of its internal monthly changes, and so on. Thus, if a Gaussian central limit theorem can be invoked for sums of GARCH processes, convergence to normality under temporal aggregation is assured. Such theorems can be invoked if the process is covariance stationary.

In closing this section, it is worth noting that the symmetry and leptokurtosis of the unconditional distribution of the GARCH process, as well as the disappearance of the leptokurtosis under temporal aggregation, provide nice independent confirmation of the accuracy of GARCH approximations to asset return volatility dynamics, insofar as GARCH was certainly not invented with the intent of explaining those features of financial asset return data. On the contrary, the unconditional distributional results emerged as unanticipated byproducts of allowing for conditional variance dynamics, thereby providing a unified explanation of phenomena that were previously believed

unrelated.

## 8.3 Extensions of ARCH and GARCH Models

There are numerous extensions of the basic GARCH model. In this section, we highlight several of the most important. One important class of extensions allows for **asymmetric response**; that is, it allows for last period's squared return to have different effects on today's volatility, depending on its sign.<sup>6</sup> Asymmetric response is often present, for example, in stock returns.

### 8.3.1 Asymmetric Response

The simplest GARCH model allowing for asymmetric response is the **threshold GARCH**, or TGARCH, model.<sup>7</sup> We replace the standard GARCH conditional variance function,  $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$ , with  $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2 D_t$ , where  $D_t = \begin{cases} 1, & \text{if } \varepsilon_t < 0 \\ 0, & \text{otherwise.} \end{cases}$

The dummy variable  $D$  keeps track of whether the lagged return is positive or negative. When the lagged return is positive (good news yesterday),  $D=0$ , so the effect of the lagged squared return on the current conditional variance is simply  $\alpha$ . In contrast, when the lagged return is negative (bad news yesterday),  $D=1$ , so the effect of the lagged squared return on the current conditional variance is  $\alpha+\gamma$ . If  $\gamma = 0$ , the response is symmetric and we have a standard GARCH model, but if  $\gamma \neq 0$  we have asymmetric response of volatility to news. Allowance for asymmetric response has proved useful for modeling “leverage effects” in stock returns, which occur when  $\gamma < 0$ .<sup>8</sup>

---

<sup>6</sup> In the GARCH model studied thus far, only the *square* of last period's return affects the current conditional variance; hence its sign is irrelevant.

<sup>7</sup> For expositional convenience, we will introduce all GARCH extensions in the context of GARCH(1,1), which is by far the most important case for practical applications. Extensions to the GARCH(p,q) case are immediate but notationally cumbersome.

<sup>8</sup> Negative shocks appear to contribute more to stock market volatility than do positive shocks. This is called the leverage effect, because a negative shock to the market value of equity increases the aggregate debt/equity ratio (other things the same), thereby increasing leverage.

Asymmetric response may also be introduced via the **exponential GARCH** (EGARCH) model,

$$\ln(\sigma_t^2) = \omega + \alpha \left| \frac{\varepsilon_{t-1}}{\sigma_{t-1}} \right| + \gamma \frac{\varepsilon_{t-1}}{\sigma_{t-1}} + \beta \ln(\sigma_{t-1}^2).$$

Note that volatility is driven by both size and sign of shocks; hence the model allows for an asymmetric response depending on the sign of news.<sup>9</sup> The log specification also ensures that the conditional variance is automatically positive, because  $\sigma_t^2$  is obtained by exponentiating  $\ln(\sigma_t^2)$ ; hence the name “exponential GARCH.”

### 8.3.2 Exogenous Variables in the Volatility Function

Just as ARMA models of conditional mean dynamics can be augmented to include the effects of exogenous variables, so too can GARCH models of conditional variance dynamics.

We simply modify the standard GARCH volatility function in the obvious way, writing

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2 + \gamma x_t,$$

where  $\gamma$  is a parameter and  $x$  is a positive exogenous variable.<sup>10</sup> Allowance for exogenous variables in the conditional variance function is sometimes useful. Financial market volume, for example, often helps to explain market volatility.

---

<sup>9</sup> The absolute “size” of news is captured by  $|r_{t-1}/\sigma_{t-1}|$ , and the sign is captured by  $r_{t-1}/\sigma_{t-1}$ .

<sup>10</sup> Extension to allow multiple exogenous variables is straightforward.

### 8.3.3 Regression with GARCH disturbances and GARCH-M

Just as ARMA models may be viewed as models for disturbances in regressions, so too may GARCH models. We write

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

$$\varepsilon_t | \Omega_{t-1} \sim N(0, \sigma_t^2)$$

$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$ . Consider now a regression model with GARCH disturbances of the usual sort, with one additional twist: the conditional variance enters as a regressor, thereby affecting the conditional mean. We write

$$y_t = \beta_0 + \beta_1 x_t + \gamma \sigma_t^2 + \varepsilon_t$$

$$\varepsilon_t | \Omega_{t-1} \sim N(0, \sigma_t^2)$$

$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$ . This model, which is a special case of the general regression model with GARCH disturbances, is called GARCH-in-Mean (GARCH-M). It is sometimes useful in modeling the relationship between risks and returns on financial assets when risk, as measured by the conditional variance, varies.<sup>11</sup>

### 8.3.4 Component GARCH

Note that the standard GARCH(1,1) process may be written as  $(\sigma_t^2 - \bar{\omega}) = \alpha(\varepsilon_{t-1}^2 - \bar{\omega})$  where  $\bar{\omega} = \frac{\omega}{1-\alpha-\beta}$  is the unconditional variance.<sup>12</sup> This is precisely the GARCH(1,1) model introduced earlier, rewritten it in a slightly different but equivalent form. In this model, short-run volatility dynamics are governed by the parameters  $\alpha$  and  $\beta$ , and there are no long-run volatility dynamics, because  $\bar{\omega}$  is constant. Sometimes we might want to allow for both long-run and

---

<sup>11</sup> One may also allow the conditional standard deviation, rather than the conditional variance, to enter the regression.

<sup>12</sup>  $\bar{\omega}$  is sometimes called the “long-run” variance, referring to the fact that the unconditional variance is the long-run average of the conditional variance.

short-run, or persistent and transient, volatility dynamics in addition to the short-run volatility dynamics already incorporated. To do this, we replace  $\bar{\omega}$  with a time-varying process, yielding  $(\sigma_t^2 - q_t) = \alpha(\varepsilon_{t-1}^2 - q_{t-1}) + \beta(\sigma_{t-1}^2 - q_{t-1})$ , where the time-varying long-run volatility,  $q_t$ , is given by  $q_t = \omega + \rho(q_{t-1} - \omega) + \phi(\varepsilon_t^2)$ . This “component GARCH” model effectively lets us decompose volatility dynamics into long-run (persistent) and short-run (transitory) components, which sometimes yields useful insights. The persistent dynamics are governed by  $\rho$ , and the transitory dynamics are governed by  $\alpha$  and  $\beta$ .<sup>13</sup>

### 8.3.5 Mixing and Matching

In closing this section, we note that the different variations and extensions of the GARCH process may of course be mixed. As an example, consider the following conditional variance function:  $(\sigma_t^2 - q_t) = \alpha(\varepsilon_{t-1}^2 - q_{t-1}) + \gamma(\varepsilon_{t-1}^2 - q_{t-1})D_{t-1}$ . This is a component GARCH specification, generalized to allow for asymmetric response of volatility to news via the sign dummy  $D$ , as well as effects from the exogenous variable  $x$ .

## 8.4 Estimating, Forecasting and Diagnosing GARCH Models

Recall that the likelihood function is the joint density function of the data, viewed as a function of the model parameters, and that maximum likelihood estimation finds the parameter values that maximize the likelihood function. This makes good sense: we choose those parameter values that maximize the likelihood of obtaining the data that were actually obtained. It turns

---

<sup>13</sup> It turns out, moreover, that under suitable conditions the component GARCH model introduced here is covariance stationary, and equivalent to a GARCH(2,2) process subject to certain nonlinear restrictions on its parameters.

out that construction and evaluation of the likelihood function is easily done for GARCH models, and maximum likelihood has emerged as the estimation method of choice.<sup>14</sup> No closed-form expression exists for the GARCH maximum likelihood estimator, so we must maximize the likelihood numerically.<sup>15</sup> Construction of optimal forecasts of GARCH processes is simple. In fact, we derived the key formula earlier but did not comment extensively on it. Recall, in particular, that

$$\sigma_{t+h,t}^2 = E[\varepsilon_{t+h}^2 | \Omega_t] = \omega \left( \sum_{i=1}^{h-1} [\alpha(1) + \beta(1)]^i \right) + [\alpha(1) + \beta(1)]^{h-1} \sigma_{t+1}^2.$$

In words, the optimal h-step-ahead forecast is proportional to the optimal 1-step-ahead forecast. The optimal 1-step-ahead forecast, moreover, is easily calculated: all of the determinants of  $\sigma_{t+1}^2$  are lagged by at least one period, so that there is no problem of forecasting the right-hand side variables. In practice, of course, the underlying GARCH parameters  $\alpha$  and  $\beta$  are unknown and so must be estimated, resulting in the feasible forecast  $\hat{\sigma}_{t+h,t}^2$  formed in the obvious way. In financial applications, volatility forecasts are often of direct interest, and the GARCH model delivers the optimal h-step-ahead point forecast,  $\hat{\sigma}_{t+h,t}^2$ . Alternatively, and more generally, we might not be intrinsically interested in volatility; rather, we may simply want to use GARCH volatility forecasts to improve h-step-ahead interval or density forecasts of  $\varepsilon_t$ , which are crucially dependent on the h-step-ahead prediction error variance,  $\sigma_{t+h,t}^2$ . Consider, for example, the case of interval forecasting. In the case of constant volatility, we earlier worked with Gaussian ninety-five percent

---

<sup>14</sup> The precise form of the likelihood is complicated, and we will not give an explicit expression here, but it may be found in various of the surveys mentioned in the Bibliographical and Computational Notes at the end of the chapter.

<sup>15</sup> Routines for maximizing the GARCH likelihood are available in a number of modern software packages such as Eviews. As with any numerical optimization, care must be taken with startup values and convergence criteria to help insure convergence to a global, as opposed to merely local, maximum.

interval forecasts of the form

$$y_{t+h,t} \pm 1.96\sigma_h ,$$

where  $\sigma_h$  denotes the unconditional h-step-ahead standard deviation (which also equals the conditional h-step-ahead standard deviation in the absence of volatility dynamics). Now, however, in the presence of volatility dynamics we use

$$y_{t+h,t} \pm 1.96\sigma_{t+h,t} .$$

The ability of the conditional prediction interval to adapt to changes in volatility is natural and desirable: when volatility is low, the intervals are naturally tighter, and conversely. In the presence of volatility dynamics, the unconditional interval forecast is correct on average but likely incorrect at any given time, whereas the conditional interval forecast is correct at all times. The issue arises as to how to detect GARCH effects in observed returns, and related, how to assess the adequacy of a fitted GARCH model. A key and simple device is the correlogram of squared returns,  $\varepsilon_t^2$ . As discussed earlier,  $\varepsilon_t^2$  is a proxy for the latent conditional variance; if the conditional variance displays persistence, so too will  $\varepsilon_t^2$ .<sup>16</sup> Once can of course also fit a GARCH model, and assess significance of the GARCH coefficients in the usual way.

Note that we can write the GARCH process for returns as  $\varepsilon_t = \sigma_t v_t$ , *iid*

where  $v_t \sim N(0, 1)$   $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$ .

Equivalently, the *standardized* return,  $v$ , is iid,

---

<sup>16</sup> Note well, however, that the converse is not true. That is, if  $\varepsilon_t^2$  displays persistence, it does not necessarily follow that the conditional variance displays persistence. In particular, neglected serial correlation associated with conditional mean dynamics may cause serial correlation in  $\varepsilon_t$  and hence also in  $\varepsilon_t^2$ . Thus, before proceeding to examine and interpret the correlogram of  $\varepsilon_t^2$  as a check for volatility dynamics, it is important that any conditional mean effects be appropriately modeled, in which case  $\varepsilon_t$  should be interpreted as the disturbance in an appropriate conditional mean model.

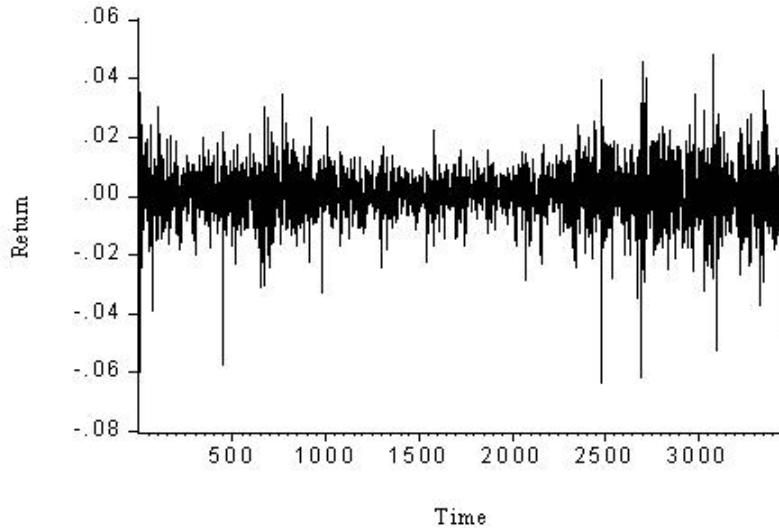


Figure 8.1: NYSE Returns

*iid*

$$\varepsilon_{\frac{t}{\sigma_t}} = v_t \sim N(0, 1).$$

This observation suggests a way to evaluate the adequacy of a fitted GARCH model: standardize returns by the conditional standard deviation from the fitted GARCH model,  $\hat{\sigma}$ , and then check for volatility dynamics missed by the fitted model by examining the correlogram of the squared *standardized* return,  $(\varepsilon_t/\hat{\sigma}_t)^2$ . This is routinely done in practice.

## 8.5 Application: Stock Market Volatility

We model and forecast the volatility of daily returns on the New York Stock Exchange (NYSE) from January 1, 1988 through December 31, 2001, excluding holidays, for a total of 3531 observations. We estimate using observations 1-3461, and then we forecast observations 3462-3531.

In Figure 8.1 we plot the daily returns,  $r_t$ . There is no visual evidence of serial correlation in the returns, but there *is* evidence of serial correlation in the *amplitude* of the returns. That is, volatility appears to cluster: large

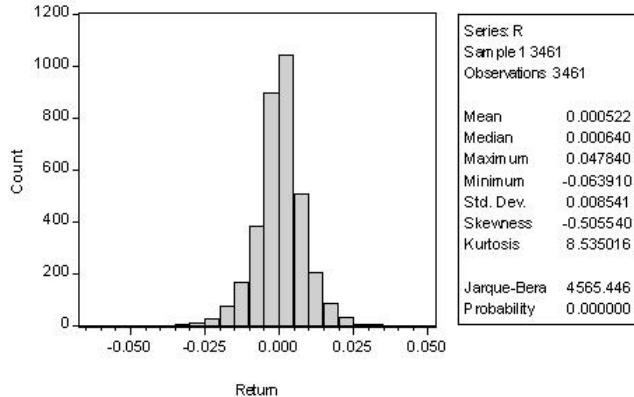


Figure 8.2: Histogram of NYSE Returns

changes tend to be followed by large changes, and small by small, *of either sign*. In Figure 8.2 we show the histogram and related statistics for  $r_t$ . The mean daily return is slightly positive. Moreover, the returns are approximately symmetric (only slightly left skewed) but highly leptokurtic. The Jarque-Bera statistic indicates decisive rejection of normality.

In Figure 8.3 we show the correlogram for  $r_t$ . The sample autocorrelations are tiny and usually insignificant relative to the Bartlett standard errors, yet the autocorrelation function shows some evidence of a systematic cyclical pattern, and the Q statistics (not shown), which cumulate the information across all displacements, reject the null of weak white noise. Despite the weak serial correlation evidently present in the returns, we will proceed for now as if returns were weak white noise, which is approximately, if not exactly, the case.<sup>17</sup>

In Figure 8.4 we plot  $r_t^2$ . The volatility clustering is even more evident than it was in the time series plot of returns. Perhaps the strongest evidence of all comes from the correlogram of  $r_t^2$ , which we show in Figure 8.5: all sample autocorrelations of  $r_t^2$  are positive, overwhelmingly larger than those of the returns themselves, and statistically significant. As a crude first pass at modeling the stock market volatility, we fit an AR(5) model directly to  $r_t^2$

<sup>17</sup> In the Exercises, Problems and Complements at the end of this chapter we model the conditional mean, as well as the conditional variance, of returns.

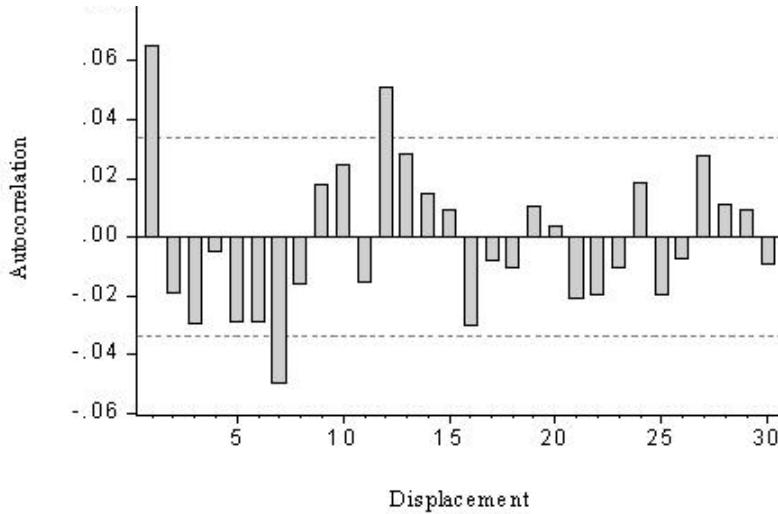


Figure 8.3: Correlogram of NYSE Returns

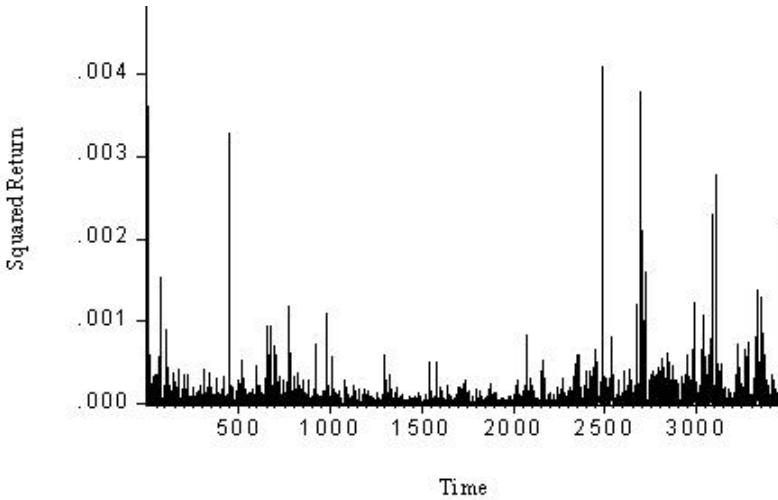


Figure 8.4: Squared NYSE Returns

; the results appear in Table 8.6. It is interesting to note that the t-statistics on the lagged squared returns are often significant, even at long lags, yet the  $R^2$  of the regression is low, reflecting the fact that  $r_t^2$  is a very noisy volatility proxy. As a more sophisticated second pass at modeling NYSE volatility, we fit an ARCH(5) model to  $r_t$  ; the results appear in Table 8.7. The lagged squared returns appear significant even at long lags. The correlogram of squared standardized residuals shown in Figure 8.8, however, displays some remaining systematic behavior, indicating that the ARCH(5) model fails to

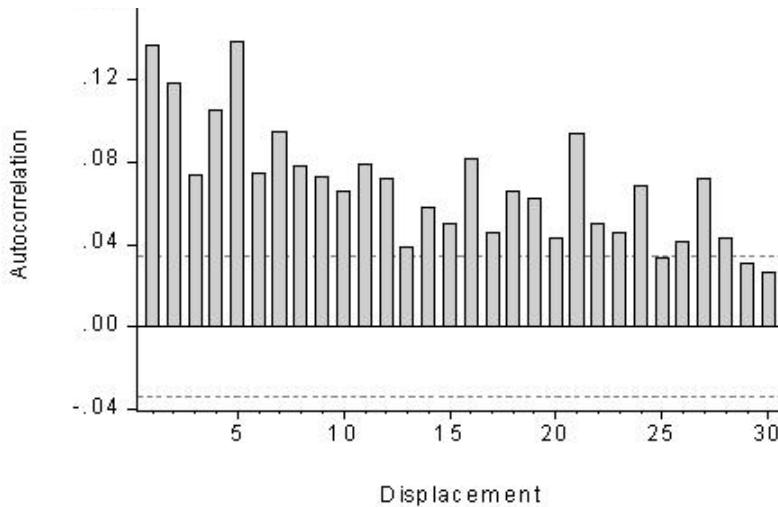


Figure 8.5: Squared NYSE Returns Correlogram

capture all of the volatility dynamics, potentially because even longer lags are needed.<sup>18</sup>

<sup>18</sup> In the Exercises, Problems and Complements at the end of this chapter we also examine ARCH(p) models with  $p > 5$ .

Dependent Variable: R2				
Method: Least Squares				
Sample(adjusted): 6 3461				
Included observations: 3456 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.40E-05	3.78E-06	11.62473	0.0000
R2(-1)	0.107900	0.016137	6.686547	0.0000
R2(-2)	0.091840	0.016186	5.674167	0.0000
R2(-3)	0.028981	0.016250	1.783389	0.0746
R2(-4)	0.039312	0.016481	2.385241	0.0171
R2(-5)	0.116436	0.016338	7.126828	0.0000
R-squared	0.052268	Mean dependent var		7.19E-05
Adjusted R-squared	0.050894	S.D. dependent var		0.000189
S.E. of regression	0.000184	Akaike info criterion		-14.36434
Sum squared resid	0.000116	Schwarz criterion		-14.35366
Log likelihood	24827.58	F-statistic		38.05372
Durbin-Watson stat	1.975672	Prob(F-statistic)		0.000000

Figure 8.6: Squared NYSE Returns, AR(5) Model

Dependent Variable: R				
Method: ML - ARCH (Marquardt)				
Sample: 1 3461				
Included observations: 3461				
Convergence achieved after 13 iterations				
Variance backcast: ON				
Coefficient	Std. Error	z-Statistic	Prob.	
C	0.000689	0.000127	5.437097	0.0000
Variance Equation				
C	3.16E-05	1.08E-06	29.28536	0.0000
ARCH(1)	0.128948	0.013847	9.312344	0.0000
ARCH(2)	0.166852	0.015055	11.08281	0.0000
ARCH(3)	0.072551	0.014345	5.057526	0.0000
ARCH(4)	0.143778	0.015363	9.358870	0.0000
ARCH(5)	0.089254	0.018480	4.829789	0.0000
R-squared	-0.000381	Mean dependent var		0.000522
Adjusted R-squared	-0.002118	S.D. dependent var		0.008541
S.E. of regression	0.008550	Akaike info criterion		-6.821461
Sum squared resid	0.252519	Schwarz criterion		-6.809024
Log likelihood	11811.54	Durbin-Watson stat		1.861036

Figure 8.7: Squared NYSE Returns, ARCH(5) Model

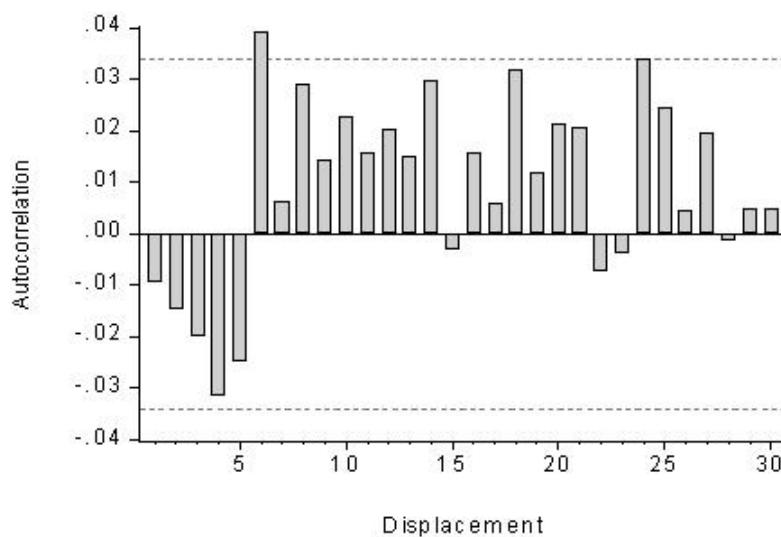


Figure 8.8: NYSE Returns, Correlogram of Squared Standardized Residuals from  $\text{ARCH}(5)$

In Table 8.9 we show the results of fitting a GARCH(1,1) model. All of the parameter estimates are highly statistically significant, and the “ARCH coefficient” ( $\alpha$ ) and “GARCH coefficient” ( $\beta$ ) sum to a value near unity (.987), with  $\beta$  substantially larger than  $\alpha$ , as is commonly found for financial asset returns. We show the correlogram of squared standardized GARCH(1,1) residuals in Figure 8.10. All sample autocorrelations are tiny and inside the Bartlett bands, and they display noticeably less evidence of any systematic pattern than for the squared standardized ARCH(5) residuals. In Figure 8.11 we show the time series of estimated conditional standard deviations implied by the estimated GARCH(1,1) model. Clearly, volatility fluctuates a great deal and is highly persistent. For comparison we show in Figure 8.12 the series of exponentially smoothed  $r_t^2$ , computed using a standard smoothing parameter of .05.<sup>19</sup> Clearly the GARCH and exponential smoothing volatility estimates behave similarly, although not at all identically. The difference reflects the fact that the GARCH smoothing parameter is effectively estimated by the method of maximum likelihood, whereas the exponential smoothing parameter is set rather arbitrarily. Now, using the model estimated using observations 1-3461, we generate a forecast of the conditional standard deviation for the out-of-sample observations 3462-3531. We show the results in Figure 8.13. The forecast period begins just following a volatility burst, so it is not surprising that the forecast calls for gradual volatility reduction. For greater understanding, in Figure 8.14 we show both a longer history and a longer forecast. Clearly the forecast conditional standard deviation is reverting exponentially to the unconditional standard deviation (.009), per the formula discussed earlier.

---

<sup>19</sup> For comparability with the earlier-computed GARCH estimated conditional standard deviation, we actually show the square root of exponentially smoothed  $r_t^2$ .

Dependent Variable: R				
Method: ML - ARCH (Marquardt)				
Sample: 1 3461				
Included observations: 3461				
Convergence achieved after 19 iterations				
Variance backcast: ON				
Coefficient	Std. Error	z-Statistic	Prob.	
C	0.000640	0.000127	5.036942	0.0000
Variance Equation				
C	1.06E-06	1.49E-07	7.136840	0.0000
ARCH(1)	0.067410	0.004955	13.60315	0.0000
GARCH(1)	0.919714	0.006122	150.2195	0.0000
R-squared	-0.000191	Mean dependent var		0.000522
Adjusted R-squared	-0.001059	S.D. dependent var		0.008541
S.E. of regression	0.008546	Akaike info criterion		-6.868008
Sum squared resid	0.252471	Schwarz criterion		-6.860901
Log likelihood	11889.09	Durbin-Watson stat		1.861389

Figure 8.9: Squared NYSE Returns, GARCH(1,1)

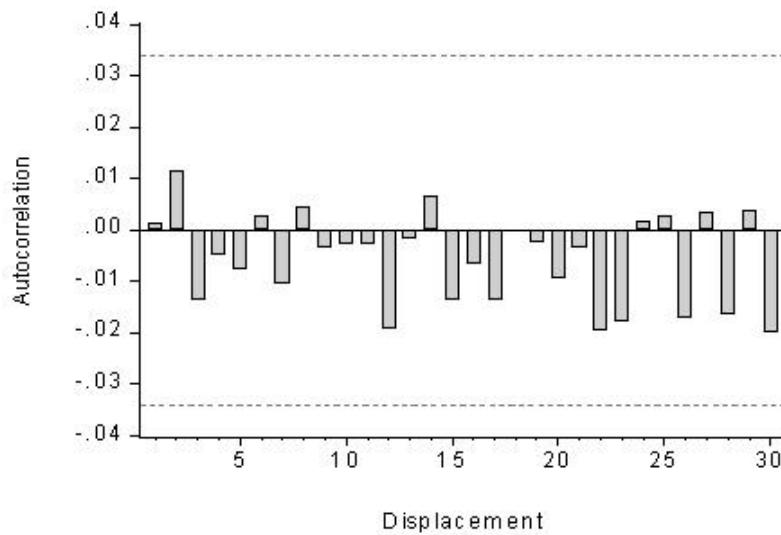


Figure 8.10: NYSE Returns, Correlogram of Squared Standardized Residuals from GARCH(1,1)

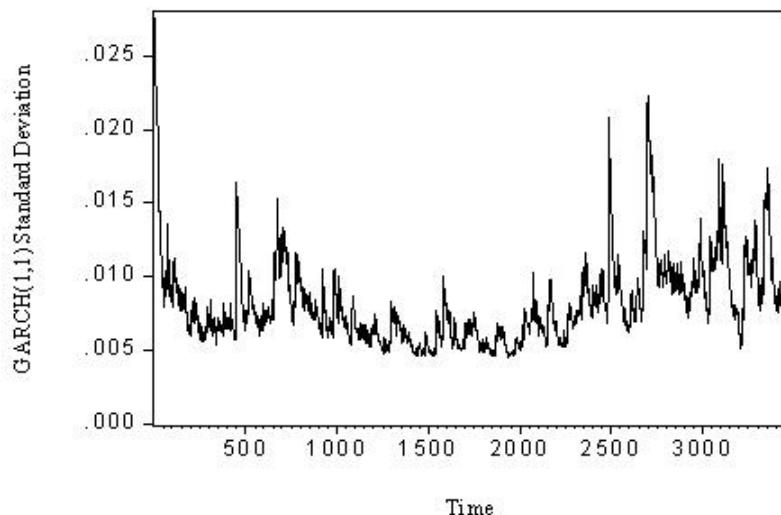


Figure 8.11: Estimated Conditional Standard Deviations from GARCH(1,1)

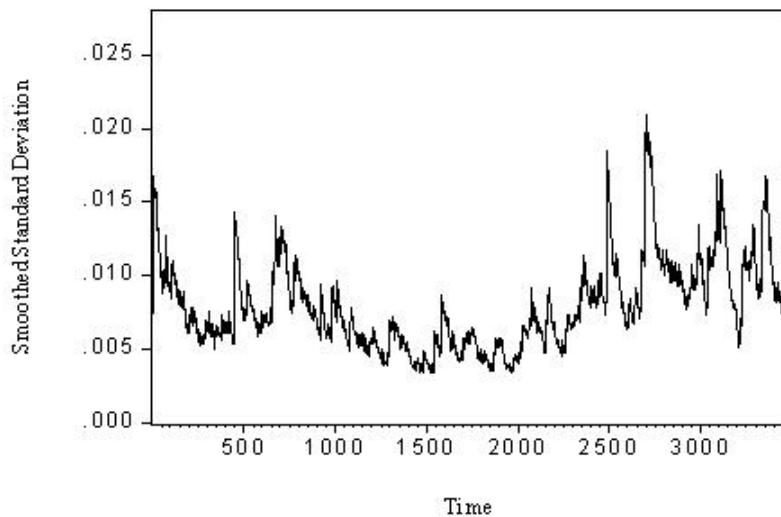


Figure 8.12: Estimated Conditional Standard Deviations - Exponential Smoothing

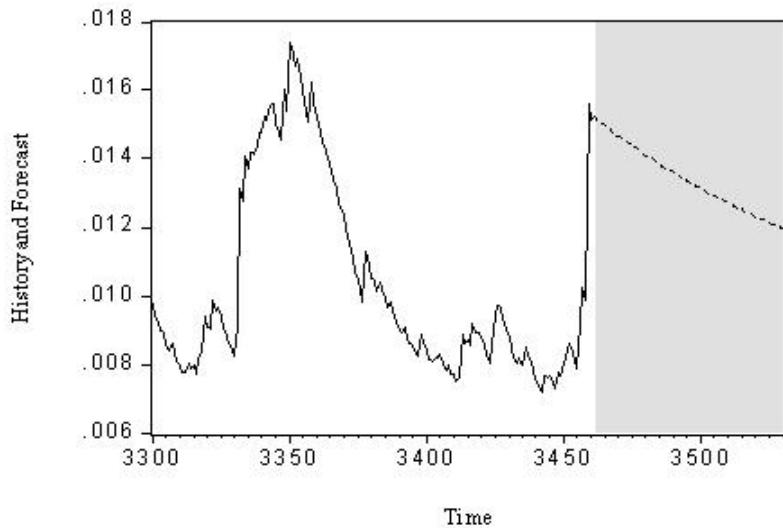


Figure 8.13: Conditional Standard Deviations, History and Forecast from GARCH(1,1)

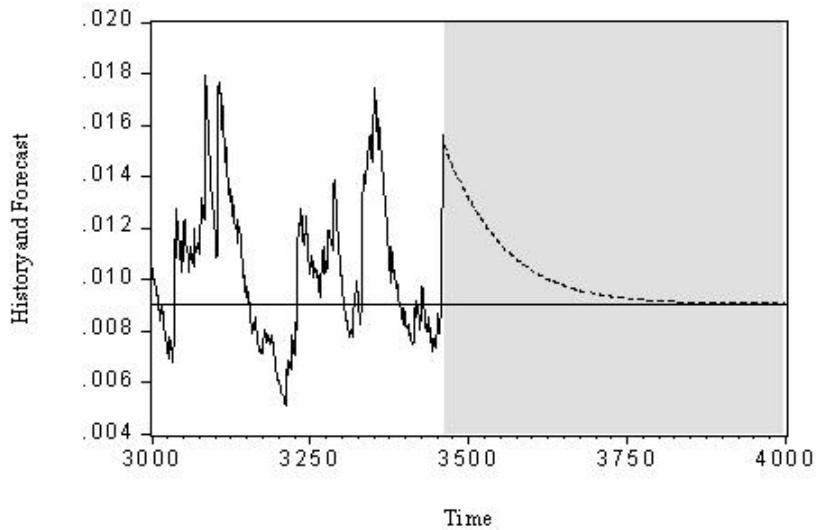


Figure 8.14: Conditional Standard Deviation, History and Extended Forecast from GARCH(1,1)

## 8.6 Exercises, Problems and Complements

1. Volatility dynamics: correlograms of squares.

In the Chapter 3 EPC, we suggested that a time series plot of a squared residual,  $e_t^2$ , can reveal serial correlation in squared residuals, which corresponds to non-constant volatility, or heteroskedasticity, in the levels of the residuals. Financial asset returns often display little systematic variation, so instead of examining residuals from a model of returns, we often examine returns directly. In what follows, we will continue to use the notation  $e_t$ , but you should interpret  $e_t$  it as an observed asset return.

- a. Find a high frequency (e.g., daily) financial asset return series,  $e_t$ , plot it, and discuss your results.
  - b. Perform a correlogram analysis of  $e_t$ , and discuss your results.
  - c. Plot  $e_t^2$ , and discuss your results.
  - d. In addition to plotting  $e_t^2$ , examining the correlogram of  $e_t^2$  often proves informative for assessing volatility persistence. Why might that be so? Perform a correlogram analysis of  $e_t^2$  and discuss your results.
2. Removing conditional mean dynamics before modeling volatility dynamics. In the application in the text we noted that NYSE stock returns appeared to have some weak conditional mean dynamics, yet we ignored them and proceeded directly to model volatility.
    - a. Instead, first fit autoregressive models using the SIC to guide order selection, and then fit GARCH models to the residuals. Redo the entire empirical analysis reported in the text in this way, and discuss any important differences in the results.

- b. Consider instead the simultaneous estimation of all parameters of AR(p)-GARCH models. That is, estimate regression models where the regressors are lagged dependent variables and the disturbances display GARCH. Redo the entire empirical analysis reported in the text in this way, and discuss any important differences in the results relative to those in the text and those obtained in part a above.
3. Variations on the basic ARCH and GARCH models.
- Using the stock return data, consider richer models than the pure ARCH and GARCH models discussed in the text.
- Estimate, diagnose and discuss a threshold GARCH(1,1) model.
  - Estimate, diagnose and discuss an EGARCH(1,1) model.
  - Estimate, diagnose and discuss a component GARCH(1,1) model.
  - Estimate, diagnose and discuss a GARCH-M model.
4. Empirical performance of pure ARCH models as approximations to volatility dynamics.
- Here we will fit pure ARCH( $p$ ) models to the stock return data, including values of  $p$  larger than  $p=5$  as done in the text, and contrast the results with those from fitting GARCH( $p,q$ ) models.
- When fitting pure ARCH( $p$ ) models, what value of  $p$  seems adequate?
  - When fitting GARCH( $p,q$ ) models, what values of  $p$  and  $q$  seem adequate?
  - Which approach appears more parsimonious?
5. Direct modeling of volatility proxies.

In the text we fit an AR(5) directly to a subset of the squared NYSE stock returns. In this exercise, use the *entire* NYSE dataset.

- a. Construct, display and discuss the fitted volatility series from the AR(5) model.
  - b. Construct, display and discuss an alternative fitted volatility series obtained by exponential smoothing, using a smoothing parameter of .10, corresponding to a large amount of smoothing, but less than done in the text.
  - c. Construct, display and discuss the volatility series obtained by fitting an appropriate GARCH model.
  - d. Contrast the results of parts a, b and c above.
  - e. Why is fitting of a GARCH model preferable in principle to the AR(5) or exponential smoothing approaches?
6. GARCH volatility forecasting.

You work for Xanadu, a luxury resort in the tropics. The daily temperature in the region is beautiful year-round, with a mean around 76 (Fahrenheit!) and no conditional mean dynamics. Occasional pressure systems, however, can cause bursts of temperature volatility. Such volatility bursts generally don't last long enough to drive away guests, but the resort still loses revenue from fees on activities that are less popular when the weather isn't perfect. In the middle of such a period of high temperature volatility, your boss gets worried and asks you to make a forecast of volatility over the next ten days. After some experimentation, you find that daily temperature  $y_t$  follows  $y_t | \Omega_{t-1} \sim N(\mu, \sigma_t^2)$ , where  $\sigma_t^2$  follows a GARCH(1,1) process,  $\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$

.

- a. Estimation of your model using historical daily temperature data yields  $\hat{\mu} = 76$ ,  $\hat{\omega} = 3$ ,  $\hat{\alpha} = .6$ , and  $\hat{\beta} = 0$ . If yesterday's temperature was 92 degrees, generate point forecasts for each of the next ten days conditional variance.

- b. According to your volatility forecasts, how many days will it take until volatility drops enough such that there is at least a 90% probability that the temperature will be within 4 degrees of 76?
  - c. Your boss is impressed by your knowledge of forecasting, and asks you if your model can predict the next spell of bad weather. How would you answer him?
7. Assessing volatility dynamics in observed returns and in standardized returns.

In the text we sketched the use of correlograms of squared observed returns for the detection of GARCH, and squared standardized returns for diagnosing the adequacy of a fitted GARCH model. Examination of Ljung-Box statistics is an important part of a correlogram analysis. McLeod and Li (1983) show that the Ljung-Box statistics may be legitimately used on squared observed returns, in which case it will have the usual

$$\chi_m^2$$

distribution under the null hypothesis of independence. Bollerslev and Mikkelsen (1996) argue that one may also use the Ljung-Box statistic on the squared standardized returns, but that a better distributional approximation is obtained in that case by using a

$$\chi_{m-k}^2$$

distribution, where  $k$  is the number of estimated GARCH parameters, to account for degrees of freedom used in model fitting.

8. Allowing for leptokurtic conditional densities.

Thus far we have worked exclusively with conditionally Gaussian GARCH

*iid*

models, which correspond to  $\varepsilon_t = \sigma_t v_t$   $v_t \sim N(0, 1)$ , or equivalently, to normality of the standardized return,  $\varepsilon_t / \sigma_t$ .

- a. The conditional normality assumption may sometimes be violated. However, Bollerslev and Wooldridge (1992) show that GARCH parameters are consistently estimated by Gaussian maximum likelihood even when the normality assumption is incorrect. Sketch some intuition for this result.
- b. Fit an appropriate conditionally Gaussian GARCH model to the stock return data. How might you use the histogram of the standardized returns to assess the validity of the conditional normality assumption? Do so and discuss your results.
- c. Sometimes the conditionally Gaussian GARCH model does indeed fail to explain all of the leptokurtosis in returns; that is, especially with very high-frequency data, we sometimes find that the conditional density is leptokurtic. Fortunately, leptokurtic conditional densities are easily incorporated into the GARCH model. For example, in Bollerslev's (1987) conditionally **Student's-t GARCH model**, the conditional density is assumed to be Student's t, with the degrees-of-freedom  $d$  treated as another parameter to be estimated. More precisely, we write

$$v_t \stackrel{iid}{\sim} \frac{t_d}{\text{std}(t_d)}.$$

$$\varepsilon_t = \sigma_t v_t$$

What is the reason for dividing the Student's t variable,  $t_d$ , by its

standard deviation,  $\text{std}(t_d)$ ? How might such a model be estimated?

## 9. Optimal prediction under asymmetric loss.

In the text we stressed GARCH modeling for improved interval and density forecasting, implicitly working under a symmetric loss function. Less obvious but equally true is the fact that, under *asymmetric* loss, volatility dynamics can be exploited to produce improved *point* forecasts, as shown by Christoffersen and Diebold (1996, 1997). The optimal predictor under asymmetric loss is not the conditional mean, but rather the conditional mean shifted by a time-varying adjustment that depends on the conditional variance. The intuition for the bias in the optimal predictor is simple – when errors of one sign are more costly than errors of the other sign, it is desirable to bias the forecasts in such a way as to reduce the chance of making an error of the more damaging type. The optimal amount of bias depends on the conditional prediction error variance of the process because, as the conditional variance grows, so too does the optimal amount of bias needed to avoid large prediction errors of the more damaging type. .

## 10. Multivariate GARCH models.

In the multivariate case, such as when modeling a *set* of returns rather than a single return, we need to model not only conditional variances, but also conditional *covariances*.

- a. Is the GARCH conditional variance specification introduced earlier, say for the  $i - \text{th}$  return,  $\sigma_{it}^2 = \omega + \alpha\varepsilon_{i,t-1}^2 + \beta\sigma_{i,t-1}^2$ , still appealing in the multivariate case? Why or why not?
- b. Consider the following specification for the conditional covariance between  $i - \text{th}$  and  $j$ -th returns:  $\sigma_{ij,t} = \omega + \alpha\varepsilon_{i,t-1}\varepsilon_{j,t-1} + \beta\sigma_{ij,t-1}$ . Is it appealing? Why or why not?

- c. Consider a fully general multivariate volatility model, in which every conditional variance and covariance may depend on lags of every conditional variance and covariance, as well as lags of every squared return and cross product of returns. What are the strengths and weaknesses of such a model? Would it be useful for modeling, say, a set of five hundred returns? If not, how might you proceed?

## 8.7 Notes

This chapter draws upon the survey by Diebold and Lopez (1995), which may be consulted for additional details. Other broad surveys include Bollerslev, Chou and Kroner

(1992), Bollerslev, Engle and Nelson (1994), Taylor (2005) and Andersen et al. (2007). Engle (1982) is the original development of the ARCH model. Bollerslev (1986) provides the important GARCH extension, and Engle (1995) contains many others. Diebold (1988) shows convergence to normality under temporal aggregation. TGARCH traces to Glosten, Jagannathan and Runkle (1993), and EGARCH to Nelson (1991). Engle, Lilien and Robins (1987) introduce the GARCH-M model, and Engle and Lee (1999) introduce component GARCH. Recently, methods of volatility measurement, modeling and forecasting have been developed that exploit the increasing availability of high-frequency financial asset return data. For a fine overview, see Dacorogna et al. (2001), and for more recent developments see Andersen, Bollerslev, Diebold and Labys (2003) and Andersen, Bollerslev and Diebold (2006). For insights into the emerging field of financial econometrics, see Diebold (2001) and many of the other papers in the same collection.



# Chapter 9

## Assembling the Components: U.S. Liquor Sales

Thus far we've focused on modeling trend, seasonals, and cycles one at a time. In Chapter 5, we introduced models and forecasts of trends and seasonality, respectively. Although cycles were likely present in the retail sales and housing starts series that we examined empirically, we simply ignored them. In Chapters 6 and 7 we introduced models and forecasts of cycles. We forecasted employment using autoregressive models. We didn't need trends or seasonals, because our employment series had no trend or seasonality.

In many forecasting situations, however, more than one component is needed to capture the dynamics in a series to be forecast – frequently they're *all* needed. Here we assemble our tools for forecasting trends, seasonals, and cycles; we use regression on a trend and calendar-effect dummies, and we capture cyclical dynamics by allowing for autoregressive effects in the regression disturbances, or by directly including lagged dependent variables in the regression.

## 9.1 Serially Correlated Disturbances

The full model is:

$$y_t = T_t(\theta) + \sum_{i=1}^s \gamma_i D_{it} + \varepsilon_t$$

$$\Phi(L)\varepsilon_t = v_t$$

$$\Phi(L) = 1 - \phi_1 L - \dots - \phi_p L^p$$

$$v_t \sim WN(0, \sigma^2).$$

$T_t(\theta)$  is a trend, with underlying parameters  $\theta$ . For example, linear trend has  $\theta = \beta_1$  and

$$T_t(\theta) = \beta_1 TIME_t,$$

and quadratic trend has  $\theta = (\beta_1, \beta_2)$  and

$$T_t(\theta) = \beta_1 TIME_t + \beta_2 TIME_t^2.$$

In addition to the trend, we include seasonal dummies.<sup>1,2</sup> The disturbances follow an  $AR(p)$  process. In any particular application, of course, various trend effects, seasonal and other calendar effects, and autoregressive cyclical effects may not be needed and so could be dropped.<sup>3</sup> Finally,  $v_t$  is the underlying white noise shock that drives everything.

Now consider constructing an  $h$ -step-ahead point forecast at time  $T$ ,  $y_{T+h,T}$ . At time  $T + h$ ,

$$y_{T+h} = T_{T+h}(\theta) + \sum_{i=1}^s \gamma_i D_{i,T+h} + \varepsilon_{T+h}.$$

Projecting the right-hand side variables on what's known at time  $T$  (that is,

<sup>1</sup>Note that, because we include a full set of seasonal dummies, the trend does not contain an intercept, and we don't include an intercept in the regression.

<sup>2</sup>Holiday and trading-day dummies could be included if relevant.

<sup>3</sup>If the seasonal dummies were dropped, then we'd include an intercept in the regression.

the time- $T$  information set,  $\Omega_T$ ), yields the point forecast

$$y_{T+h,T} = T_{T+h}(\theta) + \sum_{i=1}^s \gamma_i D_{i,T+h} + \varepsilon_{T+h,T}.$$

As with the pure trend and seasonal models discussed earlier, the trend and seasonal variables on the right-hand side are perfectly predictable. The only twist concerns the cyclical behavior that may be lurking in the disturbance term, future values of which don't necessarily project to zero, because the disturbance is no longer necessarily white noise. Instead, we construct  $\varepsilon_{T+h,T}$  using the methods we developed for forecasting cycles.

As always, we make the point forecast operational by replacing unknown parameters with estimates, yielding

$$\hat{y}_{T+h,T} = T_{T+h}(\hat{\theta}) + \sum_{i=1}^{s\hat{\gamma}_i} D_{i,T+h} + \hat{\varepsilon}_{T+h,T}.$$

To construct  $\hat{\varepsilon}_{T+h,T}$ , in addition to replacing the parameters in the formula for  $\varepsilon_{T+h,T}$  with estimates, we replace the unobservable disturbances, the  $\varepsilon_t$ 's, with the observable residuals, the  $e_t$ 's.

The complete  $h$ -step-ahead density forecast under normality is

$$N(\hat{y}_{T+h,T}, \hat{\sigma}_h^2).$$

where  $\hat{\sigma}_h^2$  is the operational estimate of the variance of the error in forecasting  $\varepsilon_{T+h}$ .

Once again, we don't actually have to *do* any of the computations just discussed; rather, the computer does them all for us. So let's get on with an application, now that we know what we're doing.

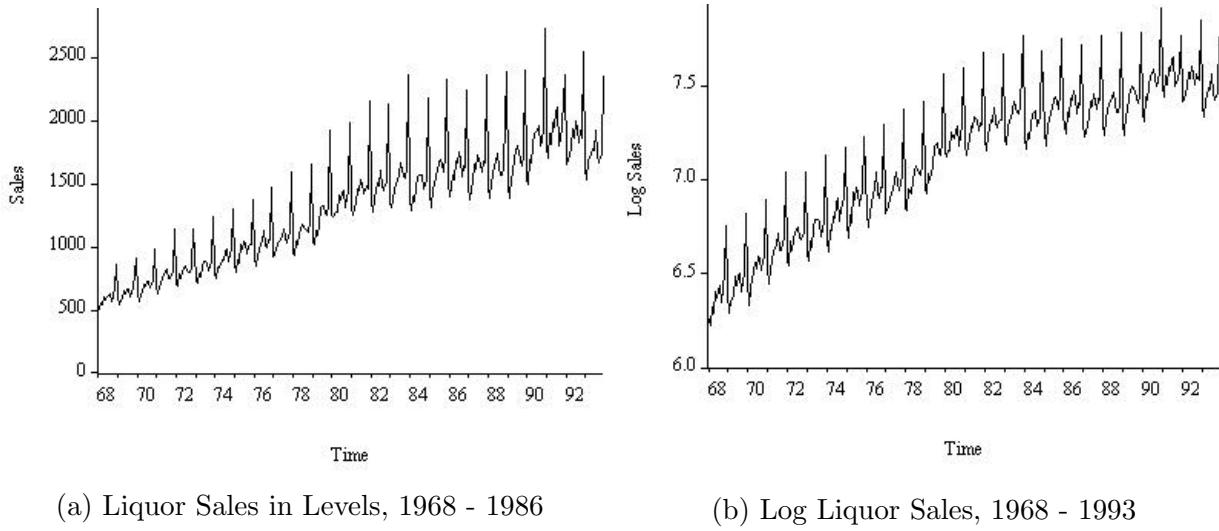


Figure 9.1: Liquor Sales

## 9.2 Lagged Dependent Variables

We use:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + T_t(\theta) + \sum_{i=1}^s \gamma_i D_{it} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

### 9.2.1 Case Study: Forecasting Liquor Sales with Deterministic Trends and Seasonals

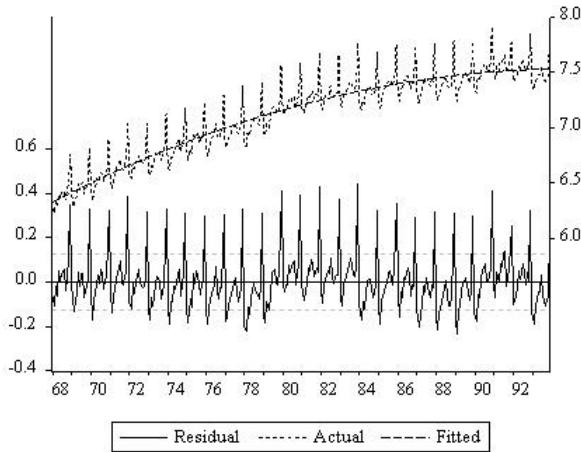
We'll forecast monthly U.S. liquor sales. In Figure 9.1a, we show the history of liquor sales, 1968.01 - 1993.12. Notice its pronounced seasonality – sales skyrocket during the Christmas season. In Figure 9.1b we show log liquor sales; we take logs to stabilize the variance, which grows over time.<sup>4</sup> The variance of log liquor sales is more stable, and it's the series for which we'll build forecasting models.<sup>5</sup>

<sup>4</sup>The nature of the logarithmic transformation is such that it “compresses” an increasing variance. Make a graph of  $\log(x)$  as a function of  $x$ , and you'll see why.

<sup>5</sup>From this point onward, for brevity we'll simply refer to “liquor sales,” but remember that we've taken logs.

LS // Dependent Variable is LSALES				
Sample: 1968:01 1993:12				
Included observations: 312				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.237356	0.024496	254.6267	0.0000
TIME	0.007690	0.000336	22.91552	0.0000
TIME2	-1.14E-05	9.74E-07	-11.72695	0.0000
R-squared	0.892394	Mean dependent var	7.112383	
Adjusted R-squared	0.891698	S.D. dependent var	0.379308	
S.E. of regression	0.124828	Akaike info criterion	-4.152073	
Sum squared resid	4.814823	Schwarz criterion	-4.116083	
Loglikelihood	208.0146	F-statistic	1281.296	
Durbin-Watson stat	1.752858	Prob(F-statistic)	0.000000	

(a) Liquor Sales, Quadratic Trend Regression



(b) Liquor Sales, Quadratic Trend Regression - Residual Plot

Figure 9.2: Liquor Sales: Quadratic Trend Model

Liquor sales dynamics also feature prominent trend and cyclical effects. Liquor sales trend upward, and the trend appears nonlinear in spite of the fact that we're working in logs. To handle the nonlinear trend, we adopt a quadratic trend model (in logs). The estimation results are in Table 9.2a. The residual plot (Figure 9.2b) shows that the fitted trend increases at a decreasing rate; both the linear and quadratic terms are highly significant. The adjusted  $R^2$  is 89%, reflecting the fact that trend is responsible for a large part of the variation in liquor sales. The standard error of the regression is .125; it's an estimate of the standard deviation of the error we'd expect to make in forecasting liquor sales if we accounted for trend but ignored

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	0.117	0.117	.056	4.3158	0.038
2	-0.149	-0.165	.056	11.365	0.003
3	-0.106	-0.069	.056	14.943	0.002
4	-0.014	-0.017	.056	15.007	0.005
5	0.142	0.125	.056	21.449	0.001
6	0.041	-0.004	.056	21.979	0.001
7	0.134	0.175	.056	27.708	0.000
8	-0.029	-0.046	.056	27.975	0.000
9	-0.136	-0.080	.056	33.944	0.000
10	-0.205	-0.206	.056	47.611	0.000
11	0.056	0.080	.056	48.632	0.000
12	0.888	0.879	.056	306.26	0.000
13	0.055	-0.507	.056	307.25	0.000
14	-0.187	-0.159	.056	318.79	0.000
15	-0.159	-0.144	.056	327.17	0.000
16	-0.059	-0.002	.056	328.32	0.000
17	0.091	-0.118	.056	331.05	0.000
18	-0.010	-0.055	.056	331.08	0.000
19	0.086	-0.032	.056	333.57	0.000
20	-0.066	0.028	.056	335.03	0.000
21	-0.170	0.044	.056	344.71	0.000
22	-0.231	0.180	.056	362.74	0.000
23	0.028	0.016	.056	363.00	0.000
24	0.811	-0.014	.056	586.50	0.000
25	0.013	-0.128	.056	586.56	0.000
26	-0.221	-0.136	.056	603.26	0.000
27	-0.196	-0.017	.056	616.51	0.000
28	-0.092	-0.079	.056	619.42	0.000
29	0.045	-0.094	.056	620.13	0.000
30	-0.043	0.045	.056	620.77	0.000
31	0.057	0.041	.056	621.89	0.000
32	-0.095	-0.002	.056	625.07	0.000
33	-0.195	0.026	.056	638.38	0.000
34	-0.240	0.088	.056	658.74	0.000
35	0.006	-0.089	.056	658.75	0.000
36	0.765	0.076	.056	866.34	0.000

Figure 9.3: Liquor Sales, Quadratic Trend - Residual Correlogram

seasonality and serial correlation. The Durbin-Watson statistic provides no evidence against the hypothesis that the regression disturbance is white noise.

The residual plot, however, shows obvious residual seasonality. The Durbin-Watson statistic missed it, evidently because it's not designed to have power against seasonal dynamics.<sup>6</sup> The residual plot also suggests that there may be a cycle in the residual, although it's hard to tell (hard for the Durbin-Watson statistic as well), because the pervasive seasonality swamps the picture and makes it hard to infer much of anything.

The residual correlogram (Table 9.3) and its graph (Figure 9.4) confirm the importance of the neglected seasonality. The residual sample autocor-

---

<sup>6</sup>Recall that the Durbin-Watson test is designed to detect simple AR(1) dynamics. It also has the ability to detect other sorts of dynamics, but evidently not those relevant to the present application, which are very different from a simple AR(1).

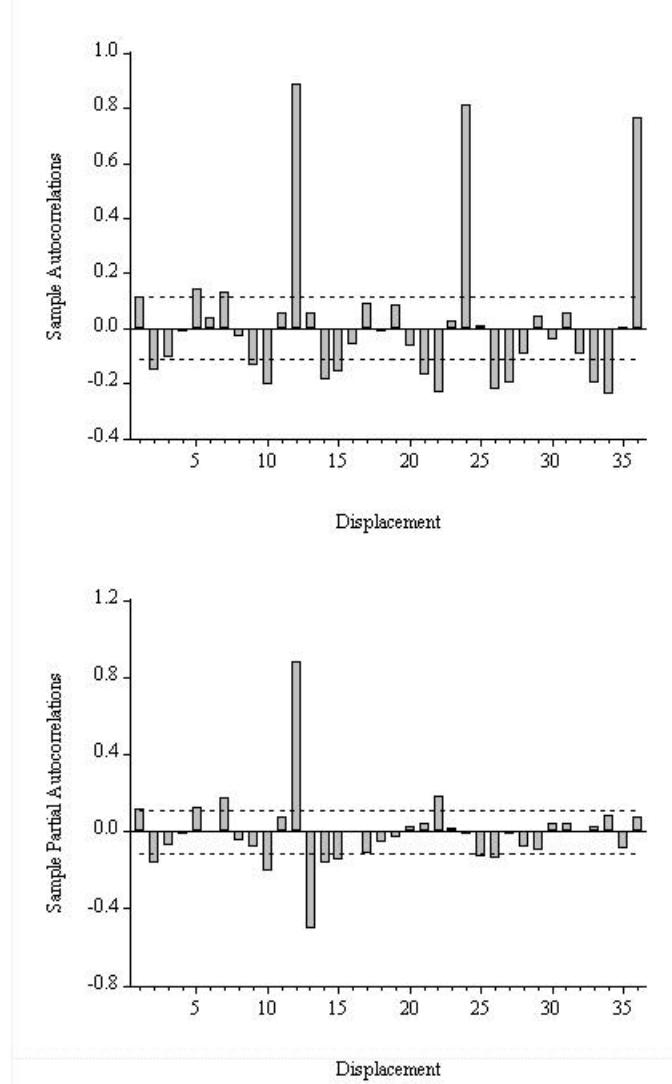


Figure 9.4: Liquor Sales, Quadratic Trend Regression - Residual Sample Autocorrelation

relation function has large spikes, far exceeding the Bartlett bands, at the seasonal displacements, 12, 24, and 36. It indicates some cyclical dynamics as well; apart from the seasonal spikes, the residual sample autocorrelation and partial autocorrelation functions oscillate, and the Ljung-Box statistic rejects the white noise null hypothesis even at very small, non-seasonal, displacements.

In Table 9.5a we show the results of regression on quadratic trend and a full set of seasonal dummies. The quadratic trend remains highly significant. The adjusted  $R^2$  rises to 99%, and the standard error of the regression falls

to .046, which is an estimate of the standard deviation of the forecast error we expect to make if we account for trend and seasonality but ignore serial correlation. The Durbin-Watson statistic, however, has greater ability to detect serial correlation now that the residual seasonality has been accounted for, and it sounds a loud alarm.

The residual plot of Figure 9.5b shows no seasonality, as that's now picked up by the model, but it confirms the Durbin-Watson's warning of serial correlation. The residuals are highly persistent, and hence predictable. We show the residual correlogram in tabular and graphical form in Table 9.6 and Figure 9.7. The residual sample autocorrelations oscillate and decay slowly, and they exceed the Bartlett standard errors throughout. The Ljung-Box test strongly rejects the white noise null at all displacements. Finally, the residual sample partial autocorrelations cut off at displacement 3. All of this suggests that an AR(3) would provide a good approximation to the disturbance's Wold representation.

In Table 9.8a, then, we report the results of estimating a liquor sales model with quadratic trend, seasonal dummies, and AR(3) disturbances. The  $R^2$  is now 100%, and the Durbin-Watson is fine. One inverse root of the AR(3) disturbance process is estimated to be real and close to the unit circle (.95), and the other two inverse roots are a complex conjugate pair farther from the unit circle. The standard error of this regression is an estimate of the standard deviation of the forecast error we'd expect to make after modeling the residual serial correlation, as we've now done; that is, it's an estimate of the standard deviation of  $v$ .<sup>7</sup>

We show the residual plot in Figure 9.8b and the residual correlogram in Table 9.9 and Figure fig: liquor sales quadratic seasonal dummies and ar(3) residual sample autocorrelation. The residual plot reveals no patterns; instead, the residuals look like white noise, as they should. The residual

---

<sup>7</sup>Recall that  $v$  is the innovation that drives the ARMA process for the regression disturbance,  $\varepsilon$ . It's a very small .027, roughly half that obtained when we ignored serial correlation.

sample autocorrelations and partial autocorrelations display no patterns and are mostly inside the Bartlett bands. The Ljung-Box statistics also look good for small and moderate displacements, although their p-values decrease for longer displacements.

All things considered, the quadratic trend, seasonal dummy, AR(3) specification seems tentatively adequate. We also perform a number of additional checks. In Figure 9.11, we show a histogram and normality test applied to the residuals. The histogram looks symmetric, as confirmed by the skewness near zero. The residual kurtosis is a bit higher than three and causes Jarque-Bera test to reject the normality hypothesis with a p-value of .02, but the residuals nevertheless appear to be fairly well approximated by a normal distribution, even if they may have slightly fatter tails.

Now we use the estimated model to produce forecasts. In Figure 9.12 we show the history of liquor sales and a 12-month-ahead extrapolation forecast for 1994.<sup>8</sup> To aid visual interpretation, we show only two years of history. The forecast looks reasonable. It's visually apparent that the model has done a good job of picking up the seasonal pattern, which dominates the local behavior of the series. In Figure 9.13, we show the history, the forecast, and the 1994 realization. The forecast was very good!

In Figure 9.14 we show four years of history together with a 60-month-ahead (five year) extrapolation forecast, to provide a better feel for the dynamics in the forecast. The figure also makes clear the trend forecast is slightly *downward*. To put the long-horizon forecast in historical context, we show in Figure 13 the 60-month-ahead forecast together with the complete history. Finally, in Figure 14, we show the history and point forecast of the level of liquor sales (as opposed to log liquor sales), which we obtain by exponentiating the forecast of log liquor sales.<sup>9</sup>

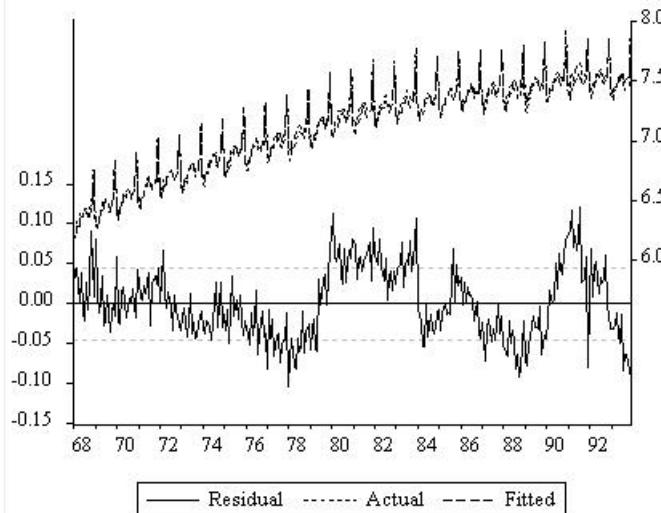
---

<sup>8</sup>We show the point forecast together with 95% intervals.

<sup>9</sup>Recall that exponentiating “undoes” a natural logarithm.

LS // Dependent Variable is LSALES				
Sample: 1968:01 1993:12				
Included observations: 312				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
TIME	0.007656	0.000123	62.35882	0.0000
TIME2	-1.14E-05	3.56E-07	-32.06823	0.0000
D1	6.147456	0.012340	498.1699	0.0000
D2	6.088653	0.012353	492.8890	0.0000
D3	6.174127	0.012366	499.3008	0.0000
D4	6.175220	0.012378	498.8970	0.0000
D5	6.246086	0.012390	504.1398	0.0000
D6	6.250387	0.012401	504.0194	0.0000
D7	6.295979	0.012412	507.2402	0.0000
D8	6.268043	0.012423	504.5509	0.0000
D9	6.203832	0.012433	498.9630	0.0000
D10	6.229197	0.012444	500.5968	0.0000
D11	6.259770	0.012453	502.6602	0.0000
D12	6.580068	0.012463	527.9819	0.0000
R-squared	0.986111	Mean dependent var		7.112383
Adjusted R-squared	0.985505	S.D. dependent var		0.379308
S.E. of regression	0.045666	Akaike info criterion		-6.128963
Sum squared resid	0.621448	Schwarz criterion		-5.961008
Log likelihood	527.4094	F-statistic		1627.567
Durbin-Watson stat	0.586187	Prob(F-statistic)		0.000000

(a) Liquor Sales, Quadratic Trend with Seasonal Dummies



(b) Liquor Sales, Quadratic Trend with Seasonal Dummies - Residual Plot

Figure 9.5: Liquor Sales - Trend and Seasonal Model

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	0.700	0.700	.056	154.34	0.000
2	0.686	0.383	.056	302.86	0.000
3	0.725	0.369	.056	469.36	0.000
4	0.569	-0.141	.056	572.36	0.000
5	0.569	0.017	.056	675.58	0.000
6	0.577	0.093	.056	782.19	0.000
7	0.460	-0.078	.056	850.06	0.000
8	0.480	0.043	.056	924.38	0.000
9	0.466	0.030	.056	994.46	0.000
10	0.327	-0.188	.056	1029.1	0.000
11	0.364	0.019	.056	1072.1	0.000
12	0.355	0.089	.056	1113.3	0.000
13	0.225	-0.119	.056	1129.9	0.000
14	0.291	0.065	.056	1157.8	0.000
15	0.211	-0.119	.056	1172.4	0.000
16	0.138	-0.031	.056	1178.7	0.000
17	0.195	0.053	.056	1191.4	0.000
18	0.114	-0.027	.056	1195.7	0.000
19	0.055	-0.063	.056	1196.7	0.000
20	0.134	0.089	.056	1202.7	0.000
21	0.062	0.018	.056	1204.0	0.000
22	-0.006	-0.115	.056	1204.0	0.000
23	0.084	0.086	.056	1206.4	0.000
24	-0.039	-0.124	.056	1206.9	0.000
25	-0.063	-0.055	.056	1208.3	0.000
26	-0.016	-0.022	.056	1208.4	0.000
27	-0.143	-0.075	.056	1215.4	0.000
28	-0.135	-0.047	.056	1221.7	0.000
29	-0.124	-0.048	.056	1227.0	0.000
30	-0.189	0.086	.056	1239.5	0.000
31	-0.178	-0.017	.056	1250.5	0.000
32	-0.139	0.073	.056	1257.3	0.000
33	-0.226	-0.049	.056	1275.2	0.000
34	-0.155	0.097	.056	1283.7	0.000
35	-0.142	0.008	.056	1290.8	0.000
36	-0.242	-0.074	.056	1311.6	0.000

Figure 9.6: Liquor Sales, Quadratic Trend with Seasonal Dummies - Residual Correlogram

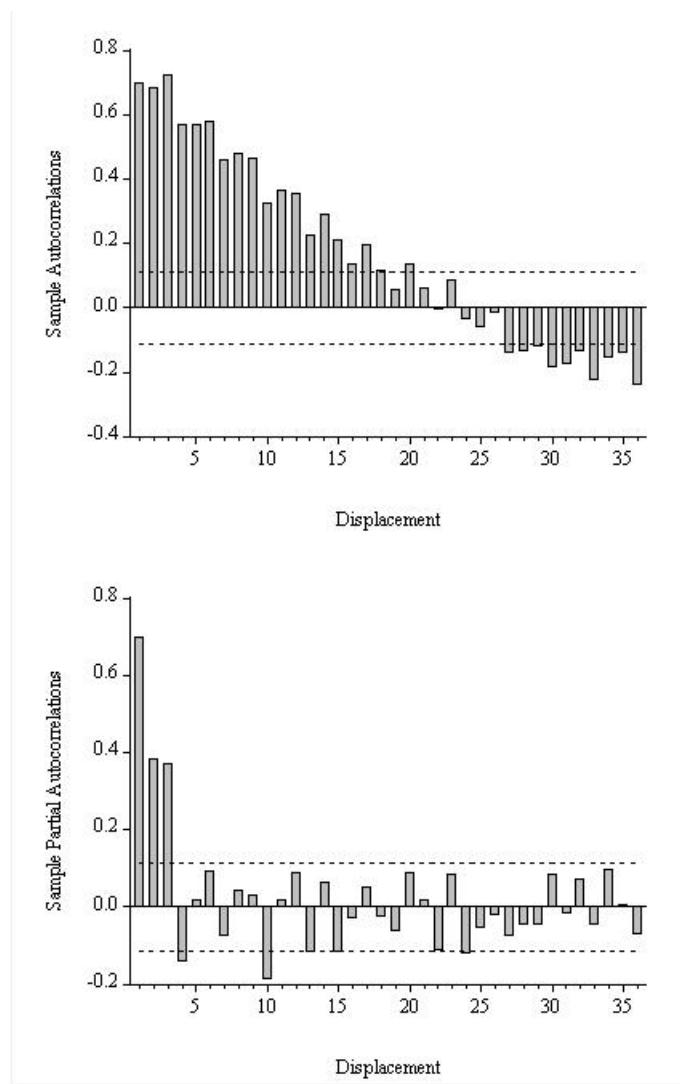
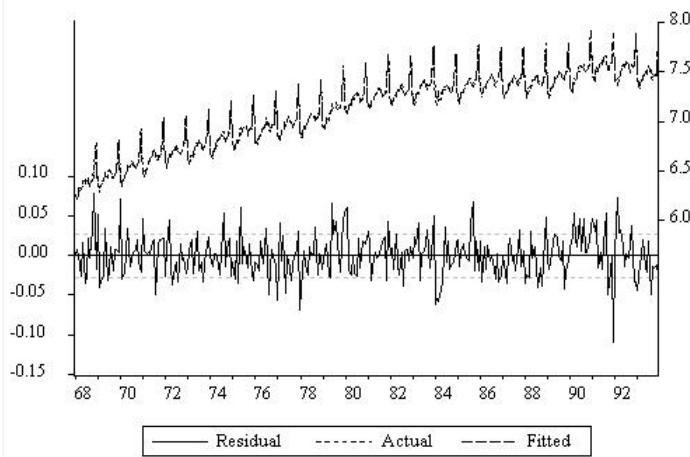


Figure 9.7: Liquor Sales, Quadratic Trend with Seasonal Dummies - Residual Sample Autocorrelation

LS // Dependent Variable is LSALES				
Sample: 1968:01 1993:12				
Included observations: 312				
Convergence achieved after 4 iterations				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
TIME	0.008606	0.000981	8.768212	0.0000
TIME2	-1.41E-05	2.53E-06	-5.556103	0.0000
D1	6.073054	0.083922	72.36584	0.0000
D2	6.013822	0.083942	71.64254	0.0000
D3	6.099208	0.083947	72.65524	0.0000
D4	6.101522	0.083934	72.69393	0.0000
D5	6.172528	0.083946	73.52962	0.0000
D6	6.177129	0.083947	73.58364	0.0000
D7	6.223323	0.083939	74.14071	0.0000
D8	6.195681	0.083943	73.80857	0.0000
D9	6.131818	0.083940	73.04993	0.0000
D10	6.157592	0.083934	73.36197	0.0000
D11	6.188480	0.083932	73.73176	0.0000
D12	6.509106	0.083928	77.55624	0.0000
AR(1)	0.268805	0.052909	5.080488	0.0000
AR(2)	0.239688	0.053697	4.463723	0.0000
AR(3)	0.395880	0.053109	7.454150	0.0000
R-squared	0.995069	Mean dependent var	7.112383	
Adjusted R-squared	0.994802	S.D. dependent var	0.379308	
S.E. of regression	0.027347	Akaike info criterion	-7.145319	
Sum squared resid	0.220625	Schwarz criteron	-6.941373	
Log likelihood	688.9610	F-statistic	3720.875	
Durbin-Watson stat	1.886119	Prob(F-statistic)	0.000000	
Inverted AR Roots	.95		-.34+.55i	-.34 -.55i

(a) Liquor Sales, Quadratic Trend with Seasonal Dummies and AR(3)



(b) Liquor Sales, Quadratic Trend with Seasonal Dummies and AR(3) Disturbances - Residual Plot

Figure 9.8: Liquor Sales - Trend, Seasonal, and AR(3) Model

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	0.056	0.056	.056	0.9779	0.323
2	0.037	0.034	.056	1.4194	0.492
3	0.024	0.020	.056	1.6032	0.659
4	-0.084	-0.088	.056	3.8256	0.430
5	-0.007	0.001	.056	3.8415	0.572
6	0.065	0.072	.056	5.1985	0.519
7	-0.041	-0.044	.056	5.7288	0.572
8	0.069	0.063	.056	7.2828	0.506
9	0.080	0.074	.056	9.3527	0.405
10	-0.163	-0.169	.056	18.019	0.055
11	-0.009	-0.005	.056	18.045	0.081
12	0.145	0.175	.056	24.938	0.015
13	-0.074	-0.078	.056	26.750	0.013
14	0.149	0.113	.056	34.034	0.002
15	-0.039	-0.060	.056	34.532	0.003
16	-0.089	-0.058	.056	37.126	0.002
17	0.058	0.048	.056	38.262	0.002
18	-0.062	-0.050	.056	39.556	0.002
19	-0.110	-0.074	.056	43.604	0.001
20	0.100	0.056	.056	46.935	0.001
21	0.039	0.042	.056	47.440	0.001
22	-0.122	-0.114	.056	52.501	0.000
23	0.146	0.130	.056	59.729	0.000
24	-0.072	-0.040	.056	61.487	0.000
25	0.006	0.017	.056	61.500	0.000
26	0.148	0.082	.056	69.024	0.000
27	-0.109	-0.067	.056	73.145	0.000
28	-0.029	-0.045	.056	73.436	0.000
29	-0.046	-0.100	.056	74.153	0.000
30	-0.084	0.020	.056	76.620	0.000
31	-0.095	-0.101	.056	79.793	0.000
32	0.051	0.012	.056	80.710	0.000
33	-0.114	-0.061	.056	85.266	0.000
34	0.024	0.002	.056	85.468	0.000
35	0.043	-0.010	.056	86.116	0.000
36	-0.229	-0.140	.056	104.75	0.000

Figure 9.9: Liquor Sales, Quadratic Trend with Seasonal Dummies and AR(3) Disturbances  
- Residual Correlogram

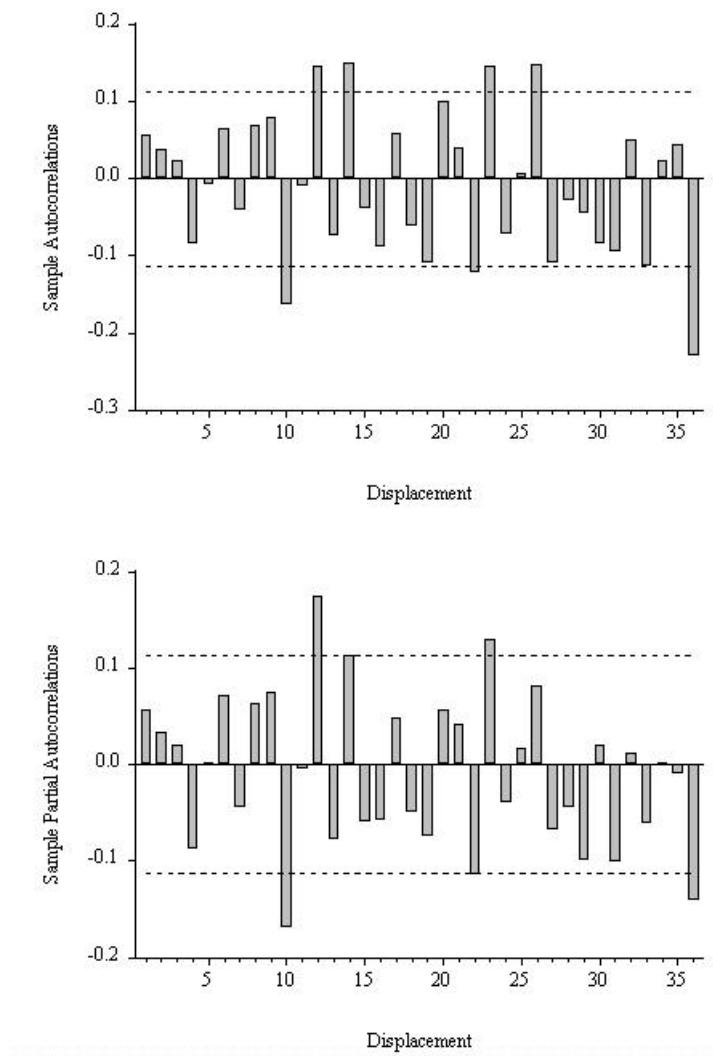


Figure 9.10: Liquor Sales, Quadratic Trend with Seasonal Dummies and AR(3) Disturbances  
- Residual Sample Autocorrelation

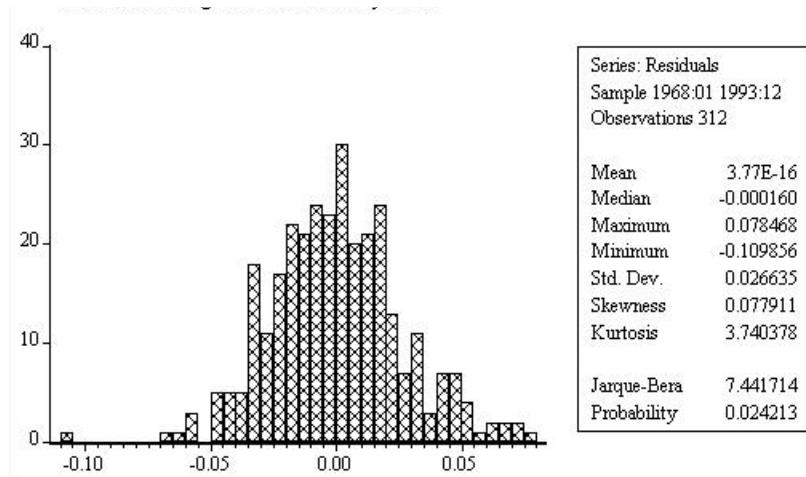


Figure 9.11: Liquor Sales, Quadratic Trend with Seasonal Dummies and AR(3) Disturbances - Residual Histogram and Normality test

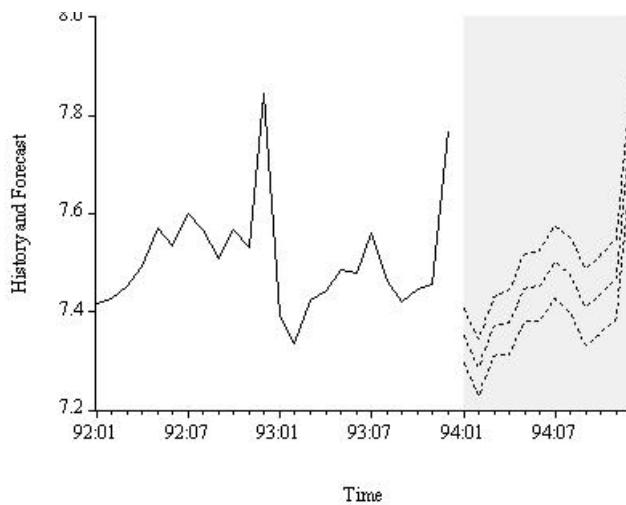


Figure 9.12: Liquor Sales: History and 12-Month-Ahead Forecast

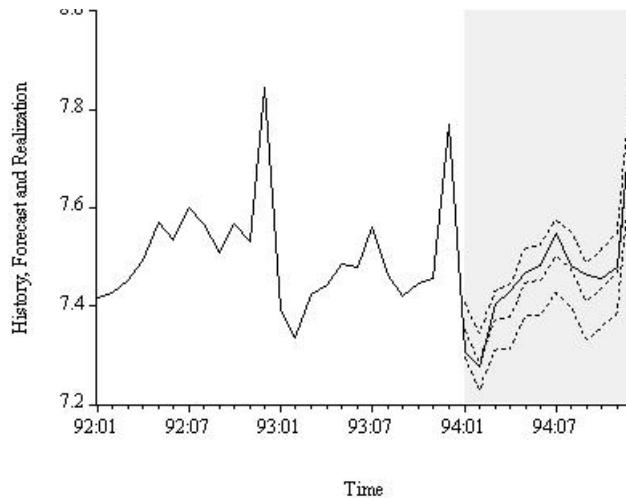


Figure 9.13: Liquor Sales: History, 12-Month-Ahead Forecast, and Realization

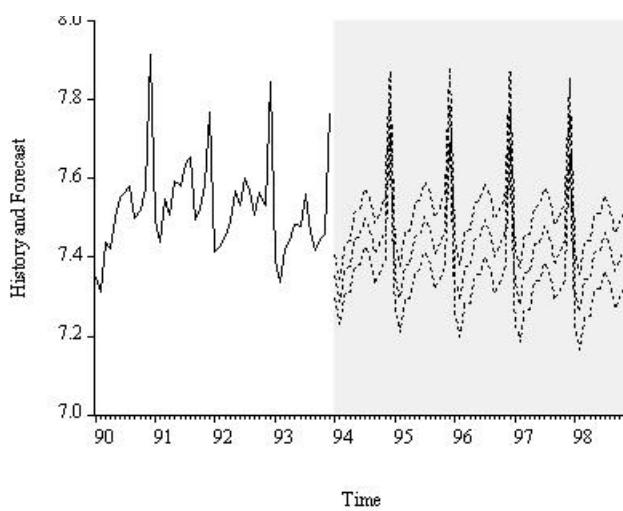


Figure 9.14: Liquor Sales: History and Four-Year-Ahead Forecast

### 9.3 Exercises, Problems and Complements

1. Serially correlated disturbances vs. lagged dependent variables. Estimate the quadratic trend model for log liquor sales with seasonal dummies and three lags of the dependent variable included directly. Discuss your results and compare them to those we obtained when we instead allowed for AR(3) disturbances in the regression. Which model is selected by AIC and SIC?
2. Assessing the adequacy of the liquor sales forecasting model deterministic trend specification. Critique the liquor sales forecasting model that we adopted (log liquor sales with quadratic trend, seasonal dummies, and AR(3) disturbances).<sup>10</sup>
  - a. If the trend is not a good approximation to the actual trend in the series, would it greatly affect short-run forecasts? Long-run forecasts?
  - b. How might you fit and assess the adequacy of a broken linear trend? How might you decide on the location of the break point?
3. Improving non-trend aspects of the liquor sales forecasting model.
  - a. Recall our argument that best practice requires using a  $\chi^2_{m-k}$  distribution rather than a  $\chi^2_m$  distribution to assess the significance of  $Q$ -statistics for model residuals, where  $m$  is the number of autocorrelations included in the  $Q$  statistic and  $k$  is the number of parameters estimated. In several places in this chapter, we failed to heed this advice when evaluating the liquor sales model. If we were instead to compare the residual  $Q$ -statistic  $p$ -values to a  $\chi^2_{m-k}$  distribution, how, if at all, would our assessment of the model's adequacy change?
  - b. Return to the log-quadratic trend model with seasonal dummies, allow for  $ARMA(p, q)$  disturbances, and do a systematic selection of  $p$  and  $q$ .

---

<sup>10</sup>I thank Ron Michener, University of Virginia, for suggesting parts d and f.

$q$  using  $AIC$  and  $SIC$ . Do  $AIC$  and  $SIC$  select the same model? If not, which do you prefer? If your preferred disturbance model differs from the  $AR(3)$  that we used, replicate the analysis in the text using your preferred model, and discuss your results.

- c. Discuss and evaluate another possible model improvement: inclusion of an additional dummy variable indicating the number of Fridays and/or Saturdays in the month. Does this model have lower  $AIC$  or  $SIC$  than the final model used in the text? Do you prefer it to the one in the text? Why or why not?

## 9.4 Notes



## **Part IV**

# **Forecast Evaluation and Combination**



# Chapter 10

## Point Forecast Evaluation

As we've stressed repeatedly, good forecasts lead to good decisions. The importance of forecast evaluation techniques follows immediately. Given a track record of forecasts,  $y_{t+h,t}$ , and corresponding realizations,  $y_{t+h}$ , we naturally want to monitor and improve forecast performance. In this chapter we show how to do so. We discuss both absolute aspects of forecast evaluation, focusing on methods for checking forecast optimality, and relative aspects, focusing on methods for ranking forecast accuracy, quite apart from optimality.

### 10.1 Absolute Standards for Point Forecasts

Think about evaluating a single forecast, in isolation. Evaluating a single forecast amounts to checking whether it has the properties expected of an optimal forecast. Denote by  $y_t$  the covariance stationary time series to be forecast. The Wold representation is

$$y_t = \mu + \varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \dots$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

Thus the  $h$ -step-ahead linear least-squares forecast is

$$y_{t+h,t} = \mu + b_h\varepsilon_t + b_{h+1}\varepsilon_{t-1} + \dots$$

and the corresponding  $h$ -step-ahead forecast error is

$$e_{t+h,t} = y_{t+h} - y_{t+h,t} = \varepsilon_{t+h} + b_1\varepsilon_{t+h-1} + \dots + b_{h-1}\varepsilon_{t+1},$$

with variance

$$\sigma_h^2 = \sigma^2 \left( 1 + \sum_{i=1}^{h-1} b_i^2 \right).$$

The key property of optimal forecast errors, from which all others follow, (including those cataloged below), is that they should be unforecastable on the basis of information available at the time the forecast was made. This **unforecastability principle** is valid in great generality; it holds, for example, regardless of whether linear-projection optimality or conditional-mean optimality is of interest, regardless of whether the relevant loss function is quadratic, and regardless of whether the series being forecast is stationary.

Many tests of aspects of optimality are based on the unforecastability principle. 1-step-ahead errors, for example, had better be white noise, because otherwise we could forecast the errors using information readily available when the forecast is made. Indeed at least four key properties of optimal forecasts, which we can easily check, follow immediately from the unforecastability principle:

- a. Optimal forecasts are unbiased
- b. Optimal forecasts have 1-step-ahead errors that are white noise
- c. Optimal forecasts have  $h$ -step-ahead errors that are at most  $MA(h-1)$
- d. Optimal forecasts have  $h$ -step-ahead errors with variances that are non-decreasing in  $h$  and that converge to the unconditional variance of the process.

### 10.1.1 Are errors zero-mean?

If the forecast is unbiased, then the forecast error has a zero mean. A variety of tests of the zero-mean hypothesis can be performed, depending on the assumptions we're willing to maintain. For example, if  $e_{t+h,t}$  is Gaussian white noise (as might be reasonably the case for 1-step-ahead errors), then the standard  $t$ -test is the obvious choice. We would simply regress the forecast error series on a constant and use the reported  $t$ -statistic to test the hypothesis that the population mean is zero. If the errors are non-Gaussian but remain iid, then the  $t$ -test is still applicable in large samples.

If the forecast errors are dependent, then more sophisticated procedures are required. We maintain the framework of regressing on a constant, but we must “correct” for any serial correlation in the disturbances. Serial correlation in forecast errors can arise for many reasons. Multi-step-ahead forecast errors will be serially correlated, even if the forecasts are optimal, because of the forecast-period overlap associated with multi-step-ahead forecasts. More generally, serial correlation in forecast errors may indicate that the forecasts are suboptimal. The upshot is simply that when regressing forecast errors on an intercept, we need to be sure that any serial correlation in the disturbance is appropriately modeled. A reasonable starting point for a regression involving  $h$ -step-ahead forecast errors is  $MA(h - 1)$  disturbances, which we'd expect if the forecast were optimal. The forecast may, of course, *not* be optimal, so we don't adopt  $MA(h - 1)$  disturbances uncritically; instead, we try a variety of models using the AIC and SIC to guide selection in the usual way.

### 10.1.2 Are 1-step-ahead errors white noise?

Under various sets of maintained assumptions, we can use standard tests of the white noise hypothesis. For example, the sample autocorrelation and

partial autocorrelation functions, together with Bartlett asymptotic standard errors, are often useful in that regard. Tests based on the first autocorrelation (e.g., the Durbin-Watson test), as well as more general tests, such as the Box-Pierce and Ljung-Box tests, are useful as well.

### 10.1.3 Are $h$ -step-ahead errors are at most $MA(h - 1)$ ?

The  $MA(h - 1)$  structure implies a cutoff in the forecast error's autocorrelation function beyond displacement  $h - 1$ . This immediately suggests examining the statistical significance of the sample autocorrelations beyond displacement  $h - 1$  using the Bartlett standard errors. In addition, we can regress the errors on a constant, allowing for  $MA(q)$  disturbances with  $q > (h - 1)$ , and test whether the moving-average parameters beyond lag  $h - 1$  are zero.

### 10.1.4 Are $h$ -step-ahead error variances non-decreasing in $h$ ?

It's often useful to examine the sample  $h$ -step-ahead forecast error variances as a function of  $h$ , both to be sure they're non-decreasing in  $h$  and to see their *pattern*, which may convey useful information.

### 10.1.5 Are errors orthogonal to available information?

The tests above make incomplete use of the unforecastability principle, insofar as they assess only the *univariate* properties of the errors. We can make a more complete assessment by broadening the information set and assessing optimality with respect to various sets of information, by estimating regressions of the form

$$e_{t+h,t} = \alpha_0 + \sum \alpha_i x_{it} + u_t.$$

The hypothesis of interest is that all the  $\alpha$ 's are zero, which is a necessary condition for forecast optimality (orthogonality) with respect to available

information.

The particular case of testing optimality with respect to  $y_{t+h,t}$  is very important in practice. (Note that  $y_{t+h,t}$  is obviously in the time- $t$  information set.) The relevant regression is

$$e_{t+h,t} = \alpha_0 + \alpha_1 y_{t+h,t} + u_t,$$

and optimality corresponds to  $(\alpha_0, \alpha_1) = (0, 0)$ .

If the above regression seems a little strange to you, consider what may seem like a more natural approach to testing optimality, regression of the realization on the forecast:

$$y_{t+h} = \beta_0 + \beta_1 y_{t+h,t} + u_t.$$

This is called a “**Mincer-Zarnowitz regression**.” If the forecast is optimal with respect to the information used to construct it, then we’d expect  $(\beta_0, \beta_1) = (0, 1)$ , in which case

$$y_{t+h} = y_{t+h,t} + u_t.$$

Note, however, that if we start with the regression

$$y_{t+h} = \beta_0 + \beta_1 y_{t+h,t} + u_t,$$

and then subtract  $y_{t+h,t}$  from each side, we obtain

$$e_{t+h,t} = \alpha_0 + \alpha_1 y_{t+h,t} + u_t,$$

where  $(\alpha_0, \alpha_1) = (0, 0)$  when  $(\beta_0, \beta_1) = (0, 1)$ . Thus, the two approaches are identical. We can regress the error on an intercept and the forecast and test  $(0, 0)$ , or we can regress the realization on an intercept and the forecast and test  $(0, 1)$ .

## 10.2 Relative Standards for Point Forecasts

Now think about ranking a set of forecasts, quite apart from how any or all of them fare regarding the absolute optimality criteria assessed in section 10.1.

### 10.2.1 Accuracy Rankings via Expected Loss

The crucial object in measuring forecast accuracy is the loss function,  $L(y_{t+h}, y_{t+h,t})$ , often restricted to  $L(e_{t+h,t})$ , which charts the “loss,” “cost,” or “disutility” associated with various pairs of forecasts and realizations.<sup>1</sup> In addition to the shape of the loss function, the forecast horizon  $h$  is of crucial importance. Rankings of forecast accuracy may of course be very different across different loss functions and different horizons.

Let’s discuss a few accuracy measures that are important and popular. Accuracy measures are usually defined on the forecast errors,

$$e_{t+h,t} = y_{t+h} - y_{t+h,t},$$

or percent errors,

$$p_{t+h,t} = (y_{t+h} - y_{t+h,t})/y_{t+h}.$$

**Mean error** measures forecast-error location, which is one component of accuracy. In population we write

$$\mu_{e_{t+h,t}} = E(e_{t+h,t}),$$

and in sample we write

$$\hat{\mu}_{e_{t+h,t}} = \frac{1}{T} \sum_{t=1}^T e_{t+h,t}.$$

The mean error is the forecast **bias**. Other things the same, we prefer a

---

<sup>1</sup>Because in many applications the loss function will be a direct function of the forecast error,  $L(y_t, y_{t+h,t}) = L(e_{t+h,t})$ , we write  $L(e_{t+h,t})$  from this point on to economize on notation, while recognizing that certain loss functions (such as direction-of-change) don’t collapse to the  $L(e_{t+h,t})$  form.

forecast with small bias.

**Error variance** measures dispersion of the forecast errors, which is another component of accuracy. In population we write

$$\sigma_{e_{t+h,t}}^2 = E(e_{t+h,t} - \mu_{e_{t+h,t}})^2,$$

and in sample we write

$$\hat{\sigma}_{e_{t+h,t}}^2 = \frac{1}{T} \sum_{t=1}^T (e_{t+h,t} - \hat{\mu}_{e_{t+h,t}})^2.$$

Other things the same, we prefer a forecast with small error variance.

Although the mean error and the error variance are components of accuracy, neither provides an overall accuracy measure. For example, one forecast might have a small  $\hat{\mu}_{e_{t+h,t}}$  but a large  $\hat{\sigma}_{e_{t+h,t}}^2$ , and another might have a large  $\hat{\mu}_{e_{t+h,t}}$  and a small  $\hat{\sigma}_{e_{t+h,t}}^2$ . Hence we would like an accuracy measure that somehow incorporates *both* the mean error and error variance.

The **mean squared error** does just that. It is the most common overall accuracy measure, by far. In population we write

$$MSE_{e_{t+h,t}} = E(e_{t+h,t})^2,$$

and in sample we write

$$\widehat{MSE}_{e_{t+h,t}} = \frac{1}{T} \sum_{t=1}^T e_{t+h,t}^2.$$

This “bias-variance tradeoff” is a crucially important insight for forecasting. Among other things, it highlights the fact that bias is not necessarily “bad,” under quadratic loss ( $MSE$ ). We’d be happy, for example, to take a small bias increase in exchange for a massive variance reduction.

We sometimes take square roots to preserve units, yielding the **root mean**

**squared error.** In population we write

$$RMSE_{e_{t+h,t}} = \sqrt{E(e_{t+h,t})^2},$$

and in sample we write

$$\widehat{RMSE}_{e_{t+h,t}} = \sqrt{\frac{1}{T} \sum_{t=1}^T e_{t+h,t}^2}.$$

To understand the meaning of “preserving units,” and why it’s sometimes helpful to do so, suppose that the forecast errors are measured in dollars. Then the mean squared error, which is built up from *squared* errors, is measured in dollars *squared*. Taking square roots – that is, moving from MSE to RMSE – brings the units back to dollars.

MSE can be decomposed into bias and variance components, reflecting the tradeoff between bias and variance forecast accuracy under quadratic loss. In particular, MSE can be decomposed into the sum of variance and squared bias. In population we write

$$MSE_{e_{t+h,t}} = \sigma_{e_{t+h,t}}^2 + \mu_{e_{t+h,t}}^2,$$

and in sample we write

$$\widehat{MSE}_{e_{t+h,t}} = \hat{\sigma}_{e_{t+h,t}}^2 + \hat{\mu}_{e_{t+h,t}}^2.$$

**Mean absolute error** is a less popular, but nevertheless common, overall accuracy measure. In population we write

$$MAE_{e_{t+h,t}} = E|e_{t+h,t}|,$$

and in sample we write

$$\widehat{MAE} = \frac{1}{T} \sum_{t=1}^T |e_{t+h,t}|.$$

When using MAE we don't have to take square roots to preserve units.

### 10.2.2 On MSE vs. MAE

Introspection suggests using MAE – not MSE – as the canonical benchmark loss function. Consider using the distribution of  $e$  directly, ranking forecasts by the distance of  $F(e)$  from  $F^*(\cdot)$ , the unit step function at 0 (the cdf of errors from a perfect forecast, which are 0 w.p. 1). That is, rank forecasts by

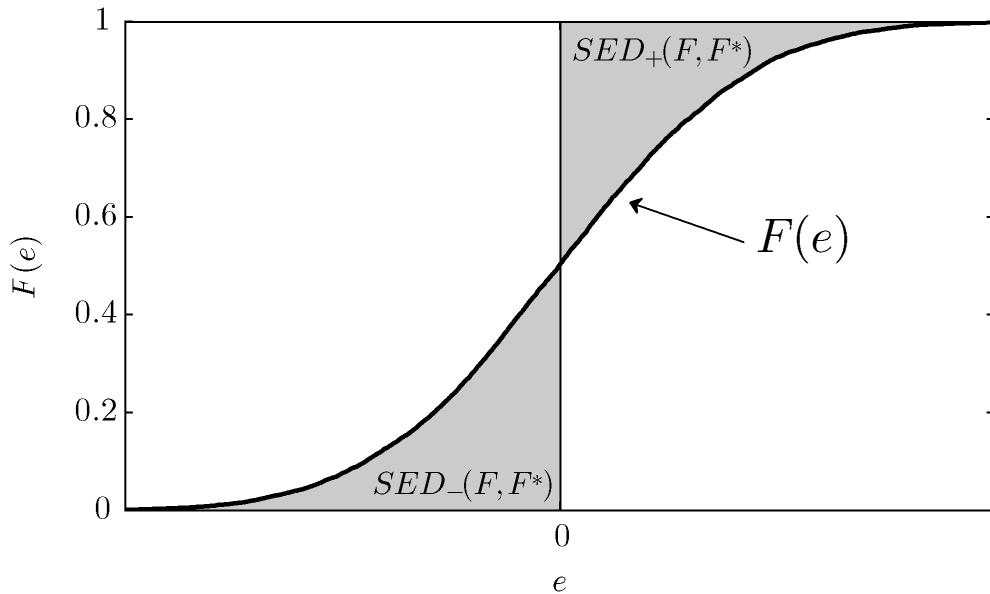
$$SED(F, F^*) = \int_{-\infty}^{\infty} |F(e) - F^*(e)| de,$$

where smaller is better. We call  $SED(F, F^*)$  the *stochastic error distance*. In Figure 10.1a we show  $SED(F, F^*)$ , and in Figure 10.1b we provide an example of two error distributions such that one would prefer  $F_1$  to  $F_2$  under  $SED(F, F^*)$ .

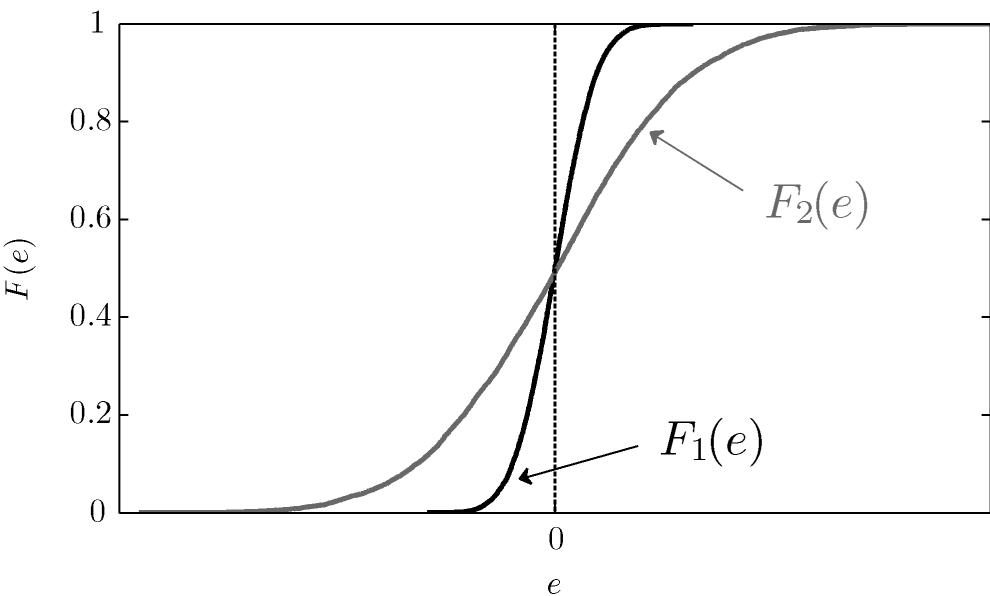
We motivated  $SED(F, F^*)$  as directly appealing and intuitive. It turns out, moreover, that  $SED(F, F^*)$  is intimately connected to one, and only one, traditionally-invoked loss function, and it is not quadratic. In particular, for any forecast error  $e$ , with cumulative distribution function  $F(e)$  such that  $E(|e|) < \infty$ , we have

$$SED(F, F^*) = \int_{-\infty}^0 F(e) de + \int_0^{\infty} [1 - F(e)] de = E(|e|). \quad (10.1)$$

That is,  $SED(F, F^*)$  equals expected absolute loss for any error distribution. Hence if one is comfortable with  $SED(F, F^*)$  and wants to use it to evaluate forecast accuracy, then one must also be comfortable with expected absolute-error loss and want to use it to evaluate forecast accuracy. The two criteria



(a) c.d.f. of  $e$ . Under the  $SED(F, F^*)$  criterion, we prefer smaller  $SED(F, F^*) = SED_-(F, F^*) + SED_+(F, F^*)$ .



(b) Two forecast error distributions. Under the  $SED(F, F^*)$  criterion, we prefer  $F_1(e)$  to  $F_2(e)$ .

Figure 10.1: Stochastic Error Distance ( $SED(F, F^*)$ )

are *identical*.

### 10.2.3 Benchmark Comparisons

It is sometimes of interest to compare forecast performance to that of an allegedly-naive benchmark.

#### Predictive $R^2$

Recall the formula for  $R^2$ ,

$$R^2 = 1 - \frac{\sum_{t=1}^T e_t^2}{\sum_{t=1}^T (y_t - \bar{y})^2},$$

where  $e_t$  is the in-sample regression residual. If we replace the  $e_t$ 's with  $e_{t,t-1}$ 's, out-of-sample 1-step forecast errors, then we get the predictive  $R^2$ ,

$$R^2 = 1 - \frac{\sum_{t=1}^T e_{t,t-1}^2}{\sum_{t=1}^T (y_t - \bar{y})^2},$$

Predictive  $R^2$  compares an estimate of 1-step-ahead out-of-sample forecast error variance to an estimate of unconditional variance. Put differently, it compares actual 1-step forecast accuracy to that of the historical mean forecast,  $\bar{y}$ . The hope is that the former is much smaller than the latter, in which case the predictive  $R^2$  will be near 1.

$h$ -step-ahead versions of predictive  $R^2$ 's are immediate. We simply replace  $e_{t,t-1}$  with  $e_{t,t-h}$  in the formulas.

#### Theil's U-Statistic

The so-called “Theil U-statistic” is just a predictive  $R^2$ , but we change the benchmark from the historical mean forecast,  $\bar{y}$ , to a “no change” forecast,

$y_{t-1}$ ,

$$U = 1 - \frac{\sum_{t=1}^T e_{t,t-1}^2}{\sum_{t=1}^T (y_t - y_{t-1})^2}.$$

In the meteorological literature measures like  $U$  are called “skill scores,” because they assess actual skill relative to a potentially-naive forecast.

It is important to note that allegedly-naive benchmarks may not be so naive. For example, many economic variables may in fact be nearly random walks, in which case forecasters will have great difficulty beating the random walk through no fault of their own (i.e., the predictive  $R^2$  relative to a random walk “no-change” forecast given by Theil’s  $U$  may be near 0)!

#### 10.2.4 Measures of Forecastability

Forecastability measures are a leading example of benchmark comparisons, as we discuss them here.

It is natural and informative to judge forecasts by their accuracy. However, actual and forecasted values will differ, even for good forecasts. To take an extreme example, consider a zero-mean white noise process. The optimal linear forecast under quadratic loss in this case is simply zero, so the paths of forecasts and realizations will clearly look different. These differences illustrate the inherent limits to predictability, even when using optimal forecasts. The extent of a series’ predictability depends on how much information the past conveys regarding future values of this series; as a result, some processes are inherently easy to forecast, and others are more difficult. Note also that predictability and volatility are different concepts; predictability is about the *ratio* of conditional to unconditional variance, whereas volatility is simply about unconditional variance.

Below we discuss some of the difficulties involved in predictability measurement and propose a simple measure of relative predictability based on the ratio of the expected loss of an optimal short-run forecast to the expected loss

of an optimal long-run forecast. Our measure allows for covariance stationary or difference stationary processes, univariate or multivariate information sets, general loss functions, and different forecast horizons of interest. First we propose parametric methods for estimating the predictability of observed series, and then we discuss alternative nonparametric measures, survey-based measures, and more.

### Population Measures

The expected loss of an optimal forecast will in general exceed zero, which illustrates the inherent limits to predictability, even when using optimal forecasts. Put differently, poor forecast accuracy does not necessarily imply that the forecaster failed. The extent of a series' predictability in population depends on how much information the past conveys regarding the future; given an information set, some processes are inherently easy to forecast, and others are more difficult.

In measuring predictability it is important to keep two points in mind. First, the question of whether a series is predictable or not should be replaced by one of *how* predictable it is. Predictability is always a matter of degree. Second, the question of how predictable a series is cannot be answered in general. We have to be clear about the relevant forecast horizon and loss function. For example, a series may be highly predictable at short horizons, but not at long horizons.

A natural measure of the forecastability of covariance stationary series under squared-error loss, patterned after the familiar regression  $R^2$ , is

$$G = 1 - \frac{\text{var}(e_{t+j,t})}{\text{var}(y_{t+j})},$$

where  $\hat{y}_{t+j,t}$  is the optimal (i.e., conditional mean) forecast and  $e_{t+j,t} = y_{t+j} - \hat{y}_{t+j,t}$ .

We can also relax several constraints that limit the broad applicability of the predictive  $R^2$  above. Its essence is basing measures of predictability on

the difference between the conditionally expected loss of an optimal short-run forecast,  $E(L(e_{t+j,t}))$ , and that of an optimal long-run forecast,  $E(L(e_{t+k,t}))$ , ,  $j \ll k$ , , where  $E(\cdot)$  denotes the mathematical expectation conditional on the information set  $\Omega$ . If  $E(L(e_{t+j,t})) \ll E(L(e_{t+k,t}))$ , we say that the series is highly predictable at horizon  $j$  relative to  $k$ , and if  $E(L(e_{t+j,t})) \approx E(L(e_{t+k,t}))$ . we say that the series is nearly unpredictable at horizon  $j$  relative to  $k$ . Thus, we define a general measure of predictability as

$$P(L, \Omega, j, k) = 1 - \frac{E(L(e_{t+j,t}))}{E(L(e_{t+k,t}))},$$

where the information set  $\Omega$  can be univariate or multivariate, as desired. The predictive  $R^2$  measure emerges when the series is covariance stationary,  $L(x) = x^2$  (and hence the optimal forecast is the conditional mean), the information set is univariate, and  $k = \infty$ . The advantages of our generalization include: (1) It is valid for both covariance stationary and difference stationary series, so long as  $k < \infty$ . (2) It allows for general loss functions. The loss function  $L(\cdot)$  need not be quadratic or even symmetric; we only require that  $L(0) = 0$  and that  $L(\cdot)$  be strictly monotone on each side of the origin. By the restrictions imposed on  $L(\cdot)$  , we have that for all covariance stationary or difference stationary processes  $P(L(\cdot), \Omega, j, k) \in [0, 1]$ , with larger values indicating greater predictability. (3) It allows for univariate or multivariate information sets, and economic theory may suggest relevant multivariate information sets. (4) It allows for flexibility in the choice of  $j$  and  $k$  and enables one to tailor the predictability measure to the horizons of economic interest.

Our predictability measure is closely related to Theil's  $U$  statistic, which we define for the 1-step-ahead horizon as

$$U = \frac{E(e_{t,t-1}^2)}{E((y_t - y_{t-1})^2)}.$$

To see this, specialize  $P$  to the quadratic, univariate,  $j = 1$  case and write it

as

$$P(\text{quadratic, univariate}, 1, k) = 1 - \frac{E(e_{t,t-1}^2)}{E(e_{t,t-k}^2)},$$

or

$$1 - P = \frac{E(e_{t,t-1}^2)}{E(e_{t,t-k}^2)}.$$

Thus, under certain conditions,  $1 - P$  is similar in spirit to Theil's  $U$ . The key difference is that Theil's  $U$  assesses 1-step forecast accuracy relative to that of a "naive" no-change forecast, whereas  $P$  assesses 1-step accuracy relative to that of a long-horizon ( $k$ -step) forecast. In the general case,

$$P(L(\cdot), \Omega, j, k) = 1 - \frac{E(L(e_{t,t-j}))}{E(L(e_{t,t-k}))}.$$

Thus,  $P(L(\cdot), \Omega, j, k)$  is effectively one minus the ratio of expected losses of two forecasts of the same object,  $y_t$ . Typically, one forecast,  $\hat{y}_{t,t-j}$ , is based on a rich information set, while the other forecast,  $\hat{y}_{t,t-k}$ , is based on a sparse information set.

The formula for  $P(L(\cdot), \Omega, j, k)$  also makes clear that the concept of predictability is related to, but distinct from, the concept of persistence of a series. Suppose, for example, that the series  $y_t$  is a random walk. Then

$$P(e^2, \text{univariate}, j, k) = 1 - \frac{j}{k},$$

as will be shown later. The corresponding  $j$ -step variance ratio, a common persistence measure, is

$$V_j = \frac{\text{var}(y_t - y_{t-j})}{\text{var}(y_t - y_{t-1})} = j.$$

It is clear, however, that although  $P(e^2, \text{univariate}, j, k)$  and  $V_j$  are deterministically related in the random walk case ( $P = 1 - V/k$ ), they are not deterministically related in more general cases.

### Sample Measures

Predictability is a population property of a series, not of any particular sample path, but predictability can be estimated from a sample path. We proceed by fitting a parametric model and then transforming estimates of the parameters into an estimate of  $P$ . To keep the discussion tractable, and in keeping with the empirical analysis of subsequent sections, we postulate a quadratic loss function  $L(e) = e^2$  for estimation, prediction, model selection, and construction of predictability measures.

It is clear that parametric measures of predictability in general will depend on the specification of the parametric model. Here we focus on univariate autoregressive models, although one could easily generalize the discussion to other parametric models, such as vector *ARMA* models. We construct  $P$  by simply reading off the appropriate diagonal elements of the forecast *MSE* matrices for forecast horizons  $j$  and  $k$ . To build intuition, consider a univariate *AR*(1) population process with innovation variance  $\Sigma_u$ :  $y_t = A_1 y_{t-1} + u_t$ . Then for  $A_1 = 0$  the model reduces to white noise, and short-run forecasts are just as accurate as long-run forecasts. As a result, relative predictability is zero:  $P(j, k) = 1 - \Sigma_u / \Sigma_u = 0$ , for all  $j$ . In contrast, for  $A_1 = 1$  the model becomes a random walk, and relative predictability steadily declines as the forecast horizon increases:  $P(j, k) = 1 - (j\Sigma_u)/(k\Sigma_u) = 1 - j/k$ .

Forecast errors from consistently estimated processes and processes with known parameters are asymptotically equivalent. In practice, we estimate  $P$  by replacing the underlying unknown parameters by their least squares estimates.

### 10.2.5 Statistical Assessment of Accuracy Rankings

Once we've decided on a loss function, it is often of interest to know whether one forecast is more accurate than another. In hypothesis testing terms, we might want to test the equal accuracy hypothesis,

$$E[L(e_{t+h,t}^a)] = E[L(e_{t+h,t}^b)],$$

against the alternative hypothesis that one or the other is better. Equivalently, we might want to test the hypothesis that the expected loss differential is zero,

$$E(d_t) = E[L(e_{t+h,t}^a)] - E[L(e_{t+h,t}^b)] = 0.$$

The hypothesis concerns population expected loss; we test it using sample average loss.

#### A Motivational Example

Consider a model-free forecasting environment, as for example with forecasts based on surveys, forecasts extracted from financial markets, forecasts obtained from prediction markets, or forecasts based on expert judgment. One routinely has competing model-free forecasts of the same object, gleaned for example from surveys or financial markets, and seeks to determine which is better.

To take a concrete example, consider U.S. inflation forecasting. One might obtain survey-based forecasts from the Survey of Professional Forecasters ( $S$ ),  $\{\pi_t^S\}_{t=1}^T$ , and simultaneously one might obtain market-based forecasts from inflation-indexed bonds ( $B$ ),  $\{\pi_t^B\}_{t=1}^T$ . Suppose that loss is quadratic and that during  $t = 1, \dots, T$  the sample mean-squared errors are  $\widehat{MSE}(\pi_t^S) = 1.80$  and  $\widehat{MSE}(\pi_t^B) = 1.92$ . Evidently “ $S$  wins,” and one is tempted to conclude that  $S$  provides better inflation forecasts than does  $B$ . The forecasting literature is filled with such horse races, with associated declarations of superiority based

on outcomes.

Obviously, however, the fact that  $\widehat{MSE}(\pi_t^S) < \widehat{MSE}(\pi_t^B)$  in a particular sample realization does not mean that  $S$  is necessarily truly better than  $B$  in population. That is, even if in population  $MSE(\pi_t^S) = MSE(\pi_t^B)$ , in any particular sample realization  $t = 1, \dots, T$  one or the other of  $S$  and  $B$  must “win,” so the question arises in any particular sample as to whether  $S$  is truly superior or merely lucky. The Diebold-Mariano test answers that question, allowing one to assess the significance of apparent predictive superiority. It provides a test of the hypothesis of equal expected loss (in our example,  $MSE(\pi_t^S) = MSE(\pi_t^B)$ ), valid under quite general conditions including, for example, wide classes of loss functions and forecast-error serial correlation of unknown form.

### The Diebold-Mariano Perspective

The essence of the *DM* approach is to take forecast errors as primitives, intentionally, and to make assumptions directly on those forecast errors. (In a model-free environment there are obviously no models about which to make assumptions.) More precisely, *DM* relies on assumptions made directly on the forecast error *loss differential*. Denote the loss associated with forecast error  $e_t$  by  $L(e_t)$ ; hence, for example, time- $t$  quadratic loss would be  $L(e_t) = e_t^2$ . The time- $t$  loss differential between forecasts 1 and 2 is then  $d_{12t} = L(e_{1t}) - L(e_{2t})$ . *DM* requires only that the loss differential be covariance stationary.<sup>2</sup> That is, *DM* assumes that:

$$Assumption\ DM : \begin{cases} E(d_{12t}) = \mu, \forall t \\ cov(d_{12t}, d_{12(t-\tau)}) = \gamma(\tau), \forall t \\ 0 < var(d_{12t}) = \sigma^2 < \infty. \end{cases} \quad (10.2)$$

---

<sup>2</sup>Actually covariance stationarity is sufficient but may not be strictly necessary, as less-restrictive types of mixing conditions could presumably be invoked.

The key hypothesis of equal predictive accuracy (i.e., equal expected loss) corresponds to  $E(d_{12t}) = 0$ , in which case, under the maintained Assumption *DM*:

$$DM_{12} = \frac{\bar{d}_{12}}{\hat{\sigma}_{\bar{d}_{12}}} \xrightarrow{d} N(0, 1), \quad (10.3)$$

where  $\bar{d}_{12} = \frac{1}{T} \sum_{t=1}^T d_{12t}$  is the sample mean loss differential and  $\hat{\sigma}_{\bar{d}_{12}}$  is a consistent estimate of the standard deviation of  $\bar{d}_{12}$  (more on that shortly). That's all: If Assumption *DM* holds, then the  $N(0, 1)$  limiting distribution of test statistic *DM* *must* hold.

*DM* is simply an asymptotic *z*-test of the hypothesis that the mean of a constructed but observed series (the loss differential) is zero. The only wrinkle is that forecast errors, and hence loss differentials, may be serially correlated for a variety of reasons, the most obvious being forecast sub-optimality. Hence the standard error in the denominator of the *DM* statistic (10.3) should be calculated robustly. A simple approach is to recognize that *DM* is just a *t*-statistic for the hypothesis of a zero population mean loss differential, adjusted to reflect the fact that the loss differential series is not necessarily, so that we can compute it via HAC regression (e.g., Newey-West or Kiefer-Vogelsang) on an intercept. Perhaps an even simpler approach is to regress the loss differential on an intercept, allowing for  $AR(p)$  disturbances, and using information criterion like *AIC* to select *p*.

*DM* is also readily extensible. The key is to recognize that the *DM* statistic can be trivially calculated by regression of the loss differential on an intercept, using heteroskedasticity and autocorrelation robust (HAC) standard errors. Immediately, then (and as noted in the original Diebold-Mariano paper), one can potentially extend the regression to condition on additional variables that may explain the loss differential, thereby moving from an un-

conditional to a conditional expected loss perspective.<sup>3</sup> For example, comparative predictive performance may differ by stage of the business cycle, in which case one might include a 0-1 NBER business cycle chronology variable (say) in the *DM* HAC regression.

### Thoughts on Assumption *DM*

Thus far I have praised *DM* rather effusively, and its great simplicity and wide applicability certainly *are* virtues: There is just one Assumption *DM*, just one *DM* test statistic, and just one *DM* limiting distribution, always and everywhere. But of course everything hinges on Assumption *DM*. Here I offer some perspectives on the validity of Assumption *DM*.

First, as George Box (1979) famously and correctly noted, “All models are false, but some are useful.” Precisely the same is true of *assumptions*. Indeed all areas of economics benefit from assumptions that are surely false if taken literally, but that are nevertheless useful. So too with Assumption *DM*. Surely  $d_t$  is likely never *precisely* covariance stationary, just as surely *no* economic time series is likely precisely covariance stationary. But in many cases Assumption *DM* may be a useful approximation.

Second, special forecasting considerations lend support to the validity of Assumption *DM*. Forecasters strive to achieve forecast optimality, which corresponds to unforecastable covariance-stationary errors (indeed white-noise errors in the canonical 1-step-ahead case), and hence unforecastable covariance-stationary loss differentials. Of course forecasters may not achieve optimality, resulting in serially-correlated, and indeed forecastable, forecast errors. But  $I(1)$  non-stationarity of forecast errors takes serial correlation to the extreme.<sup>4</sup>

Third, even in the extreme case where nonstationary components somehow *do* exist in forecast errors, there is reason to suspect that they may be shared.

---

<sup>3</sup>Important subsequent work takes the conditional perspective farther; see Giacomini and White (2006).

<sup>4</sup>Even with apparent nonstationarity due to apparent breaks in the loss differential series, Assumption *DM* may nevertheless hold if the breaks have a stationary rhythm, as for example with hidden-Markov processes in the tradition of Hamilton (1989).

In particular, information sets overlap across forecasters, so that forecast-error nonstationarities may vanish from the loss differential. For example, two loss series, each integrated of order one, may nevertheless be cointegrated with cointegrating vector  $(1, -1)$ . Suppose for example that  $L(e_{1t}) = x_t + \varepsilon_{1t}$  and  $L(e_{2t}) = x_t + \varepsilon_{2t}$ , where  $x_t$  is a common nonstationary  $I(1)$  loss component, and  $\varepsilon_{1t}$  and  $\varepsilon_{2t}$  are idiosyncratic stationary  $I(0)$  loss components. Then  $d_{12t} = L(e_{1t}) - L(e_{2t}) = \varepsilon_{1t} - \varepsilon_{2t}$  is  $I(0)$ , so that the loss differential series is covariance stationary despite the fact that neither individual loss series is covariance stationary.

Fourth, and most importantly, standard and powerful tools enable empirical assessment of Assumption *DM*. That is, the approximate validity of Assumption *DM* is ultimately an empirical matter, and a wealth of diagnostic procedures are available to help assess its validity. One can plot the loss differential series, examine its sample autocorrelations and spectrum, test it for unit roots and other nonstationarities including trend, structural breaks or evolution, and so on.

## 10.3 OverSea Shipping

We'll work with an application to OverSea Services, Inc., a major international cargo shipper. To help guide fleet allocation decisions, each week OverSea makes forecasts of volume shipped over each of its major trade lanes, at horizons ranging from 1-week ahead through 16-weeks-ahead. In fact, OverSea produces two sets of forecasts – a quantitative forecast is produced using modern quantitative techniques, and a judgmental forecast is produced by soliciting the opinion of the sales representatives, many of whom have years of valuable experience.

Here we'll examine the realizations and 2-week-ahead forecasts of volume on the Atlantic East trade lane (North America to Europe). We have

nearly ten years of data on weekly realized volume ( $VOL$ ) and weekly 2-week-ahead forecasts (the quantitative forecast  $VOLQ$ , and the judgmental forecast  $VOLJ$ ), from January 1988 through mid-July 1997, for a total of 499 weeks.

In Figure 1, we plot realized volume vs. the quantitative forecast, and in Figure 2 we show realized volume vs. the judgmental forecast. The two plots look similar, and both forecasts appear quite accurate; it's not too hard to forecast shipping volume just two weeks ahead.

In Figures 3 and 4, we plot the errors from the quantitative and judgmental forecasts, which are more revealing. The quantitative error, in particular, appears roughly centered on zero, whereas the judgmental error seems to be a bit higher than zero on average. That is, the judgmental forecast appears biased in a pessimistic way – on average, actual realized volume is a bit higher than forecasted volume.

In Figures 5 and 6, we show histograms and related statistics for the quantitative and judgmental forecast errors. The histograms confirm our earlier suspicions based on the error plots; the histogram for the quantitative error is centered on a mean of -.03, whereas that for the judgmental error is centered on 1.02. The error standard deviations, however, reveal that the judgmental forecast errors vary a bit less around their mean than do the quantitative errors. Finally, the Jarque-Bera test can't reject the hypothesis that the errors are normally distributed.

In Tables 1 and 2 and Figures 7 and 8, we show the correlograms of the quantitative and judgmental forecast errors. In each case, the errors appear to have  $MA(1)$  structure; the sample autocorrelations cut off at displacement 1, whereas the sample partial autocorrelations display damped oscillation, which is reasonable for 2-step-ahead forecast errors.

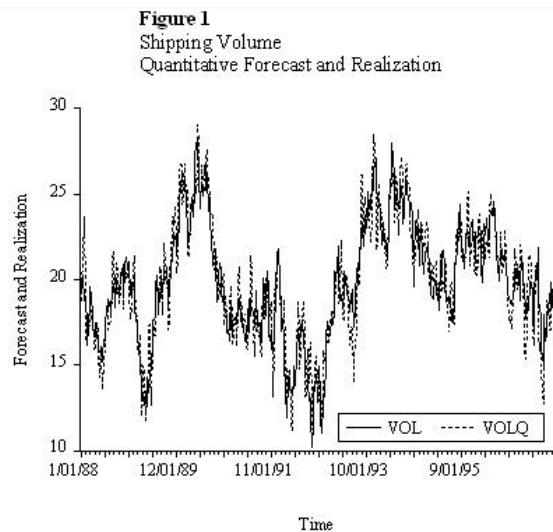
To test for the statistical significance of bias, we need to account for the  $MA(1)$  serial correlation. To do so, we regress the forecast errors on a con-

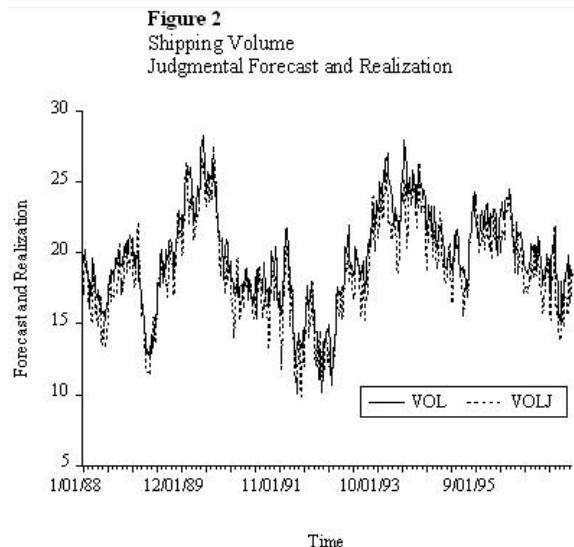
stant, allowing for  $MA(1)$  disturbances. We show the results for the quantitative forecast errors in Table 3, and those for the judgmental forecast errors in Table 4. The t-statistic indicates no bias in the quantitative forecasts, but sizable and highly statistically significant bias in the judgmental forecasts.

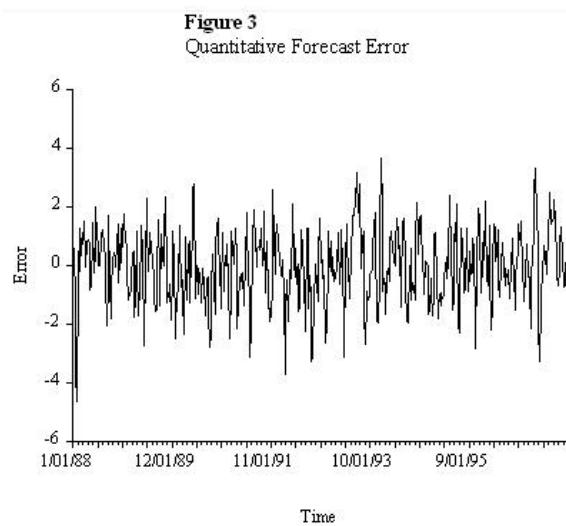
In Tables 5 and 6, we show the results of Mincer-Zarnowitz regressions; both forecasts fail miserably. We expected the judgmental forecast to fail, because it's biased, but until now no defects were found in the quantitative forecast.

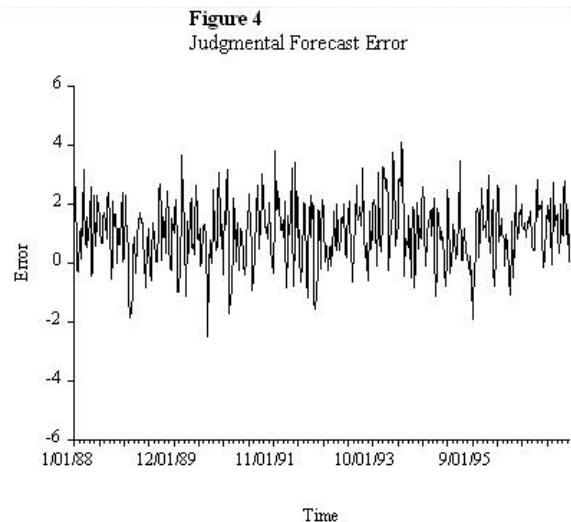
Now let's compare forecast accuracy. We show the histogram and descriptive statistics for the squared quantitative and judgmental errors in Figures 9 and 10. The histogram for the squared judgmental error is pushed rightward relative to that of the quantitative error, due to bias. The  $RMSE$  of the quantitative forecast is 1.26, while that of the judgmental forecast is 1.48.

In Figure 11 we show the (quadratic) loss differential; it's fairly small but looks a little negative. In Figure 12 we show the histogram of the loss differential; the mean is -.58, which is small relative to the standard deviation of the loss differential, but remember that we have not yet corrected for serial correlation. In Table 7 we show the correlogram of the loss differential, which strongly suggests  $MA(1)$  structure. The sample autocorrelations and partial autocorrelations, shown in Figure 13, confirm that impression. Thus, to test for significance of the loss differential, we regress it on a constant and allow for  $MA(1)$  disturbances; we show the results in Table 8. The mean loss differential is highly statistically significant, with a  $p$ -value less than .01; we conclude that the quantitative forecast is more accurate than the judgmental forecast under quadratic loss.

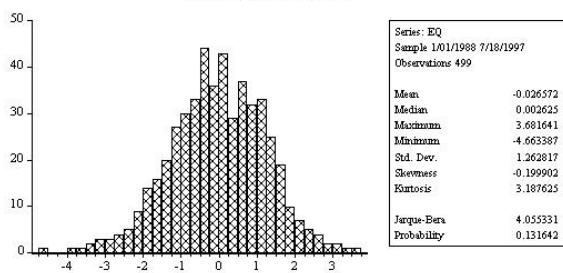




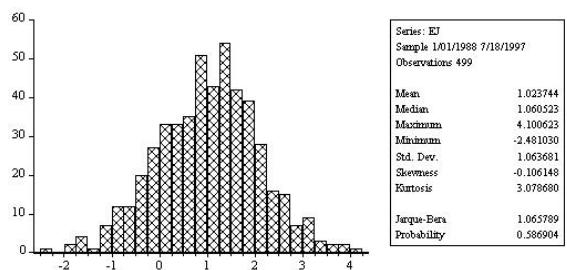




**Figure 5**  
Histogram and Related Statistics  
Quantitative Forecast Error



**Figure 6**  
Histogram and Related Statistics  
Judgmental Forecast Error

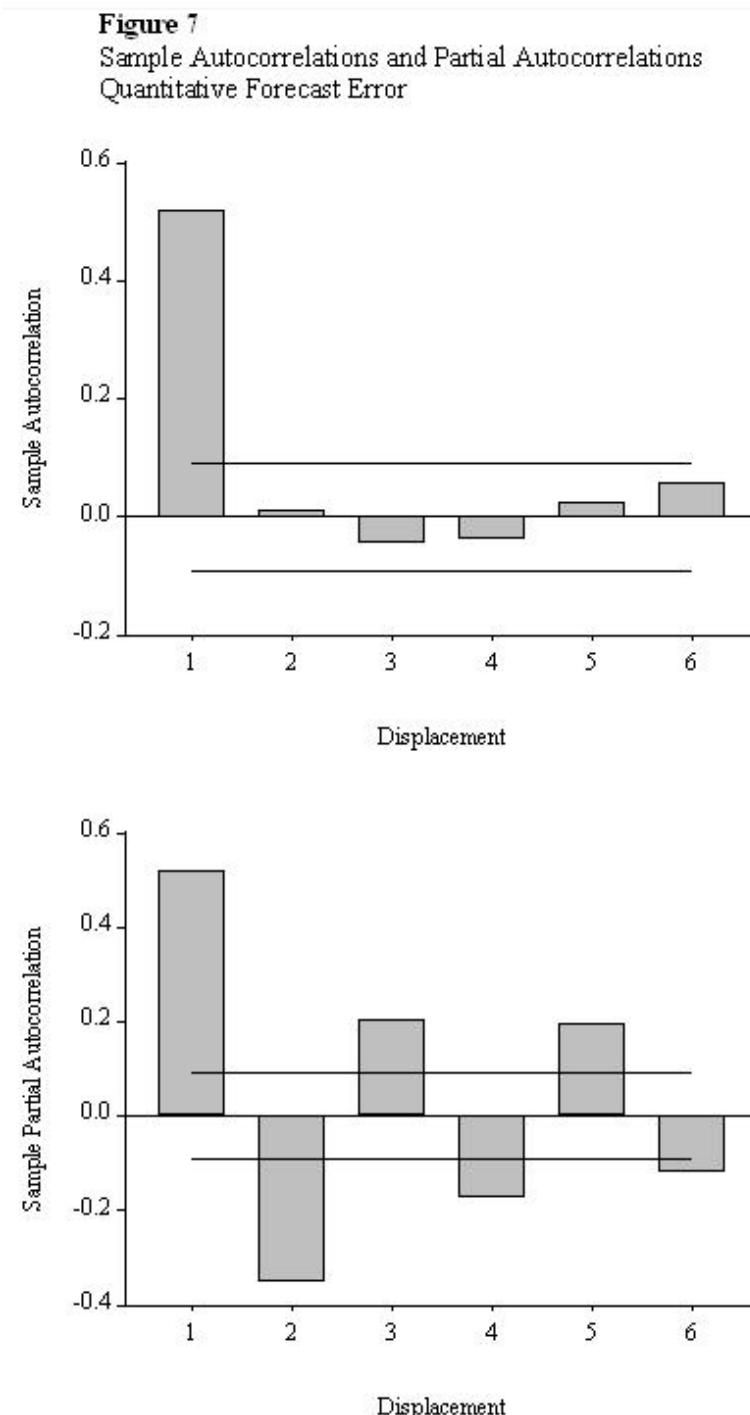


**Table 1**  
Correlogram, Quantitative Forecast Error

Sample: 1/01/1988 7/18/1997

Included observations: 499

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	0.518	0.518	.045	134.62	0.000
2	0.010	-0.353	.045	134.67	0.000
3	-0.044	0.205	.045	135.65	0.000
4	-0.039	-0.172	.045	136.40	0.000
5	0.025	0.195	.045	136.73	0.000
6	0.057	-0.117	.045	138.36	0.000



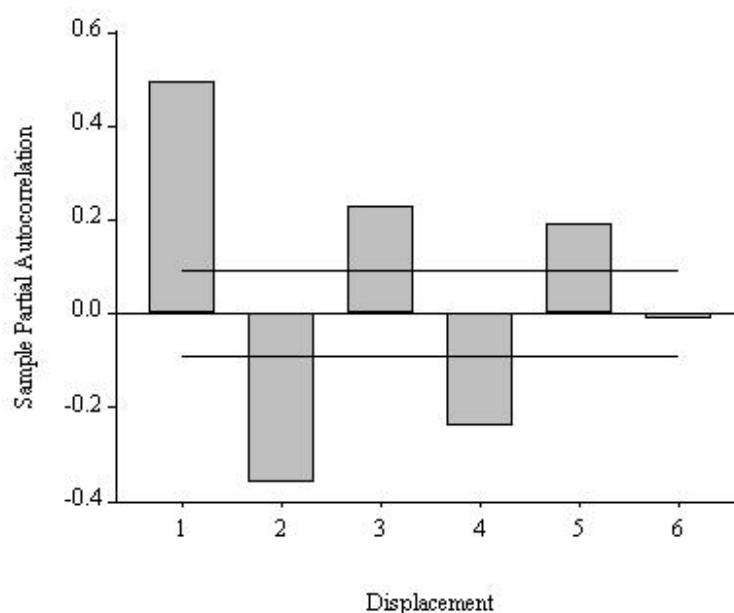
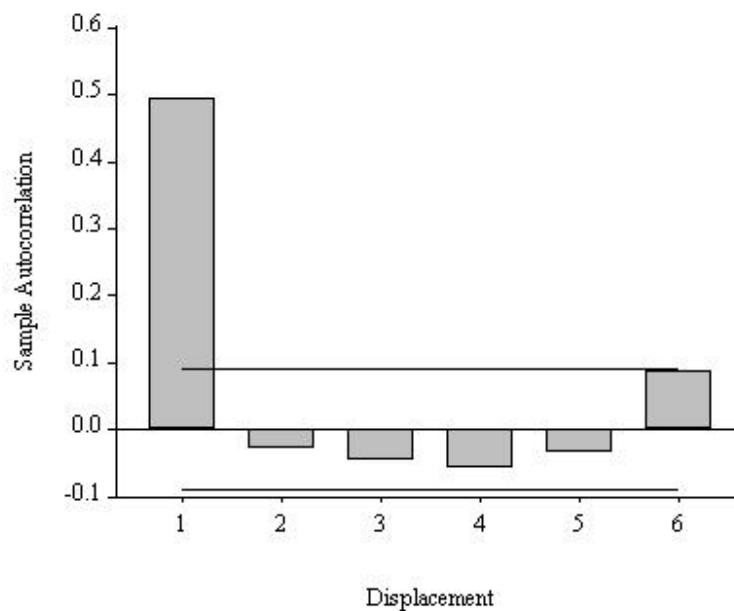
**Table 2**  
Correlogram, Judgmental Forecast Error

Sample: 1/01/1988 7/18/1997

Included observations: 499

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	0.495	0.495	.045	122.90	0.000
2	-0.027	-0.360	.045	123.26	0.000
3	-0.045	0.229	.045	124.30	0.000
4	-0.056	-0.238	.045	125.87	0.000
5	-0.033	0.191	.045	126.41	0.000
6	0.087	-0.011	.045	130.22	0.000

**Figure 8**  
Sample Autocorrelations and Partial Autocorrelations  
Judgmental Forecast Error



**Table 3**  
**Quantitative Forecast Error**  
 Regression on Intercept, MA(1) Disturbances

LS // Dependent Variable is EQ  
Sample: 1/01/1988 7/18/1997  
Included observations: 499  
Convergence achieved after 6 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.024770	0.079851	-0.310200	0.7565
MA(1)	0.935393	0.015850	59.01554	0.0000
R-squared	0.468347		Mean dependent var	-0.026572
Adjusted R-squared	0.467277		S.D. dependent var	1.262817
S.E. of regression	0.921703		Akaike info criterion	-0.159064
Sum squared resid	422.2198		Schwarz criterion	-0.142180
Log likelihood	-666.3639		F-statistic	437.8201
Durbin-Watson stat	1.988237		Prob(F-statistic)	0.000000
Inverted MA Roots	.94			

**Table 4**  
 Judgmental Forecast Error  
 Regression on Intercept, MA(1) Disturbances

LS // Dependent Variable is EJ

Sample: 1/01/1988 7/18/1997

Included observations: 499

Convergence achieved after 7 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.026372	0.067191	15.27535	0.0000
MA(1)	0.961524	0.012470	77.10450	0.0000
R-squared	0.483514	Mean dependent var		1.023744
Adjusted R-squared	0.482475	S.D. dependent var		1.063681
S.E. of regression	0.765204	Akaike info criterion		-0.531226
Sum squared resid	291.0118	Schwarz criterion		-0.514342
Log likelihood-573.5094		F-statistic	465.2721	
Durbin-Watson stat	1.968750	Prob(F-statistic)		0.000000
Inverted MA Roots				-.96

**Table 5**  
 Mincer-Zarnowitz Regression  
 Quantitative Forecast Error

LS // Dependent Variable is VOL

Sample: 1/01/1988 7/18/1997

Included observations: 499

Convergence achieved after 10 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.958191	0.341841	8.653696	0.0000
VOLQ	0.849559	0.016839	50.45317	0.0000
MA(1)	0.912559	0.018638	48.96181	0.0000
R-squared	0.936972		Mean dependent var	19.80609
Adjusted R-squared	0.936718		S.D. dependent var	3.403283
S.E. of regression	0.856125		Akaike info criterion	-0.304685
Sum squared resid	363.5429		Schwarz criterion	-0.279358
Log likelihood	-629.0315		F-statistic	3686.790
Durbin-Watson stat	1.815577		Prob(F-statistic)	0.000000
Inverted MA Roots				
				.91
Wald Test:				
Null Hypothesis:	C(1)=0C(2)=1			
F-statistic	39.96862		Probability	0.000000
Chi-square	79.93723		Probability	0.000000

**Table 6**  
 Mincer-Zarnowitz Regression  
 Judgmental Forecast Error

LS // Dependent Variable is VOL

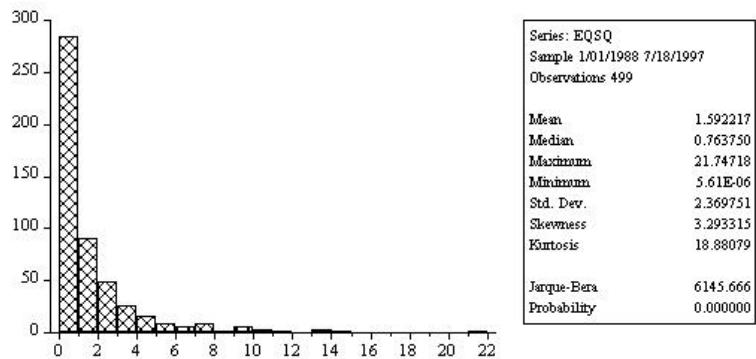
Sample: 1/01/1988 7/18/1997

Included observations: 499

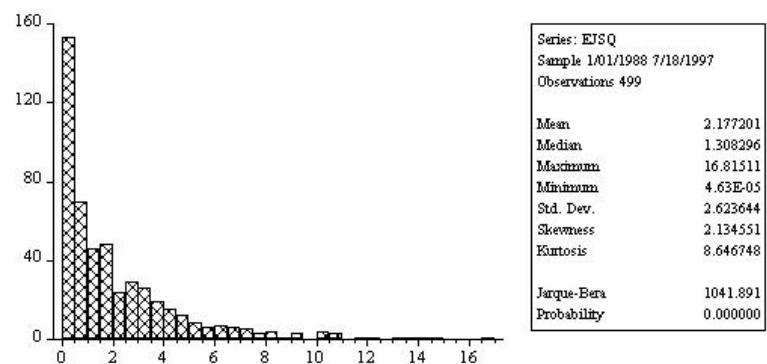
Convergence achieved after 11 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.592648	0.271740	9.540928	0.0000
VOLJ	0.916576	0.014058	65.20021	0.0000
MA(1)	0.949690	0.014621	64.95242	0.0000
R-squared	0.952896		Mean dependent var	19.80609
Adjusted R-squared	0.952706		S.D. dependent var	3.403283
S.E. of regression	0.740114		Akaike info criterion	-0.595907
Sum squared resid	271.6936		Schwarz criterion	-0.570581
Log likelihood	-556.3715		F-statistic	5016.993
Durbin-Watson stat	1.917179		Prob(F-statistic)	0.000000
Inverted MA Roots				
Wald Test:				
Null Hypothesis:	C(1)=0C(2)=1			
F-statistic	143.8323		Probability	0.000000
Chi-square	287.6647		Probability	0.000000

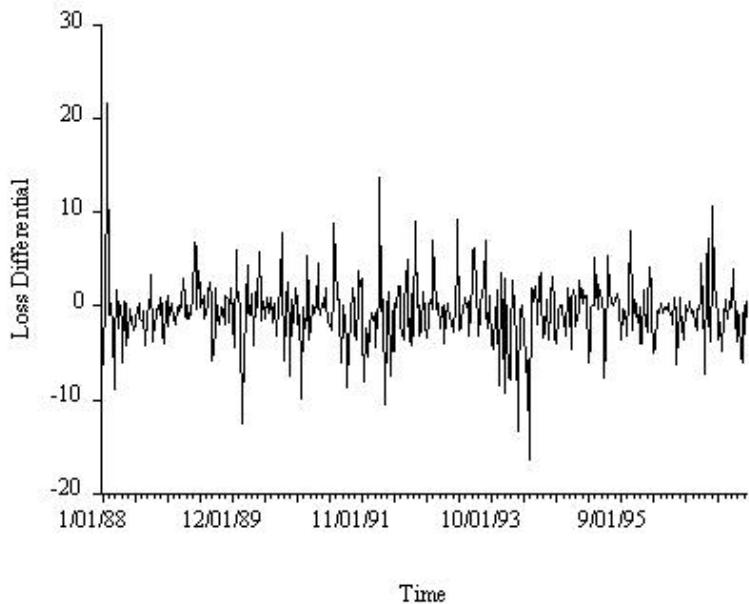
**Figure 9**  
Histogram and Related Statistics  
Squared Quantitative Forecast Error



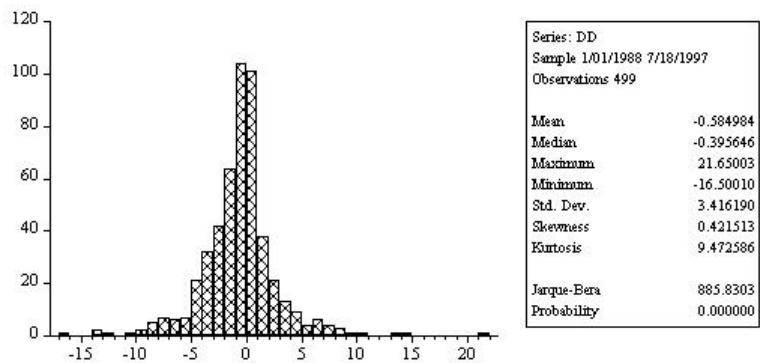
**Figure 10**  
Histogram and Related Statistics  
Squared Judgmental Forecast Error



**Figure 11**  
Loss Differential



**Figure 12**  
Histogram and Related Statistics  
Loss Differential



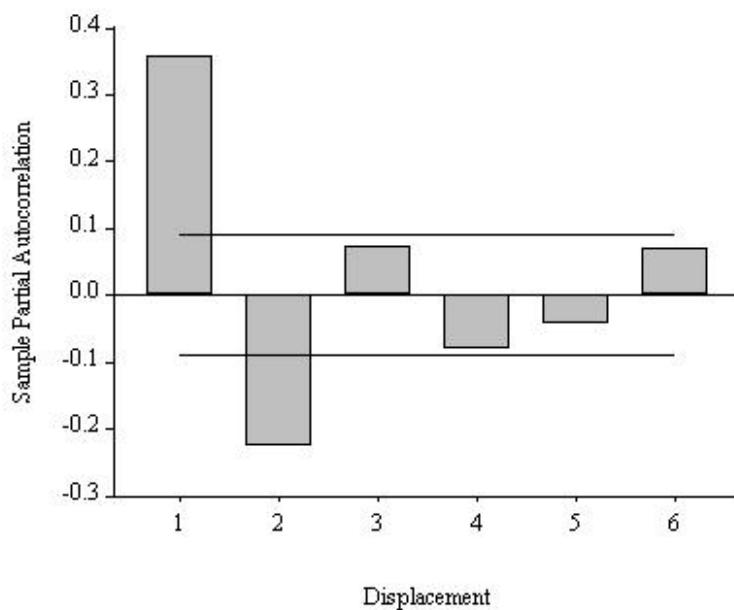
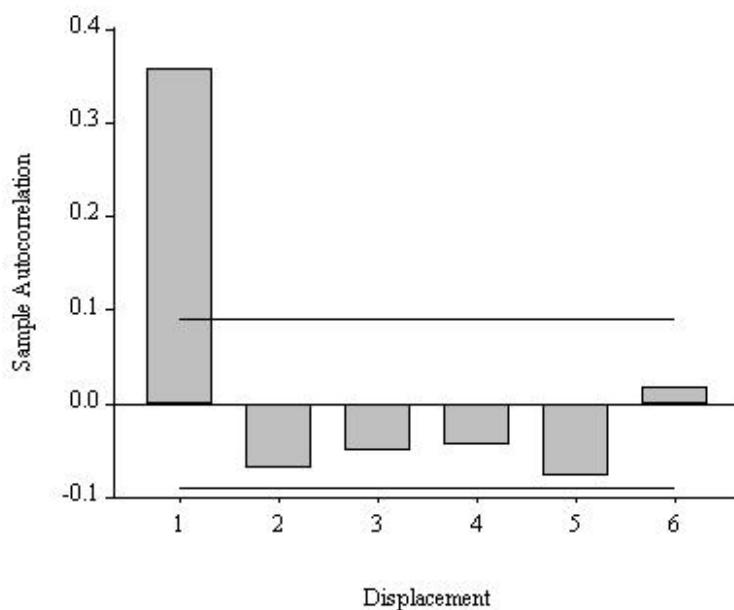
**Table 7**  
Loss Differential Correlogram

Sample: 1/01/1988 7/18/1997

Included observations: 499

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	0.357	0.357	.045	64.113	0.000
2	-0.069	-0.226	.045	66.519	0.000
3	-0.050	0.074	.045	67.761	0.000
4	-0.044	-0.080	.045	68.746	0.000
5	-0.078	-0.043	.045	71.840	0.000
6	0.017	0.070	.045	71.989	0.000

**Figure 13**  
Sample Autocorrelations and Partial Autocorrelations  
Loss Differential



**Table 8**

Loss Differential

Regression on Intercept with MA(1) Disturbances

LS // Dependent Variable is DD

Sample: 1/01/1988 7/18/1997

Included observations: 499

Convergence achieved after 4 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.585333	0.204737	-2.858945	0.0044
MA(1)	0.472901	0.039526	11.96433	0.0000
R-squared	0.174750	Mean dependent var		-0.584984
Adjusted R-squared	0.173089	S.D. dependent var		3.416190
S.E. of regression	3.106500	Akaike info criterion		2.270994
Sum squared resid	4796.222	Schwarz criterion		2.287878
Log likelihood-1272.663		F-statistic	105.2414	
Durbin-Watson stat	2.023606	Prob(F-statistic)		0.000000
Inverted MA Roots				.47

## 10.4 Exercises, Problems and Complements

### 1. Forecast evaluation in action.

Discuss in detail how you would use forecast evaluation techniques to address each of the following questions.

- a. Are asset returns (e.g., stocks, bonds, exchange rates) forecastable over long horizons?
- b. Do forward exchange rates provide unbiased forecasts of future spot exchange rates at all horizons?
- c. Are government budget projections systematically too optimistic, perhaps for strategic reasons?
- d. Can interest rates be used to provide good forecasts of future inflation?

### 2. Forecast error analysis.

You work for a London-based hedge fund, Thompson Energy Investors, and your boss has assigned you to assess a model used to forecast U.S. crude oil imports. On the last day of each quarter, the model is used to forecast oil imports at horizons of 1-quarter-ahead through 4-quarters-ahead. Thompson has done this for each of 80 quarters and has kept the corresponding four forecast error series, which appear on the book's web page.

- a. Based on a correlogram analysis, assess whether the 1-quarter-ahead forecast errors are white noise. (Be sure to discuss all parts of the correlogram: sample autocorrelations, sample partial autocorrelations, Bartlett standard errors and Ljung-Box statistics.) Why care?
- b. Regress each of the four forecast error series on constants, in each case allowing for a  $MA(5)$  disturbances. Comment on the significance of

the  $MA$  coefficients in each of the four cases and use the results to assess the optimality of the forecasts at each of the four horizons. Does your 1-step-ahead  $MA(5)$ -based assessment match the correlogram-based assessment obtained in part a? Do the multi-step forecasts appear optimal?

- c. Overall, what do your results suggest about the model's ability to predict U.S. crude oil imports?
- 3. The mechanics of practical forecast evaluation.

For the following, use the time series of shipping volume, quantitative forecasts, and judgmental forecasts used in this chapter.

- a. Replicate the empirical results reported in this chapter. Explore and discuss any variations or extensions that you find interesting.
- b. Using the first 250 weeks of shipping volume data, specify and estimate a univariate autoregressive model of shipping volume (with trend and seasonality if necessary), and provide evidence to support the adequacy of your chosen specification.
- c. Use your model each week to forecast two weeks ahead, each week estimating the model using all available data, producing forecasts for observations 252 through 499, made using information available at times 250 through 497. Calculate the corresponding series of 248 2-step-ahead recursive forecast errors.
- d. Using the methods of this chapter, evaluate the quality of your forecasts, both in isolation and relative to the original quantitative and judgmental forecasts. Discuss.
- 4. Forecasting Competitions.

There are many forecasting competitions. Kaggle.com, for example, is a well-known online venue. Participants are given a “training sample” of

data and asked to forecast a “test sample”; that is, to make an out-of-sample forecast of hold-out data, which they are not shown

- (a) Check out Kaggle. Also read “A Site for Data Scientists to Prove Their Skills and Make Money,” by Claire Cain Miller, *New York Times*, November 3, 2011. What’s good about the Kaggle approach? What’s bad? What happened to Kaggle since its launch in 2011?
- (b) “Kaggle competitions” effectively outsource forecasting. What are pros and cons of in-house experts vs. outsourcing?
- (c) Kaggle strangely lets people peek at the test sample by re-submitting forecasts once per day.
- (d) Kaggle scores extrapolation forecasts rather than h-step. This blends apples and oranges.
- (e) Kaggle is wasteful from a combining viewpoint. One doesn’t just want to find the “winner.”

## 5. The Peso Problem.

Suppose someone assigns a very high probability to an event that fails to occur, or a very low probability to an event that does occur. Is the person a bad probability forecaster? The answer is perhaps, but not at all necessarily. Even events *correctly* forecast to occur with high probability may simply fail to occur, and conversely.

Thus, for example, a currency might sell forward at a large discount, indicating that the market has assigned a high probability of a large depreciation. In the event, that depreciation might fail to occur, but that does not necessarily mean that the market was in any sense “wrong” in assigning a high depreciation probability. The term “**Peso problem**” refers to exactly such issues in a long-ago situation involving the Mexican Peso.

## 6. Measuring forecastability with canonical correlations.

One can measure forecastability via canonical correlation between “past” and “future,” as in Jewell and Bloomfield 1983, Hannan and Poskitt 1988.

## 7. Forecast Evaluation When Realizations are Unobserved.

Sometimes we never see the realization of the variable being forecast. This occurs for example in forecasting ultimate resource recovery, such as the total amount of oil in an underground reserve. The actual value, however, won’t be known until the reserve is depleted, which may be decades away. Such situations obviously make for difficult accuracy evaluation!

If the resource recovery example sounds a bit exotic, rest assured that it’s not. In volatility forecasting, for example, “true” volatility is never observed. And in any sort of state-space model, such as a dynamic factor model, the true state vector is never observed. (See Chapters \*\*\*.)

### (a) Nordhaus tests.

– Some optimality tests can be obtained even when the forecast target is unobservable (Patton and Timmermann 2010, building on Nordhaus 1987). In particular, (1) forecast revisions (for fixed target date) should be MDS, and (2) forecast variance (not error variance, but forecast variance) should decrease with distance from terminal date. PT 2010 also have a nice generalized MZ test that builds on these ideas.

### (b) Patton tests.

## 8. Nonparametric predictability assessment.

We presented our autoregressive modeling approach as a parametric method. However, in general, we need not assume that the fitted autore-

gression is the true data-generating process; rather, it may be considered an approximation, the order of which can grow with sample size. Thus the autoregressive model can be viewed as a sieve, so our approach actually *is* nonparametric.

Nevertheless, the sieve approach has a parametric flavor. For any fixed sample size, we assess predictability through the lens of a particular autoregressive model. Hence it may be of interest to develop an approach with a more thoroughly nonparametric flavor by exploiting Kolmogorov's well-known spectral formula for the univariate innovation variance,

$$\sigma^2 = \exp\left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \ln 2\pi f(\omega) d\omega\right),$$

where  $f$  is the spectral density function. Kolmogorov's result has been extended to univariate  $h$ -step-ahead forecast error variances by Bhansali (1992).

9. Can unskilled density forecasters successfully disguise themselves as skilled?
10. Cross section forecast evaluation.

Most of the basic lessons for time-series forecast evaluation introduced in this chapter are also relevant for cross-section forecast evaluation. Cross-section forecast errors (appropriately standardized if heteroskedasticity is present) should be iid white noise over space, and unpredictable using any available covariates. DM-type tests can be done for point forecasts, and DGT-type test for density forecasts.

11. Turning point forecasts into density forecasts.

As we have shown, Mincer-Zarnowitz corrections can be used to “correct” sub-optimal point forecasts. They can also be used to produce density forecasts, by drawing from an estimate of the density of the MZ regression disturbances, as we did in a different context in section 4.1.

## **10.5 Notes**

# Chapter 11

## Interval and Density Forecast Evaluation

### 11.1 Interval Forecast Evaluation

Interval forecast evaluation is largely, but not entirely, subsumed by density forecast evaluation. There is a simple method for absolute interval forecast evaluation that must be mentioned. It is of great practical use, and moreover establishes the proper notion of a 1-step-ahead interval forecast error (which should be unforecastable), and which then translates into the proper notion of a 1-step-ahead density forecast error (which should also be unforecastable).

#### 11.1.1 Absolute Standards

##### On Correct Unconditional vs. Conditional Coverage

A  $(1 - \alpha)\%$  interval is correctly *unconditionally* calibrated if it brackets the truth  $(1 - \alpha)\%$  of the time, on average over the long run. But an interval can be correctly unconditionally calibrated and still poorly *conditionally* calibrated insofar as it's poorly calibrated at any given time, despite being correct on average. In environments of time-varying conditional variance, for example, constant-width intervals may be correctly unconditionally calibrated, but they cannot be correctly conditionally calibrated, because they

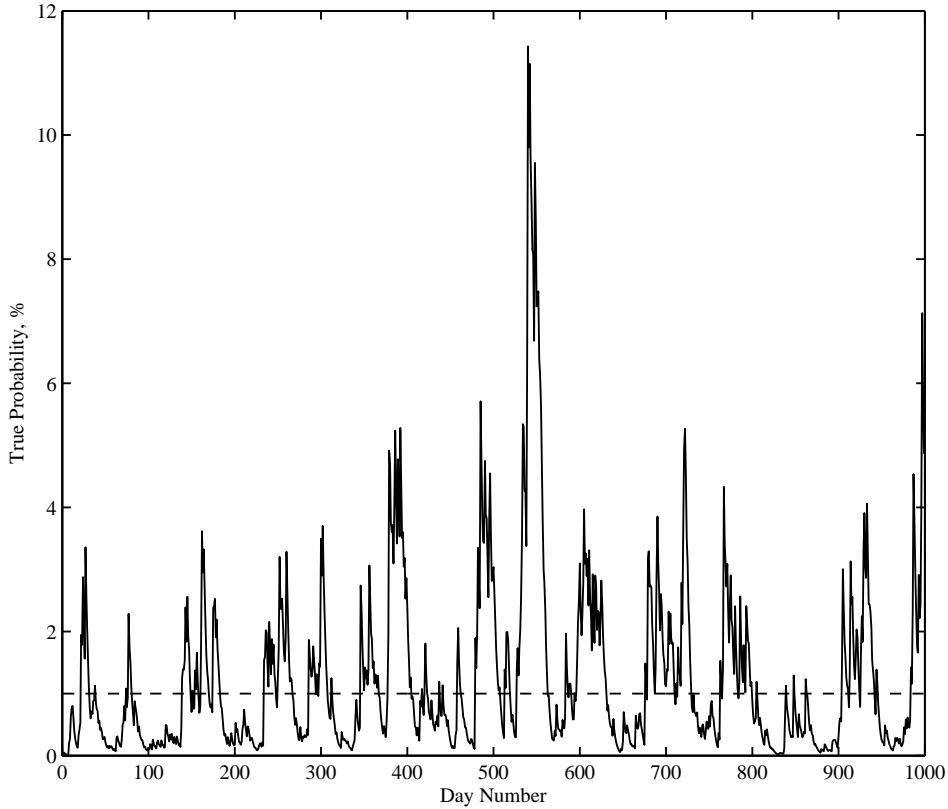


Figure 11.1: True Exceedance Probabilities of Nominal one-sided 1% Interval When Volatility is Persistent. We simulate returns from a realistically-calibrated dynamic volatility model. We plot the series of true conditional exceedance probabilities, which we infer from the model. For visual reference we include a horizontal line at the desired 1% probability level. Adapted from Andersen et al. 2013.

fail to tighten appropriately in low-volatility times and widen appropriately in high-volatility times. Intervals can be completely mis-calibrated, correctly calibrated unconditionally but not conditionally, or correctly conditionally calibrated (which automatically implies correct conditional calibration). Figure 11.1 says it all

### Christoffersen's Absolute Interval Forecast Evaluation

Christoffersen (1998) considers likelihood-ratio tests of correct  $(1 - \alpha)\%$  conditional coverage. Define the sequence of hit indicators of a 1-step-ahead forecast interval (the “hit series”) as

$$I_t^{(1-\alpha)} = 1\{\text{realized } y_t \text{ falls inside the interval}\}$$

Under the null hypothesis of correct conditional calibration,

$$I_t^{(1-\alpha)} \sim \text{iid Bernoulli}(1 - \alpha).$$

Note well the two-part characterization. The hit series must have the correct mean,  $(1 - \alpha)$ , which corresponds to correct unconditional calibration. But there's more: the hit series must also be *iid*.<sup>1</sup> When both hold, we have correct conditional calibration. Conversely, rejection of the *iid Bernoulli* null could be due to rejection of *iid*, rejection of the *Bernoulli* mean of  $(1 - \alpha)$ , or both. Hence it is advisable to use constructive procedures, which, when rejections occur, convey information as to *why* rejections occur.

### On Testing *iid* in Forecast Evaluation

Note that in (1-step) forecast evaluation we're always testing some sort of 1-step error for *iid* (or at least white noise) structure.

For point forecasts the forecast errors are immediately at hand. If they're dependent, then, in general, today's error is informative regarding tomorrow's likely error, and we could we could generally use that information to adjust today's point forecast to make it better, which means something is wrong.

For interval forecasts, the correct notion of "error" is the hit sequence, which is readily constructed. If the hit sequence is dependent, then, in general, today's hit value (0 or 1) is informative regarding tomorrow's likely hit value, and we could we could generally use that information to adjust today's interval forecast to make it better conditionally calibrated, which means something is wrong.

Soon in section 11.2.1 we will introduce yet another generalized "forecast error" series for *density* forecasts, which again should be *iid* if all is well.

---

<sup>1</sup>In  $h$ -step-ahead contests the hit sequence need not be *iid* but should have  $h$ -dependent structure.

### 11.1.2 Relative Standards

Little studied. It seems clear that for two correctly conditionally calibrated interval forecasts, one should prefer the one with shorter average length. But, just as with bias-variance tradeoffs for point forecast evaluation, presumably one should willing to accept a little mis-calibration in exchange for a big length reduction. One would have to define a loss function over miscalibration and length.

## 11.2 Density Forecast Evaluation

### 11.2.1 Absolute Standards

#### Theory

We seek to characterize the properties of a density forecast that is optimal with respect to an information set, that is, a density forecast that coincides with the true conditional expectation.

The task of determining whether  $\{p_t(y_t|\Omega_t)\}_{t=1}^m = \{f_t(y_t|\Omega_t)\}_{t=1}^m$  appears difficult, perhaps hopeless, because  $\{f_t(y_t|\Omega_t)\}_{t=1}^m$  is never observed, even after the fact. Moreover, and importantly, the true density  $f_t(y_t|\Omega_t)$  may exhibit structural change, as indicated by its time subscript. As it turns out, the challenges posed by these subtleties are not insurmountable.

Our methods are based on the relationship between the data generating process,  $f_t(y_t)$ , and the sequence of density forecasts,  $p_t(y_t)$ , as related through the probability integral transform,  $z_t$ , of the realization of the process taken with respect to the density forecast. The probability integral transform is simply the cumulative density function corresponding to the density  $p_t(y_t)$  evaluated at  $y_t$ ,

$$\begin{aligned} z_t &= \int_{-\infty}^{y_t} p_t(u) du \\ &= P_t(y_t). \end{aligned}$$

The density of  $z_t$ ,  $q_t(z_t)$ , is of particular significance. Assuming that  $\frac{\partial P_t^{-1}(z_t)}{\partial z_t}$  is continuous and nonzero over the support of  $y_t$ , then because  $p_t(y_t) = \frac{\partial P_t(y_t)}{\partial y_t}$  and  $y_t = P_t^{-1}(z_t)$ ,  $z_t$  has support on the unit interval with density

$$\begin{aligned} q_t(z_t) &= \left| \frac{\partial P_t^{-1}(z_t)}{\partial z_t} \right| f_t(P_t^{-1}(z_t)) \\ &= \frac{f_t(P_t^{-1}(z_t))}{p_t(P_t^{-1}(z_t))}. \end{aligned}$$

Note, in particular, that if  $p_t(y_t) = f_t(y_t)$ , then  $q_t(z_t)$  is simply the  $U(0, 1)$  density.

Now we go beyond the one-period characterization of the density of  $z$  when  $p_t(y_t) = f_t(y_t)$  and characterize both the density and dependence structure of the entire  $z$  sequence when  $p_t(y_t) = f_t(y_t)$ .

**Proposition** Suppose  $\{y_t\}_{t=1}^m$  is generated from  $\{f_t(y_t|\Omega_t)\}_{t=1}^m$  where  $\Omega_t = \{y_{t-1}, y_{t-2}, \dots\}$ . If a sequence of density forecasts  $\{p_t(y_t)\}_{t=1}^m$  coincides with  $\{f_t(y_t|\Omega_t)\}_{t=1}^m$ , then under the usual condition of a non-zero Jacobian with continuous partial derivatives, the sequence of probability integral transforms of  $\{y_t\}_{t=1}^m$  with respect to  $\{p_t(y_t)\}_{t=1}^m$  is *iid*  $U(0, 1)$ . That is,

$$\{z_t\}_{t=1}^m \sim U(0, 1).$$

The intuition for the above result may perhaps be better understood from the perspective of Christoffersen's method for interval forecast evaluation. If a sequence of density forecasts is correctly conditionally calibrated, then *every* interval will be correctly conditionally calibrated and will generate an *iid* Bernoulli hit sequence. This fact manifests itself in the *iid* uniformity of the corresponding probability integral transforms.

## Practical Application

The theory developed thus far suggests that we evaluate density forecasts by assessing whether the probability integral transform series,  $\{z_t\}_{t=1}^m$ , is *iid*  $U(0, 1)$ . Simple tests of *iid*  $U(0, 1)$  behavior are readily available, such as those of Kolmogorov-Smirnov and Cramer-vonMises. Alone, however, such tests are not likely to be of much value in the practical applications that we envision, because they are not constructive; that is, when rejection occurs, the tests generally provide no guidance as to *why*. If, for example, a Kolmogorov-Smirnov test rejects the hypothesis of *iid*  $U(0, 1)$  behavior, is it because of violation of unconditional uniformity, violation of *iid*, or both? Moreover, even if we know that rejection comes from violation of uniformity, we would like to know more: What, precisely, is the nature of the violation of uniformity, and how important is it? Similarly, even if we know that rejection comes from a violation of *iid*, what precisely is its nature? Is  $z$  heterogeneous but independent, or is  $z$  dependent? If  $z$  is dependent, is the dependence operative primarily through the conditional mean, or are higher-ordered conditional moments, such as the variance, relevant? Is the dependence strong and important, or is *iid* an economically adequate approximation, even if strictly false?

Hence we adopt less formal, but more revealing, graphical methods, which we *supplement* with more formal tests. First, as regards unconditional uniformity, we suggest visual assessment using the obvious graphical tool, a density estimate. Simple histograms are attractive in the present context because they allow straightforward imposition of the constraint that  $z$  has support on the unit interval, in contrast to more sophisticated procedures such as kernel density estimates with the standard kernel functions. We visually compare the estimated density to a  $U(0, 1)$ , and we compute confidence intervals under the null hypothesis of *iid*  $U(0, 1)$  exploiting the binomial structure, bin-by-bin.

Second, as regards evaluating whether  $z$  is *iid*, we again suggest visual assessment using the obvious graphical tool, the correlogram, supplemented with the usual Bartlett confidence intervals. The correlogram assists with the detection of particular dependence patterns in  $z$  and can provide useful information about the deficiencies of density forecasts. For example, serial correlation in the  $z$  series indicates that conditional mean dynamics have been inadequately modeled captured by the forecasts. Because we are interested in potentially sophisticated nonlinear forms of dependence, not simply linear dependence, we examine not only the correlogram of  $(z - \bar{z})$ , but also those of powers of  $(z - \bar{z})$ . Examination of the correlograms of  $(z - \bar{z})$ ,  $(z - \bar{z})^2$ ,  $(z - \bar{z})^3$  and  $(z - \bar{z})^4$  should be adequate; it will reveal dependence operative through the conditional mean, conditional variance, conditional skewness, or conditional kurtosis.

### 11.2.2 Additional Discussion

#### Parameter Estimation Uncertainty

Our decision to ignore parameter estimation uncertainty was intentional. In our framework, the forecasts are the primitives, and we do not require that they be based on a model. This is useful because many density forecasts of interest, such as those from surveys, do not come from models. A second and very important example of model-free density forecasts is provided by the recent finance literature, which shows how to use options written at different strike prices to extract a model-free estimate of the market's risk-neutral density forecast of returns on the underlying asset. Moreover, many density forecasts based on estimated models already incorporate the effects of parameter estimation uncertainty, for example by using simulation techniques. Finally, sample sizes are often so large as to render negligible the effects of parameter estimation uncertainty, as for example in our simulation study.

### Improving Mis-Calibrated Density Forecasts

It is apparent that our methods can be used to improve defective density forecasts, in a fashion parallel to standard procedures for improving defective point forecasts. Recall that in the case of defective point forecasts case we can regress the  $y$ 's on the  $\hat{y}$ 's (the point forecasts) and use the estimated relationship to construct improved point forecasts. Similarly, in the context of density forecasts that are defective in that they produce an *iid* but non-uniform  $z$  sequence, we can exploit the fact that (in period  $m + 1$ , say)

$$\begin{aligned} f_{m+1}(y_{m+1}) &= p_{m+1}(y_{m+1}) q_{m+1}(P(y_{m+1})) \\ &= p_{m+1}(y_{m+1}) q_{m+1}(z_{m+1}). \end{aligned}$$

Thus if we know  $q_{m+1}(z_{m+1})$ , we would know the actual distribution  $f_{m+1}(y_{m+1})$ . Because  $q_{m+1}(z_{m+1})$  is unknown, we obtain an estimate  $\hat{q}_{m+1}(z_{m+1})$  using the historical series of  $z_{t=1}^m$ , and we use that estimate to construct an improved estimate,  $\hat{f}_{m+1}(y_{m+1})$ , of the true distribution. Standard density estimation techniques can be used to produce the estimate  $\hat{q}_{m+1}(z_{m+1})$ .<sup>2</sup>

### Multi-Step Density Forecasts

Our methods may be generalized to handle multi-step-ahead density forecasts, so long as we make provisions for serial correlation in  $z$ , in a fashion to the usual  $MA(h - 1)$  structure for optimal  $h$ -step ahead point forecast errors. It may prove most effective to partition the  $z$  series into groups for which we expect *iid* uniformity if the density forecasts were indeed correct. For instance, for correct 2-step ahead forecasts, the sub-series  $z_1, z_3, z_5, \dots$  and  $z_2, z_4, z_6, \dots$  should each be *iid*  $U(0, 1)$ , although the full series would not be *iid*  $U(0, 1)$ . If a formal test is desired, it may be obtained via Bonferroni

---

<sup>2</sup>In finite samples, of course, there is no guarantee that the “improved” forecast will actually be superior to the original, because it is based on an estimate of  $q$  rather than the true  $q$ , and the estimate could be very poor. In large samples, however, very precise estimation should be possible.

bounds, as suggested in a different context by Campbell and Ghysels (1995). Under the assumption that the  $z$  series is  $(h - 1)$ -dependent, each of the following  $h$  sub-series will be *iid*:  $\{z_1, z_{1+h}, z_{1+2h}, \dots\}$ ,  $\{z_2, z_{2+h}, z_{2+2h}, \dots\}$ , ...,  $\{z_h, z_{2h}, z_{3h}, \dots\}$ . Thus, a test with size bounded by  $\alpha$  can be obtained by performing  $h$  tests, each of size  $\alpha/h$ , on each of the  $h$  sub-series of  $z$ , and rejecting the null hypothesis of *iid* uniformity if the null is rejected for *any* of the  $h$  sub-series. With the huge high-frequency datasets now available in finance, such sample splitting, although inefficient, is not likely to cause important power deterioration.

### 11.2.3 Relative Standards

The time- $t$  one-step-ahead point predictive likelihood is

$$P_t = p_{t,t-1}(y_t)$$

It is simply the height of the earlier-made density forecast,  $p_{t,t-1}(\cdot)$  at the realized value,  $y_t$ . The full predictive likelihood is then

$$P = \prod_{i=1}^N P_t.$$

We can rank density forecasts using  $P$ . The sequence of density forecasts with the largest  $P$  is the the sequence for which the subsequently-observed realizations were most likely.

## 11.3 Stock Return Density Forecasting

### 11.3.1 A Preliminary GARCH Simulation

Before proceeding to apply our density forecast evaluation methods to real data, it is useful to examine their efficacy on simulated data, for which we

know the true data-generating process. We examine a simulated sample of length 8000 from the  $t$ -GARCH(1,1) process:

$$y_t = \sqrt{\frac{2h_t}{3}} t(6)$$

$$h_t = .01 + .13y_{t-1}^2 + .86h_{t-1}.$$

Both the sample size and the parameter values are typical for financial asset returns.<sup>3</sup> Throughout, we split the sample in half and use the “in-sample” observations 1 through 4000 for estimation, and the “out-of-sample” observations 4001 through 8000 for density forecast evaluation.

We will examine the usefulness of our density forecast evaluation methods in assessing four progressively better density forecasts. To establish a benchmark, we first evaluate forecasts based on the naive and incorrect assumption that the process is *iid*  $N(0, 1)$ .<sup>4</sup> That is, in each of the periods 4001-8000, we simply issue the forecast “ $N(0, 1)$ .”

In Figure \*\*\* we show two histograms of  $z$ , one with 20 bins and one with 40 bins.<sup>5</sup> The histograms have a distinct non-uniform “butterfly” shape – a hump in the middle and two wings on the sides – indicating that too many of the realizations fall in middle and in the tails of the forecast densities relative to what we would expect if the data were really *iid* normal. This is exactly what we hope the histograms would reveal, given that the data-generating process known to be unconditionally leptokurtic.

In Figure \*\*\* we show the correlograms of  $(z - \bar{z})$ ,  $(z - \bar{z})^2$ ,  $(z - \bar{z})^3$  and  $(z - \bar{z})^4$ .<sup>6</sup> The strong serial correlation in  $(z - \bar{z})^2$  (and hence  $(z - \bar{z})^4$ )

<sup>3</sup>The conditional variance function intercept of .01 is arbitrary but inconsequential; it simply amounts to a normalization of the unconditional variance to 1 (.01/(1-.13-.86)).

<sup>4</sup>The process as specified does have mean zero and variance 1, but it is neither *iid* nor unconditionally Gaussian.

<sup>5</sup>The dashed lines superimposed on the histogram are approximate 95% confidence intervals for the individual bin heights under the null that  $z$  is *iid*  $U(0, 1)$ .

<sup>6</sup>The dashed lines superimposed on the correlograms are Bartlett's approximate 95% confidence intervals under the null that  $z$  is *iid*.

makes clear another key deficiency of the  $N(0, 1)$  forecasts – they fail to capture the volatility dynamics operative in the process. Again, this is what we hope the correlograms would reveal, given our knowledge of the true data-generating process.

Second, we evaluate forecasts produced under the incorrect assumption that the process is *iid* but not necessarily Gaussian. We estimate the unconditional distribution from observations 1 through 4000, freeze it, and then issue it as the density forecast in each of the periods 4001 through 8000. Figures \*\*\* and \*\*\* contain the results. The  $z$  histogram is now almost perfect (as it must be, apart from estimation error, which is small in a sample of size 4000), but the correlograms correctly continue to indicate neglected volatility dynamics.

Third, we evaluate forecasts that are based on a  $GARCH(1, 1)$  model estimated under the incorrect assumption that the conditional density is Gaussian. We use observations 1 through 4000 to estimate the model, freeze the estimated model, and then use it to make (time-varying) density forecasts from 4001 through 8000. Figures \*\*\* and \*\*\* contain the  $z$  histograms and correlograms. The histograms are closer to uniform than those of Figure \*\*\*, but they still display slight peaks at either end and a hump in the middle. We would expect to see such a reduction, but not elimination, of the butterfly pattern, because allowance for conditionally Gaussian  $GARCH$  effects should account for some, but not all, unconditional leptokurtosis.<sup>7</sup> The correlograms now show no evidence of neglected conditional volatility dynamics, again as expected because the conditionally Gaussian  $GARCH$  model delivers consistent estimates of the conditional variance parameters, despite the fact that the conditional density is misspecified, so that the estimated model tracks the volatility dynamics well.

Finally, we forecast with an estimated correctly-specified  $t-GARCH(1, 1)$

---

<sup>7</sup>Recall that the data generating process is *conditionally*, as well as unconditionally, fat-tailed.

model. We show the  $z$  histogram and correlograms in Figures \*\*\* and \*\*\*. Because we are forecasting with a correctly specified model, estimated using a large sample, we would expect that the histogram and correlograms would fail to find flaws with the density forecasts, which is the case.

In closing this section, we note that at each step of the above simulation exercise, our density forecast evaluation procedures clearly and correctly revealed the strengths and weaknesses of the various density forecasts. The results, as with all simulation results, are specific to the particular data-generating process examined, but the process and the sample size were chosen to be realistic for the leading applications in high-frequency finance. This gives us confidence that the procedures will perform well on real financial data, to which we now turn, and for which we do not have the luxury of knowing the true data-generating process.

### 11.3.2 Daily S&P 500 Returns

We study density forecasts of daily value-weighted S&P 500 returns, with dividends, from 02/03/62 through 12/29/95. As before, we split the sample into in-sample and out-of-sample periods for model estimation and density forecast evaluation. There are 4133 in-sample observations (07/03/62 - 12/29/78) and 4298 out-of-sample observations (01/02/79 - 12/29/95). As before, we assess a series of progressively more sophisticated density forecasts.

As in the simulation example, we begin with an examination of  $N(0, 1)$  density forecasts, in spite of the fact that high-frequency financial data are well-known to be unconditionally leptokurtic and conditionally heteroskedastic. In Figures \*\*\* and \*\*\* we show the histograms and correlograms of  $z$ . The histograms have the now-familiar butterfly shape, indicating that the S&P realizations are leptokurtic relative to the  $N(0, 1)$  density forecasts, and the correlograms of  $(z - \bar{z})^2$  and  $(z - \bar{z})^4$  indicate that the  $N(0, 1)$  forecasts are severely deficient, because they neglect strong conditional volatility

dynamics.

Next, we generate density forecasts using an apparently much more sophisticated model. Both the Akaike and Schwarz information criteria select an  $MA(1) - GARCH(1, 1)$  model for the in-sample data, which we estimate, freeze, and use to generate out-of-sample density forecasts.

Figures \*\*\* and \*\*\* contain the  $z$  histograms and correlograms. The histograms are closer to uniform and therefore improved, although they still display slight butterfly pattern. The correlograms look even better; all evidence of neglected conditional volatility dynamics has vanished.

Finally, we estimate and then forecast with an  $MA(1) - t - GARCH(1, 1)$  model. We show the  $z$  histogram and correlograms in Figures \*\*\* and \*\*\*. The histogram is improved, albeit slightly, and the correlograms remain good.

## 11.4 Exercises, Problems and Complements

1. xxx

## 11.5 Notes



# Chapter 12

## Model-Based Forecast Combination

In forecast accuracy comparison, we ask which forecast is best with respect to a particular loss function. Such “horse races” arise constantly in practical work. Regardless of whether one forecast is significantly better than the others, however, the question arises as to whether competing forecasts may be fruitfully combined to produce a composite forecast superior to all the original forecasts. Thus, forecast combination, although obviously related to forecast accuracy comparison, is logically distinct and of independent interest. We start with what one might call “model-based” forecast combination, and then we proceed to “survey-based” combination and “market-based” combination (financial markets, prediction markets, ...).

### 12.1 Forecast Encompassing

Whether there are gains from forecast combination turns out to be fundamentally linked to the notion of forecast encompassing, with which we now begin. We use forecast encompassing tests to determine whether one forecast incorporates (or encompasses) all the relevant information in competing forecasts. If one forecast incorporates all the relevant information, nothing can be gained by combining forecasts. For simplicity, let’s focus on the case

of two forecasts,  $y_{a,t+h,t}$  and  $y_{b,t+h,t}$ . Consider the regression

$$y_{t+h} = \beta_a y_{a,t+h,t} + \beta_b y_{b,t+h,t} + \varepsilon_{t+h,t}.$$

If  $(\beta_a, \beta_b) = (1, 0)$ , we'll say that model  $a$  forecast-encompasses model  $b$ , and if  $(\beta_a, \beta_b) = (0, 1)$ , we'll say that model  $b$  forecast-encompasses model  $a$ . For other  $(\beta_a, \beta_b)$  values, neither model encompasses the other, and both forecasts contain useful information about  $y_{t+h}$ . In covariance stationary environments, encompassing hypotheses can be tested using standard methods.<sup>1</sup> If neither forecast encompasses the other, forecast combination is potentially desirable.

We envision an ongoing, iterative process of model selection and estimation, forecasting, and forecast evaluation. What is the role of forecast combination in that paradigm? In a world in which information sets can be instantaneously and costlessly combined, there is no role; it is always optimal to combine information sets rather than forecasts. That is, if no model forecast-encompasses the others, we might hope to eventually figure out what's gone wrong, learn from our mistakes, and come up with a model based on a combined information set that *does* forecast-encompass the others. But in the short run – particularly when deadlines must be met and timely forecasts produced – pooling of information sets is typically either impossible or prohibitively costly. This simple insight motivates the pragmatic idea of forecast combination, in which forecasts rather than models are the basic object of analysis, due to an assumed inability to combine information sets. Thus, forecast combination can be viewed as a key link between the short-run, real-time forecast production process, and the longer-run, ongoing process of model development.

---

<sup>1</sup>Note that  $\varepsilon_{t+h,t}$  may be serially correlated, particularly if  $h > 1$ , and any such serial correlation should be accounted for.

## 12.2 Model-Based Combined Forecasts I: Variance-Covariance Forecast Combination

In forecast accuracy comparison, we ask which forecast is best with respect to a particular loss function. Such “horse races” arise constantly in practical work. Regardless of whether one forecast is significantly better than the others, however, the question arises as to whether competing forecasts may be fruitfully combined to produce a composite forecast superior to all the original forecasts. Thus, forecast combination, although obviously related to forecast accuracy comparison, is logically distinct and of independent interest.

Failure of each model’s forecasts to encompass other model’s forecasts indicates that both models are misspecified, and that there may be gains from combining the forecasts. It should come as no surprise that such situations are typical in practice, because forecasting models are *likely* to be misspecified – they are intentional abstractions of a much more complex reality. Many combining methods have been proposed, and they fall roughly into two groups, “variance-covariance” methods and “regression” methods. As we’ll see, the variance-covariance forecast combination method is in fact a special case of the regression-based forecast combination method, so there’s really only one method. However, for historical reasons – and more importantly, to build valuable intuition – it’s important to understand the variance-covariance forecast combination, so let’s begin with it.

### 12.2.1 Bivariate Case

Suppose we have two unbiased forecasts. First assume that the errors in  $y_a$  and  $y_b$  are uncorrelated. Consider the convex combination

$$y_C = \lambda y_a + (1 - \lambda) y_b,$$

where  $\lambda \in [0, 1]$ .<sup>2</sup> Then the associated errors follow the same weighting,

$$e_C = \lambda e_a + (1 - \lambda) e_b,$$

where  $e_C = y - y_C$ ,  $e_a = y - y_a$  and  $e_b = y - y_b$ . Assume that both  $y_a$  and  $y_b$  are unbiased for  $y$ , in which case  $y_C$  is also unbiased, because the combining weights sum to unity.

Given the unbiasedness assumption, the minimum-MSE combining weights are just the minimum-variance weights. Immediately, using the assumed zero correlation between the errors,

$$\sigma_C^2 = \lambda^2 \sigma_a^2 + (1 - \lambda)^2 \sigma_b^2, \quad (12.1)$$

where  $\sigma_C^2 = \text{var}(e_C)$ ,  $\sigma_a^2 = \text{var}(e_a)$  and  $\sigma_b^2 = \text{var}(e_b)$ . Minimization with respect to  $\lambda$  yields the optimal combining weight,

$$\lambda^* = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_a^2} = \frac{1}{1 + \phi^2}, \quad (12.2)$$

where  $\phi = \sigma_a/\sigma_b$ .

As  $\sigma_a^2$  approaches 0, forecast a becomes progressively more accurate. The formula for  $\lambda^*$  indicates that as  $\sigma_a^2$  approaches 0,  $\lambda^*$  approaches 1, so that all weight is put on forecast a, which is desirable. Similarly, as  $\sigma_b^2$  approaches 0, forecast b becomes progressively more accurate. The formula for  $\lambda^*$  indicates that as  $\sigma_b^2$  approaches 0,  $\lambda^*$  approaches 0, so that all weight is put on forecast b, which is also desirable. In general, the forecast with the smaller error variance receives the higher weight, with the precise size of the weight depending on the disparity between variances.

Now consider the more general and empirically-relevant case of correlated

---

<sup>2</sup>Strictly speaking, we need not even impose  $\lambda \in [0, 1]$ , but  $\lambda \notin [0, 1]$  would be highly nonstandard for two valuable and sophisticated y estimates such as  $y_a$  and  $y_b$ .

errors. Under the same conditions as earlier,

$$\sigma_C^2 = \lambda^2 \sigma_a^2 + (1 - \lambda)^2 \sigma_b^2 + 2\lambda(1 - \lambda)\sigma_{ab}, \quad (12.3)$$

so

$$\begin{aligned}\lambda^* &= \frac{\sigma_b^2 - \sigma_{ab}}{\sigma_b^2 + \sigma_a^2 - 2\sigma_{ab}} \\ &= \frac{1 - \phi\rho}{1 + \phi^2 - 2\phi\rho},\end{aligned}$$

where  $\sigma_{ab} = \text{cov}(e_a, e_b)$  and  $\rho = \text{corr}(e_a, e_b)$ .

The optimal combining weight is a simple function of the variances and covariances of the underlying forecast errors. The forecast error variance associated with the optimally combined forecast is less than or equal to the smaller of  $\sigma_a^2$  and  $\sigma_b^2$ ; thus, in population, we have nothing to lose by combining forecasts, and potentially much to gain. In practical applications, the unknown variances and covariances that underlie the optimal combining weights are unknown, so we replace them with consistent estimates; that is, we estimate  $\lambda^*$  by replacing unknown error variances and covariances with estimates, yielding

$$\hat{\lambda}^* = \frac{\hat{\sigma}_b^2 - \hat{\sigma}_{ab}^2}{\hat{\sigma}_b^2 + \hat{\sigma}_a^2 - 2\hat{\sigma}_{ab}^2}.$$

The full formula for the optimal combining weight indicates that the variances *and* the covariance are relevant, but the basic intuition remains valid. Effectively, we're forming a portfolio of forecasts, and as we know from standard results in finance, the optimal shares in a portfolio depend on the variances *and* covariances of the underlying assets.

### 12.2.2 General Case

The optimal combining weight solves the following problem:

$$\begin{aligned} \min_{\lambda} & \lambda' \Sigma_t \lambda \\ \text{s.t. } & \lambda' \iota = 1. \end{aligned} \tag{12.4}$$

where  $\Sigma$  is the  $N \times N$  covariance matrix of forecast errors and  $\iota$  is a  $N \times 1$  vector of ones. The solution is

$$\lambda^* = (\iota' \Sigma_t^{-1} \iota)^{-1} \Sigma_t^{-1} \iota.$$

## 12.3 Model-Based Combined Forecasts II: Regression-Based Forecast Combination

Now consider the regression method of forecast combination. The form of forecast-encompassing regressions immediately suggests combining forecasts by simply regressing realizations on forecasts. This intuition proves accurate, and in fact the optimal variance-covariance combining weights have a regression interpretation as the coefficients of a linear projection of  $y_{t+h}$  onto the forecasts, subject to two constraints: the weights sum to unity, and the intercept is excluded.

In practice, of course, population linear projection is impossible, so we simply run the regression on the available data. Moreover, it's usually preferable *not* to force the weights to add to unity, or to exclude an intercept. Inclusion of an intercept, for example, facilitates bias correction and allows biased forecasts to be combined. Typically, then, we simply estimate the regression,

$$y_{t+h} = \beta_0 + \beta_a y_{a,t+h,t} + \beta_b y_{b,t+h,t} + \varepsilon_{t+h,t}.$$

Extension to the fully general case of more than two forecasts is immediate.

In general, the regression method is simple and flexible. There are many variations and extensions, because any regression tool is potentially applicable. The key is to use generalizations with sound motivation. We'll give four examples in an attempt to build an intuitive feel for the sorts of extensions that are possible: time-varying combining weights, dynamic combining regressions, shrinkage of combining weights toward equality, and nonlinear combining regressions.

### 12.3.1 Time-Varying Combining Weights

Relative accuracies of different forecasts may change, and if they do, we naturally want to weight the improving forecasts progressively more heavily and the worsening forecasts less heavily. Relative accuracies can change for a number of reasons. For example, the design of a particular forecasting model may make it likely to perform well in some situations, but poorly in others. Alternatively, people's decision rules and firms' strategies may change over time, and certain forecasting techniques may be relatively more vulnerable to such change.

We allow for time-varying combining weights in the regression framework by using weighted or rolling estimation of combining regressions, or by allowing for explicitly time-varying parameters. If, for example, we suspect that the combining weights are evolving over time in a trend-like fashion, we might use the combining regression

$$\begin{aligned} y_{t+h} = & (\beta_0^0 + \beta_0^1 TIME) + (\beta_a^0 + \beta_a^1 TIME)y_{a,t+h,t} \\ & + (\beta_b^0 + \beta_b^1 TIME)y_{b,t+h,t} + \varepsilon_{t+h,t}, \end{aligned}$$

which we estimate by regressing the realization on an intercept, time, each of the two forecasts, the product of time and the first forecast, and the product of time and the second forecast. We assess the importance of time variation

by examining the size and statistical significance of the estimates of  $\beta_0^1$ ,  $\beta_a^1$ , and  $\beta_b^1$ .

### 12.3.2 Serial Correlation

It's a good idea to allow for serial correlation in combining regressions, for two reasons. First, as always, even in the best of conditions we need to allow for the usual serial correlation induced by overlap when forecasts are more than 1-step-ahead. This suggests that instead of treating the disturbance in the combining regression as white noise, we should allow for  $MA(h - 1)$  serial correlation,

$$\begin{aligned} y_{t+h} &= \beta_0 + \beta_a y_{a,t+h,t} + \beta_b y_{b,t+h,t} + \varepsilon_{t+h,t} \\ \varepsilon_{t+h,t} &\sim MA(h - 1). \end{aligned}$$

Second, and very importantly, the  $MA(h - 1)$  error structure is associated with forecasts that are optimal with respect to their information sets, of which there's no guarantee. That is, although the primary forecasts were designed to capture the dynamics in  $y$ , there's no guarantee that they do so. Thus, just as in standard regressions, it's important in combining regressions that we allow either for serially correlated disturbances or lagged dependent variables, to capture any dynamics in  $y$  not captured by the various forecasts. A combining regression with  $ARMA(p, q)$  disturbances,

$$\begin{aligned} y_{t+h} &= \beta_0 + \beta_a y_{a,t+h,t} + \beta_b y_{b,t+h,t} + \varepsilon_{t+h,t} \\ \varepsilon_{t+h,t} &\sim ARMA(p, q), \end{aligned}$$

with  $p$  and  $q$  selected using information criteria in conjunction with other diagnostics, is usually adequate.

### 12.3.3 Shrinkage of Combining Weights Toward Equality

Simple arithmetic averages of forecasts – that is, combinations in which the weights are constrained to be equal – sometimes perform very well in out-of-sample forecast competitions, even relative to “optimal” combinations. The equal-weights constraint eliminates sampling variation in the combining weights at the cost of possibly introducing bias. Sometimes the benefits of imposing equal weights exceed the cost, so that the *MSE* of the combined forecast is reduced.

The equal-weights constraint associated with the arithmetic average is an example of extreme shrinkage; regardless of the information contained in the data, the weights are forced into equality. We’ve seen before that shrinkage can produce forecast improvements, but typically we want to *coax* estimates in a particular direction, rather than to force them. In that way we guide our parameter estimates toward reasonable values when the data are uninformative, while nevertheless paying a great deal of attention to the data when they are informative.

Thus, instead of imposing a *deterministic* equal-weights constraint, we might like to impose a *stochastic* constraint. With this in mind, we sometimes coax the combining weights toward equality without forcing equality. A simple way to do so is to take a weighted average of the simple average combination and the least-squares combination. Let the shrinkage parameter  $\gamma$  be the weight put on the simple average combination, and let  $(1-\gamma)$  be the weight put on the least-squares combination, where  $\gamma$  is chosen by the user. The larger is  $\gamma$ , the more the combining weights are shrunk toward equality. Thus the combining weights are coaxed toward the arithmetic mean, but the data are still allowed to speak, when they have something important to say.

### 12.3.4 Nonlinear Combining Regressions

There is no reason to force linearity of combining regressions, and various of the nonlinear techniques that we've already introduced may be used. We might, for example, regress realizations not only on forecasts, but also on squares and cross products of the various forecasts, in order to capture quadratic deviations from linearity,

$$\begin{aligned} y_{t+h} = & \beta_0 + \beta_a y_{a,t+h,t} + \beta_b y_{b,t+h,t} \\ & + \beta_{aa}(y_{a,t+h,t})^2 + \beta_{bb}(y_{b,t+h,t})^2 + \beta_{ab} y_{a,t+h,t} y_{b,t+h,t} + \varepsilon_{t+h,t}. \end{aligned}$$

We assess the importance of nonlinearity by examining the size and statistical significance of estimates of  $\beta_{aa}$ ,  $\beta_{bb}$ , and  $\beta_{ab}$ ; if the linear combining regression is adequate, those estimates should differ significantly from zero. If, on the other hand, the nonlinear terms are found to be important, then the full nonlinear combining regression should be used.

### 12.3.5 Regularized Regression for Combining Large Numbers of Forecasts

Another, related, approach, involving both shrinkage and selection, is lasso and other “regularization” methods. Lasso can be used to shrink and select, and it’s a simple matter to make the shrinkage/selection direction “equal weights” rather than the standard lasso “zero weights.”

## 12.4 Application: OverSea Shipping Volume Revisited

Now let's combine the forecasts. Both failed Mincer-Zarnowitz tests, which suggests that there may be scope for combining. The correlation between the two forecast errors is .54, positive but not too high. In Table 9 we show the results of estimating the unrestricted combining regression with

$MA(1)$  errors (equivalently, a forecast encompassing test). Neither forecast encompasses the other; both combining weights, as well as the intercept, are highly statistically significantly different from zero. Interestingly, the judgmental forecast actually gets *more* weight than the quantitative forecast in the combination, in spite of the fact that its  $RMSE$  was higher. That's because, after correcting for bias, the judgmental forecast appears a bit more accurate.

**Table 9**  
Shipping Volume Combining Regression

LS // Dependent Variable is VOL

Sample: 1/01/1988 7/18/1997

Included observations: 499

Convergence achieved after 11 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.181977	0.259774	8.399524	0.0000
VOLQ	0.291577	0.038346	7.603919	0.0000
VOLJ	0.630551	0.039935	15.78944	0.0000
MA(1)	0.951107	0.014174	67.10327	0.0000
R-squared	0.957823	Mean dependent var	19.80609	
Adjusted R-squared	0.957567	S.D. dependent var	3.403283	
S.E. of regression	0.701049	Akaike info criterion	-0.702371	
Sum squared resid	243.2776	Schwarz criterion	-0.668603	
Log likelihood-528.8088		F-statistic	3747.077	
Durbin-Watson stat	1.925091	Prob(F-statistic)	0.000000	
Inverted MA Roots	-.95			

## 12.5 On the Optimality of Equal Weights

### 12.5.1 Under Quadratic Loss

In Figure 12.1 we graph  $\lambda^*$  as a function of  $\phi$ , for  $\phi \in [.75, 1.45]$ .  $\lambda^*$  is of course decreasing in  $\phi$ , but interestingly, it is only mildly sensitive to  $\phi$ . Indeed, for our range of  $\phi$  values, the optimal combining weight remains close to 0.5, varying from roughly 0.65 to 0.30. At the midpoint  $\phi = 1.10$ , we have  $\lambda^* = 0.45$ .

It is instructive to compare the error variance of combined  $y$ ,  $\sigma_C^2$ , to  $\sigma_a^2$  for a range of  $\lambda$  values (including  $\lambda = \lambda^*$ ,  $\lambda = 0$ , and  $\lambda = 1$ ).<sup>3</sup> From (12.1) we have:

$$\frac{\sigma_C^2}{\sigma_a^2} = \lambda^2 + \frac{(1 - \lambda)^2}{\phi^2}.$$

In Figure 12.2 we graph  $\sigma_C^2/\sigma_a^2$  for  $\lambda \in [0, 1]$  with  $\phi = 1.1$ . Obviously the max-

<sup>3</sup>We choose to examine  $\sigma_C^2$  relative to  $\sigma_a^2$ , rather than to  $\sigma_b^2$ , because  $y_a$  is the “standard”  $y$  estimate used in practice almost universally. A graph of  $\sigma_C^2/\sigma_b^2$  would be qualitatively identical, but the drop below 1.0 would be less extreme.

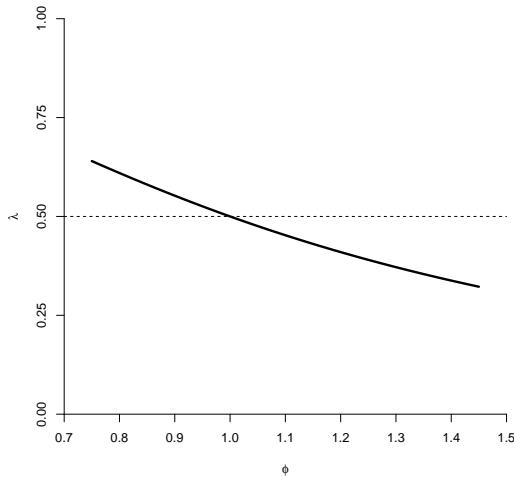


Figure 12.1:  $\lambda^*$  vs.  $\phi$ .  $\lambda^*$  constructed assuming uncorrelated errors. The horizontal line for visual reference is at  $\lambda^* = .5$ . See text for details.

imum variance reduction is obtained using  $\lambda^* = 0.45$ , but even for nonoptimal  $\lambda$ , such as simple equal-weight combination ( $\lambda = 0.5$ ), we achieve substantial variance reduction relative to using  $y_a$  alone. Indeed, a key result is that *for all*  $\lambda$  (except those very close to 1, of course) we achieve substantial variance reduction.

In Figure 12.3 we show  $\lambda^*$  as a function of  $\phi$  for  $\rho = 0, 0.3, 0.45$  and  $0.6$ ; in Figure 12.4 we show  $\lambda^*$  as a function of  $\rho$  for  $\phi = 0.95, 1.05, 1.15$  and  $1.25$ ; and in Figure 12.5 we show  $\lambda^*$  as a bivariate function of  $\phi$  and  $\rho$ . For  $\phi = 1$  the optimal weight is 0.5 for all  $\rho$ , but for  $\phi \neq 1$  the optimal weight differs from 0.5 and is more sensitive to  $\phi$  as  $\rho$  grows. The crucial observation remains, however, that under a wide range of conditions it is optimal to put significant weight on both  $y_a$  and  $y_b$ , with the optimal weights not differing radically from equality. Moreover, for all  $\phi$  values greater than one, so that less weight is optimally placed on  $y_a$  under a zero-correlation assumption, allowance for positive correlation further decreases the optimal weight placed on  $y_a$ . For a benchmark calibration of  $\phi = 1.1$  and  $\rho = 0.45$ ,  $\lambda^* \approx 0.41$ .

Let us again compare  $\sigma_C^2$  to  $\sigma_a^2$  for a range of  $\lambda$  values (including  $\lambda = \lambda^*$ ,

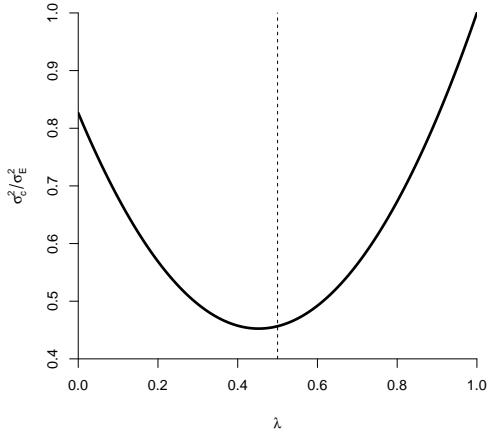


Figure 12.2:  $\sigma_C^2/\sigma_a^2$  for  $\lambda \in [0, 1]$ . We assume  $\phi = 1.1$  and uncorrelated errors. See text for details.

$\lambda = 0$ , and  $\lambda = 1$ ). From (12.3) we have:

$$\frac{\sigma_C^2}{\sigma_a^2} = \lambda^2 + \frac{(1 - \lambda)^2}{\phi^2} + 2\lambda(1 - \lambda)\frac{\rho}{\phi}.$$

In Figure 12.6 we graph  $\sigma_C^2/\sigma_a^2$  for  $\lambda \in [0, 1]$  with  $\phi = 1.1$  and  $\rho = 0.45$ . Obviously the maximum variance reduction is obtained using  $\lambda^* = 0.41$ , but even for nonoptimal  $\lambda$ , such as simple equal-weight combination ( $\lambda = 0.5$ ), we achieve substantial variance reduction relative to using  $y_a$  alone.

The “equal weights puzzle.” It is clear from our analysis above that in realistic situations (similar variances, small or moderate correlations) the gains from optimally combining can be massive, and that the loss from combining with equal weights relative to optimal weights is small. That is, optimal weights are not generally equal, but combining with equal weights is often not far from the optimum, and *much* better than any primary forecast. Equal weights are fully optimal, moreover, in the equi-correlation case, or more generally, in the Elliott case. Also, from an estimation perspective, equal weights may be slightly biased, but they have no variance! So the equal weight puzzle is perhaps not such a puzzle.

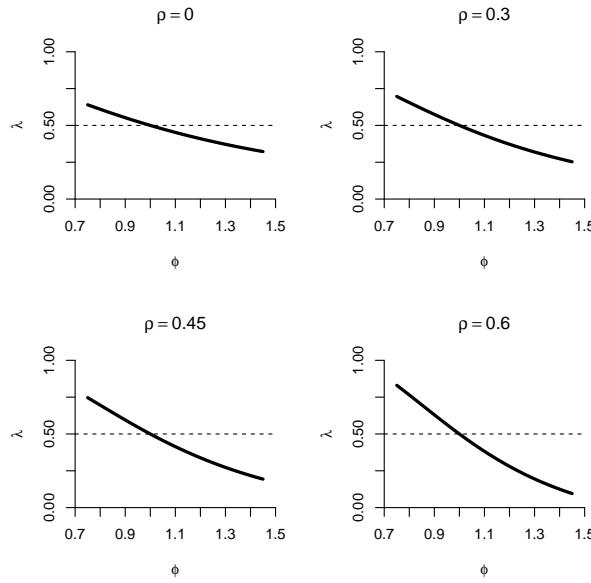


Figure 12.3:  $\lambda^*$  vs.  $\phi$  for Various  $\rho$  Values. The horizontal line for visual reference is at  $\lambda^* = .5$ . See text for details.

### 12.5.2 Under Minimax Loss

Here we take a more conservative perspective on forecast combination, solving a different but potentially important optimization problem. We utilize the minimax framework of ?, which is the main decision-theoretic approach for imposing conservatism and therefore of intrinsic interest. We solve a game between a benevolent scholar (the Econometrician) and a malevolent opponent (Nature). In that game the Econometrician chooses the combining weights, and Nature selects the stochastic properties of the forecast errors. The minimax solution yields the combining weights that deliver the smallest chance of the worst outcome for the Econometrician. Under the minimax approach knowledge or calibration of objects like  $\phi$  and  $\rho$  is unnecessary, enabling us to dispense with judgment, for better or worse.

We obtain the minimax weights by solving for the Nash equilibrium in a two-player zero-sum game. Nature chooses the properties of the forecast errors and the Econometrician chooses the combining weights  $\lambda$ . For exposi-

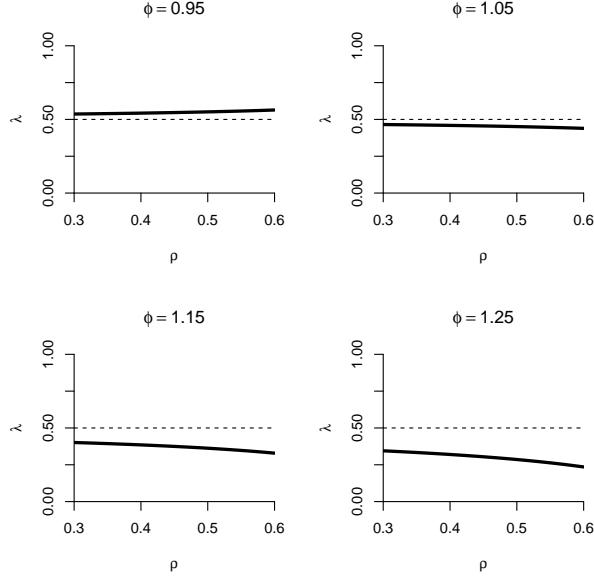


Figure 12.4:  $\lambda^*$  vs.  $\rho$  for Various  $\phi$  Values. The horizontal line for visual reference is at  $\lambda^* = .5$ . See text for details.

tional purposes, we begin with the case of uncorrelated errors, constraining Nature to choose  $\rho = 0$ . To impose some constraints on the magnitude of forecast errors that Nature can choose, it is useful to re-parameterize the vector  $(\sigma_b, \sigma_a)'$  in terms of polar coordinates; that is, we let  $\sigma_b = \psi \cos \varphi$  and  $\sigma_a = \psi \sin \varphi$ . We restrict  $\psi$  to the interval  $[0, \bar{\psi}]$  and let  $\varphi \in [0, \pi/2]$ . Because  $\cos^2 \varphi + \sin^2 \varphi = 1$ , the sum of the forecast error variances associated with  $y_a$  and  $y_b$  is constrained to be less than or equal to  $\bar{\psi}^2$ . The error associated with the combined forecast is given by

$$\sigma_C^2(\psi, \varphi, \lambda) = \psi^2 [\lambda^2 \sin^2 \varphi + (1 - \lambda)^2 \cos^2 \varphi]. \quad (12.5)$$

so that the minimax problem is

$$\max_{\psi \in [0, \bar{\psi}], \varphi \in [0, \pi/2]} \min_{\lambda \in [0, 1]} \sigma_C^2(\psi, \varphi, \lambda). \quad (12.6)$$

The best response of the Econometrician was derived in (12.2) and can be expressed in terms of polar coordinates as  $\lambda^* = \cos^2 \varphi$ . In turn, Nature's

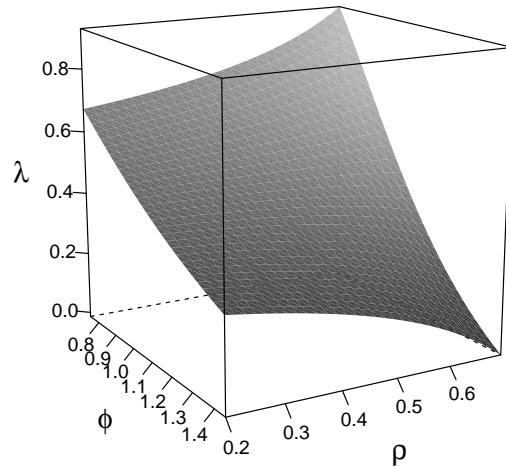


Figure 12.5:  $\lambda^*$  vs.  $\rho$  and  $\phi$ . See text for details.

problem simplifies to

$$\max_{\psi \in [0, \bar{\psi}], \varphi \in [0, \pi/2]} \psi^2 (1 - \sin^2 \varphi) \sin^2 \varphi,$$

which leads to the solution

$$\varphi^* = \arcsin \sqrt{1/2}, \quad \psi^* = \bar{\psi}, \quad \lambda^* = 1/2. \quad (12.7)$$

Nature's optimal choice implies a unit forecast error variance ratio,  $\phi = \sigma_a/\sigma_b = 1$ , and hence that the optimal combining weight is  $1/2$ . If, instead, Nature set  $\varphi = 0$  or  $\varphi = \pi/2$ , that is  $\phi = 0$  or  $\phi = \infty$ , then either  $y_a$  or  $y_b$  is perfect and the Econometrician could choose  $\lambda = 0$  or  $\lambda = 1$  to achieve a perfect forecast leading to a suboptimal outcome for Nature.

Now we consider the case in which Nature can choose a nonzero correlation between the forecast errors of  $y_a$  and  $y_b$ . The loss of the combined forecast

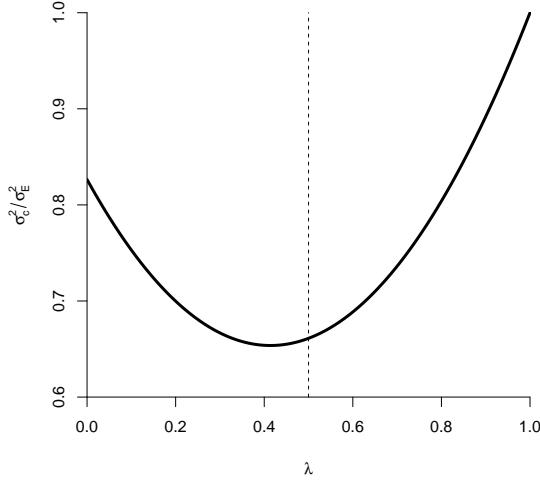


Figure 12.6:  $\sigma_C^2/\sigma_a^2$  for  $\lambda \in [0, 1]$ . We assume  $\phi = 1.1$  and  $\rho = 0.45$ . See text for details.

can be expressed as

$$\sigma_C^2(\psi, \rho, \varphi, \lambda) = \psi^2 [\lambda^2 \sin^2 \varphi + (1 - \lambda)^2 \cos^2 \varphi + 2\lambda(1 - \lambda)\rho \sin \varphi \cos \varphi]. \quad (12.8)$$

It is apparent from (12.8) that as long as  $\lambda$  lies in the unit interval the most devious choice of  $\rho$  is  $\rho^* = 1$ . We will now verify that conditional on  $\rho^* = 1$  the solution in (12.7) remains a Nash Equilibrium. Suppose that the Econometrician chooses equal weights,  $\lambda^* = 1/2$ . In this case

$$\sigma_C^2(\psi, \rho^*, \varphi, \lambda^*) = \psi^2 \left[ \frac{1}{4} + \frac{1}{2} \sin \varphi \cos \varphi \right].$$

We can deduce immediately that  $\psi^* = \bar{\psi}$ . Moreover, first-order conditions for the maximization with respect to  $\varphi$  imply that  $\cos^2 \varphi^* = \sin^2 \varphi^*$  which in turn leads to  $\varphi^* = \arcsin \sqrt{1/2}$ . Conditional on Nature choosing  $\rho^*$ ,  $\psi^*$ , and  $\varphi^*$ , the Econometrician has no incentive to deviate from the equal-weights combination  $\lambda^* = 1/2$ , because

$$\sigma_C^2(\psi^*, \rho^*, \varphi^*, \lambda) = \frac{\bar{\psi}}{2} \left[ \lambda^2 + (1 - \lambda)^2 + 2\lambda(1 - \lambda) \right] = \frac{\bar{\psi}}{2}.$$

In sum, the minimax analysis provides a rational for combining  $y_a$  and  $y_b$  with equal weights of  $\lambda = 1/2$ . Of course it does not resolve the equal weights puzzle, which refers to quadratic loss, but it puts equal weights on an even higher pedestal, and from a very different perspective.

## 12.6 Interval Forecast Combination

## 12.7 Density Forecast Combination

### 12.7.1 Choosing Weights to Optimize a Predictive Likelihood

Has Bayesian foundations. Geweke-Amisano.

### 12.7.2 Choosing Weights Optimize Conditional Calibration

Maximize a test statistic for iid uniformity of the PIT.

## 12.8 Exercises, Problems and Complements

### 1. Combining Forecasts.

You are a managing director at Paramex, a boutique investment bank in Los Angeles. Each day during the summer your two interns give you a 1-day-ahead forecast of the Euro/Dollar exchange rate. At the end of the summer, you calculate each intern's series of daily forecast errors. You find that the mean errors are zero, and the error variances and covariances are  $\hat{\sigma}_{AA}^2 = 153.76$ ,  $\hat{\sigma}_{BB}^2 = 92.16$ , and  $\hat{\sigma}_{AB}^2 = .2$ .

- (a) If you were forced to choose between the two forecasts, which would you choose? Why?
- (b) If instead you had the opportunity to combine the two forecasts by forming a weighted average, what would be the optimal weights

according to the variance-covariance method? Why?

- (c) Is it guaranteed that a combined forecast formed using the “optimal” weights calculated in part 1b will have lower mean squared prediction error? Why or why not?
- 2. The algebra of forecast combination.

Consider the combined forecast,

$$y_{t+h,t}^c = \lambda y_{t+h,t}^a + (1 - \lambda) y_{t+h,t}^b.$$

Verify the following claims made in the text:

- a. The combined forecast error will satisfy the same relation as the combined forecast; that is,

$$e_{t+h,t}^c = \lambda e_{t+h,t}^a + (1 - \lambda) e_{t+h,t}^b$$

- b. Because the weights sum to unity, if the primary forecasts are unbiased then so too is the combined forecast.
- c. The variance of the combined forecast error is

$$\sigma_c^2 = \lambda^2 \sigma_{aa}^2 + (1 - \lambda)^2 \sigma_{bb}^2 + 2\lambda(1 - \lambda)\sigma_{ab}^2,$$

where  $\sigma_{aa}^2$  and  $\sigma_{bb}^2$  are unconditional forecast error variances and  $\sigma_{ab}^2$  is their covariance.

- d. The combining weight that minimizes the combined forecast error variance (and hence the combined forecast error  $MSE$ , by unbiasedness) is

$$\lambda^* = \frac{\sigma_{bb}^2 - \sigma_{ab}^2}{\sigma_{bb}^2 + \sigma_{aa}^2 - 2\sigma_{ab}^2}.$$

- e. If neither forecast encompasses the other, then

$$\sigma_c^2 < \min(\sigma_{aa}^2, \sigma_{bb}^2).$$

- f. If one forecast encompasses the other, then

$$\sigma_c^2 = \min(\sigma_{aa}^2, \sigma_{bb}^2).$$

3. Quantitative forecasting, judgmental forecasting, forecast combination, and shrinkage.

Interpretation of the modern quantitative approach to forecasting as eschewing judgment is most definitely misguided. How is judgment used routinely and informally to modify quantitative forecasts? How can judgment be formally used to modify quantitative forecasts via forecast combination? How can judgment be formally used to modify quantitative forecasts via shrinkage? Discuss the comparative merits of each approach.

4. The empirical success of forecast combination.

In the text we mentioned that we have nothing to lose by forecast combination, and potentially much to gain. That's certainly true in population, with optimal combining weights. However, in finite samples of the size typically available, sampling error contaminates the combining weight estimates, and the problem of sampling error may be exacerbated by the collinearity that typically exists between  $y_{t+h,t}^a$  and  $y_{t+h,t}^b$ . Thus, while we hope to reduce out-of-sample forecast *MSE* by combining, there is no guarantee. Fortunately, however, in practice forecast combination often leads to very good results. The efficacy of forecast combination is well-documented in a vast literature.

5. Regression forecasting models with expectations, or anticipatory, data.

A number of surveys exist of anticipated market conditions, investment intentions, buying plans, advance commitments, consumer sentiment, and so on.

- (a) Search the World Wide Web for such series and report your results. A good place to start is the Resources for Economists page mentioned in Chapter ??.
- (b) How might you use the series you found in a regression forecasting model of  $y$ ? Are the implicit forecast horizons known for all the anticipatory series you found? If not, how might you decide how to lag them in your regression forecasting model?
- (c) How would you test whether the anticipatory series you found provide incremental forecast enhancement, relative to the own past history of  $y$ ?

## 6. Crowd-sourcing via internet activity.

How, in a sense, are trends identified by search data (on Google, YouTube, ...), tweets, etc. “combined forecasts”?

## 7. Turning a set of point forecasts into a combined density forecast.

We can produce a combined density forecast by drawing from an estimate of the density of the combining regression disturbances, as we did in a different context in section 4.1.

## 12.9 Notes

The idea of forecast encompassing dates at least to Nelson (1972), and was formalized and extended by Chong and Hendry (1986) and Fair and Shiller (1990). The variance-covariance method of forecast combination is due to Bates and Granger (1969), and the regression interpretation is due to Granger

and Ramanathan (1984). Surveys of econometric forecast combination include Diebold and Lopez (1996) and Timmermann (2006). Surveys of survey-based combination include Pesaran and Weale (2006). Snowberg et al. (2013) (prediction markets) provide a nice review of prediction markets.



# Chapter 13

## Market-Based Forecast Combination

### 13.1 Financial Markets

Markets can be a spectacularly effective method of information aggregation, as shown in both classic theoretical treatments of price systems as information aggregators (e.g., Koopmans (1957)) and similarly classic experimental work e.g., Plott (2000)). Hence one might suspect that markets would be useful for combining forecasts. In this section we explore that idea.

Markets and surveys are in certain respects opposite extremes. Markets are loved by economists (as market participants have real money on the line), and surveys are often unloved by economists (as survey participants typically have nothing on the line). That is, because market participants have real money on the line, markets may be more likely than otherwise to truthfully reveal traders' views, via their trading decisions). In any event, both market-based combined forecasts and survey-based combined forecasts are very different from forecasts from a simple single model.

Financial markets are naturally forward-looking, and market forecasts can sometimes be extracted from financial asset prices. There are many examples.

### 13.1.1 General Principles

#### Point Forecasts From Forward Markets

A classic example is using forward foreign exchange rates as forecasts of future spot exchange rates. Under risk neutrality, agents will always buy forward foreign exchange when it's "cheap" (under-priced) relative to their expectations of the future spot rate, and sell when it's "dear" (over-priced). Immediately then, we have that under risk neutrality:

$$F_t(t+h) = E_t(S_{t+h}),$$

where  $F_t(t+h)$  is the  $h$ -day forward rate prevailing at time  $t$  and  $E_t(S_{t+h})$  is the optimal (conditional mean) spot-rate forecast made at time  $t$  for time  $t+h$ .

Note well that in this example, and in financial markets more generally, typically only "risk neutral" forecasts are easy to extract from financial markets, the real-world usefulness of which remains an issue. That is, risk premia, which moreover are likely time-varying due to the time-varying financial-market volatility emphasized in Chapter 8, are always a potential issue.<sup>1</sup>

#### Point Forecasts From Futures Markets

Futures markets exist for many many things, trading contracts not only for standard financials (e.g., currencies, interest rates, ...), but also for myriad other things, including aspects of energy, agriculture, metals and other commodities – even weather, real estate, and stock market volatility!

Because futures are traded on exchanges, you can learn a lot about the contracts traded and currently-prevailing prices (and hence implied forecasts) by visiting exchanges' websites. (Some are listed at the end of this chapter.)

---

<sup>1</sup>But much of the evidence looks good, a point to which we will return in some detail.

### Density Forecasts From Options Markets (Using Sets of Options)

We can infer market-based density forecasts of future spot price  $S$  by looking at currently-prevailing options prices across a (hopefully-wide) wide range of strike prices,  $k$ .

### Event Probability Forecasts From Digital Options Markets

“Contingent claims,” or “Arrow-Debreu securities,” or “binary options,” or “digital options” simply pay \$1 if a certain event occurs, and 0 otherwise. Hence we can infer the market’s event probability assessment from the price at which the digital option sells. Digital options are now written on a variety of “underlyings,” from the  $S&P$  500, to the  $VIX$ , to the weather.

But digital options can be written on anything and traded by anyone (if only the regulators would stay away). Effectively they’re just gambles, in a financial-market disguise.<sup>2</sup> This brings up the general idea of so-called “prediction markets,” which have always been viewed as gambling markets (e.g., sports betting), to which we now turn.

### Density Forecasts From Digital Options Markets (Using Sets of Digital Options)

Estimate sets of probabilities (i.e., a density or cdf) using sets of contracts.

#### 13.1.2 More

##### Volatility Forecasts From Options Markets

Using no-arbitrage arguments (i.e., not even requiring risk neutrality), we can price options given a view on volatility, and conversely we can use market prices of options to infer the market volatility view. The famous Black-Scholes formula for pricing European options, although surely incorrect, as it

---

<sup>2</sup>Of course all financial markets are effectively casinos in significant part.

assumes that spot prices follow Gaussian random walks with constant volatility, nevertheless conveys all of the relevant lessons. We have

$$P_t = G(\sigma_t, i_t, S_t, k, \tau),$$

where  $P_t$  is a call price,  $\sigma_t$  is volatility,  $i_t$  is the risk-free rate corresponding to the remaining lifespan of the option,  $S_t$  is current spot price,  $k$  is strike price, and  $\tau$  is time to maturity. Alternatively we can invert the equation and write

$$\sigma_t = G^{-1}(P_t, i_t, S_t, k, \tau).$$

This equation gives the current market view (forecast) of  $\sigma_t$  as a function of observed market price  $P_t$ .

### Correlation Forecasts From Trios of Implied Volatilities

By a no-arbitrage argument (i.e., not even requiring risk neutrality), we have

$$\text{cov}(\Delta \ln Y/\$, \Delta \ln D/\$) = \frac{1}{2} (\text{var}(\Delta \ln Y/\$) + \text{var}(\Delta \ln D/\$) - \text{var}(\Delta \ln Y/D)).$$

To see why, note that in the absence of triangular arbitrage,

$$\text{var}(\Delta \ln(Y/D)) = \text{var}\left(\Delta \ln \frac{Y/\$}{D/\$}\right).$$

But

$$\text{var}\left(\Delta \ln \frac{Y/\$}{D/\$}\right) = \text{var}(\Delta \ln Y/\$) + \text{var}(\Delta \ln D/\$) - 2\text{cov}(\Delta \ln Y/\$, \Delta \ln D/\$),$$

so that

$$\text{cov}(\Delta \ln Y/\$, \Delta \ln D/\$) = \frac{1}{2} \left( \text{var}(\Delta \ln Y/\$) + \text{var}(\Delta \ln D/\$) - \text{var}\left(\Delta \ln \frac{Y/\$}{D/\$}\right) \right).$$

This means that, given exchange-rate volatility forecasts extracted from financial markets via options as discussed above, we can also produce market-

based covariance and correlation forecasts.

### Skewness Forecasts From Risk Reversals

In a risk reversal, one buys/sells a call and sells/buys a put, both out of the money.

### Inflation Forecasts From Indexed Bonds

The difference between yields on non-indexed vs. indexed bonds is an immediate risk-neutral forecast of inflation.

### Inflation Forecasts from Bond Yields

Under risk neutrality, nominal government bond yields equal real yields plus expected inflation (the famous “Fisher equation”),

$$i_t(t+h) = r_t(t+h) + E_t(\pi_{t+h}),$$

where  $i_t(t+h)$  is the nominal bond yield from time  $t$  to time  $t+h$ ,  $r_t(t+h)$  is the corresponding real yield, and  $E_t(\pi_{t,t+h})$  the optimal (conditional mean) forecast of inflation between time  $t$  and time  $t+h$ . Hence under an assumption about the real rate one can extract expected inflation.

### Bond Yield Forecasts From the Term Premium

Long rates always involve averages of expected future short rates. Under risk neutrality we get the famous Hicksian “expectations theory” of the yield curve, in which long rates are *precisely* averages of expected future short rates, so that borrowers are indifferent between issuing a long bond or issuing a short bond and sequentially rolling it over.<sup>3</sup>

---

<sup>3</sup>Put differently (for bond market aficionados), another way to state the expectations theory is that currently-prevailing forward interest rates should equal expected future short interest rates.

We have

$$i_t(t+h) = \frac{i_t(t+1) + E_t i_{t+1}(t+2) + \dots + E_t i_{t+h-1}(t+h)}{h}.$$

This suggests that a useful predictive regression would relate changes in short yields to the currently-prevailing long-short spread (“term spread”), a so-called Campbell-Shiller regression.

### **Real Activity Forecasts From the Term Premium**

Unexpectedly tight monetary policy now, by raising short rates, produces an inverted (negatively-sloped) yield curve now, and recession often follows (both theoretically and empirically). Conversely, loose monetary policy now produces an upward-sloping yield curve now, and a boom later. This suggests that the shape of the yield curve now, and in particular a long-short term spread, has predictive content for real activity (real GDP and its components, and more generally “the business cycle.”)

### **Real Activity Forecasts From the Default Premium**

A simple direct argument regarding market-perceived recession risk suggests that we compare the prevailing yield on an  $N$ -year (“risk-free”) government bond to an index of  $N$ -year (risky, defaultable) corporate bond yields.<sup>4</sup> The larger the spread, the larger the market-perceived corporate bond default probability, presumably driven by an increase in market-perceived recession probability.

---

<sup>4</sup>The corporate bond yield index can moreover be broken down by grade.

## Long-Run Equity Return Forecasts from the Dividend Yield

### 13.2 “Prediction Markets”

#### 13.2.1 Arrow-Debreu Contingent Claims

The name “prediction markets” is a misnomer, as predictions are not traded in those markets; rather, Arrow-Debreu contracts are traded. Hence there’s really nothing new relative to digital options. As with digital options, we interpret the prices in markets for Arrow-Debreu securities as market-based combined forecasts of event probabilities.

But prediction markets are run purely for the purpose of inferring predictions, so we don’t simply have to take markets “as is,” as with financial markets. Instead, we *design* the markets to provide exactly what we want.

Prediction markets are proving useful in many forecasting situations, and they may be unusually useful in the very hardest forecasting situations, such as assessing the probability of events like “an earthquake hits Tehran before December 31, 2050 and kills 100,000 or more people.”

#### 13.2.2 Parimutual Betting Markets

Parimutual is like Arrow-Debreu but without the ability to resell a security once bought. So it suffers in terms of dynamic tracking of market-based probabilities. A way to fix it would be to have a secondary market in pari “receipts”.

### 13.3 Issues with Market-Based Forecasts

There are many interesting issues yet to be thoroughly explored.

### 13.3.1 Market Inefficiencies and No-Arbitrage Conditions

Issues may arise with composite bets (e.g., Duke wins it all in March Madness). In particular if the market is arbitrage-free, then the price of the composite bet should equal that of a replicating portfolio of simple bets. Another no-arbitrage issue is that the bid price on one exchange should never be higher than the ask price on another.

A related issue is apparent mis-pricing of extreme events, such as overpricing of far out-of-the-money puts. This is often called “favorite / longshot bias,” in reference to parimutual betting markets.

### 13.3.2 Moral Hazard and Market Manipulation

Moral hazard – the temptation to be less vigilant against risks that are insured – is always an issue in “insurance” markets such as those under discussion. Related, incentives arise to manipulate markets so as to increase the likelihood of payoffs.

### 13.3.3 True Moral Issues

The public sometimes finds trading on extreme events immoral (e.g., “we shouldn’t let someone profit from an earthquake that kills 100,000 people”). But the profiting hedgers are precisely those that need help – earthquake victims who bought earthquake contracts!

### 13.3.4 Risk Neutrality

The main issue is that market-assessed probabilities are risk neutral, so they may not be reliable guides to the physical world. Perhaps market can behave as risk neutral even if individual agents are not. Many of the bets are for small entertainment purposes, so risk neutrality may not be unreasonable. At any

rate, as an empirical matter, event probabilities extracted from prediction markets are often highly accurate.

### 13.3.5 Beyond Risk Neutrality

Charles Manski argues that there are problems with market-assessed probabilities even if traders are risk-neutral, as long as they have heterogeneous beliefs. In particular, he argues that

... the price of a contract in a prediction market reveals nothing about the dispersion of traders' beliefs and only partially identifies the central tendency of beliefs. Most persons have beliefs higher than price when price is above 0.5, and most have beliefs lower than price when price is below 0.5. The mean belief of traders lies in an interval whose midpoint is the equilibrium price.

The first part of Manski's critique (that price reveals nothing about the dispersion of traders' beliefs) seems disingenuous. Why would anyone think that price *would* reveal anything about dispersion of beliefs?

The second part of Manski's critique (that price only partially identifies the central tendency of beliefs) seems relevant, if not particularly trenchant. Indeed he shows that the mean belief of traders nevertheless "lies in an interval whose midpoint is the equilibrium price." And as an empirical matter, market-based probability assessments are typically highly accurate, for whatever reason. For a broad overview of these and related issues, see [Snowberg et al. \(2013\)](#).

### 13.3.6 A Bit More on Market Efficiency

We have seen that groupthink may wreak havoc in certain survey environments (Delphi, focus-group), but we hasten to add that it can similarly pollute market-based probability assessments, as price bubbles and the like

(“panics, manias and crashes,” in the colorful language of Kindleberger and Aliber (2011)) may seriously compromise the independence of the opinions being aggregated.

In addition, if surveys may suffer from the fact that “no money is on the line,” markets may suffer from selection bias – markets aggregate only the views of those who choose to trade, and traders may be a very special group with special characteristics, whereas randomized surveys cover entire populations.

### 13.4 Exercises, Problems and Complements

1. “Parimutual” market-based forecasts inside Intel.

Read Gillen, Plott and Shum (2014) (GPS), “A Parimutual-Like Mechanism for Information Aggregation: A Field Test Inside Intel”.

- (a) GPS aggregate information (combine forecasts) using a “parimutual betting market” as opposed to an Arrow-Debreu (AD) securities market. Discuss the similarities and differences, pros and cons, etc. Clearly the GPS parimutual market-based information aggregation mechanism is *different* from the AD mechanism, but is it necessarily *better*? Why or why not?
- (b) Why do GPS reveal the bet distribution in real time? Might that not promote groupthink? Discuss.
- (c) GPS admirably try to provide new insight into why parimutual prediction markets work. But isn’t it basically the usual story, namely that people in prediction markets behave in approximately risk-neutral fashion, for whatever reason, allowing us to infer market-assessed conditional event probabilities?
- (d) Recall Manski’s critique of AD markets: Even under risk neutral-

ity AD markets identify only a range for the conditional probability (centered at the true conditional probability). Perhaps the parimutual mechanism has provably better properties under risk neutrality, nailing the conditional expectation as opposed to a range? Discuss.

- (e) Density forecasts are often evaluated using the sequence of probability integral transforms (PIT's). What is the PIT sequence, what two properties should it have, and why?
  - (f) Do GPS check the two PIT sequence conditions for their parimutual density forecast? One? None? Discuss in detail.
  - (g) Density forecasts are often compared using predictive likelihood (PL). What is the PL, and how do such comparisons proceed? Do GPS do a PL comparison of their parimutual forecast to the official Intel forecast? Why or why not?
  - (h) GPS compare parimutual forecasts to official Intel forecasts, but they neglect a key comparison parimutual vs. AD. How would you do it?
  - (i) In a forecast combination exercise (GPS vs. official Intel), Intel receives a *negative* combining weight. Discuss.
2. Comparing parimutual and AD information aggregation mechanisms.

<http://authors.library.caltech.edu/44358/1/wp1131.pdf>

3. Are combined prediction markets likely valuable?

PredictWise aggregates prices from alternative prediction markets. But prediction market forecasts are effectively combined forecasts, so averages of different prediction markets are effectively combined combined forecasts. Are such averages likely to be superior to any single prediction market? And isn't the existence of different prices for the same contract in different prediction markets a violation of the law of one price, and

hence of market efficiency, enabling arbitrage? If the answer is that the contracts in different markets aren't identical, then does it really make sense to average their prices?

4. A diary of experiences trading in prediction markets (among other things).

See the material under “STATISTICS/Predicting and Prediction Markets” at [gwern.net](#). Much other material on the site is interesting and also worth a look.

5. Interesting people working on prediction markets.

Miro Dudik, Dan Goldstein, Jake Hofman, Patrick Hummel, Adam Isen, Neil Malhotra, David Pennock, David Rothchild, Florian Teschner, Duncan Watts, Justin Wolfers, [Eric Zitzewitz](#). For links to names not hyperlinked, see [Rothchild's site](#).

6. Prediction markets encourage “foxy” behavior.

Read [Silver \(2012\)](#), [Tetlock and Gardner \(2015\)](#) and [Tetlock \(2006\)](#) on forecasting “foxes” and “hedgehogs” (e.g., Silver, pp. 53-54). Note the key to the success of prediction markets – particularly as typically used for complicated event forecasting – may be their encouragement of foxy behavior, by virtue of their making non-foxy (hedgehog) behavior explicitly *costly!* Note also, however, that it’s important that prediction-market wagers be set at levels not so small as to encourage strong risk-seeking, or so high as to encourage strong risk aversion. We need approximate risk neutrality to be able to credibly interpret prediction-market prices as probabilities.

## 13.5 Notes

### Financial Markets

**Macroeconomic derivatives.** CME futures and options on the S&P Case-Shiller house price index.

**VIX implied volatility index**

**CBOE options exchange**

**CME futures exchange**

### Prediction Markets

**ipredict.** New Zealand. Wide range of contracts.

**Tradesports.** In trouble. Re-emerging with virtual currency?

### **The Journal of Prediction Markets**

**Lumenologic.** “Collective intelligence solutions.” Consultancy providing prediction markets to client firms. Uses virtual currency.

**European Betting Markets**

**Microsoft Prediction Lab**

**PredictWise** Prediction/Betting Market Aggregator. Combined prediction-market forecasts.

From their FAQ’s (regarding the markets that they follow / combine):

Q:What are the Iowa Electronic Markets, <http://tippie.uiowa.edu/iem/index.cfm>?

A:The Iowa Electronic Markets (IEM) is an exchange of real-money prediction markets operated by the University of Iowa Tippie College of Business. The IEM is not-for-profit; the markets are run for educational and research purposes. Because of the small sums wagered and the academic focus, the IEM has received no-action relief from the United States government, meaning U.S.-based speculators can legally risk up to \$500 on the exchange.

Q:What is Betfair, <https://www.betfair.com/us>?

A:Betfair, based in the United Kingdom, is the world’s largest internet betting exchange. Rather than having a bookmaker create odds, the odds for every bet are determined by the market of bettors, working similarly to a stock market. Bettors can either ”Back” (buy) or ”Lay” (sell) a given bet at

certain odds, and the odds move as Backs and Lays are matched. Betfair is legal in the UK and other countries, but it is illegal to bet money on Betfair as a resident of the United States.

Q:What is Intrade, <https://prev.intrade.com/v4/home/>?

A:Intrade, based in Ireland, is a prediction market which allows individuals to trade contracts on whether future events will or will not occur. For any given contract, the value at expiration is either 100 (if the event happens) or 0 (if it does not). Contracts therefore trade between 0 and 100 at all times, with the price representing the market's prediction for the likelihood of that event. Intrade is legal in the Republic of Ireland and other countries, but it is illegal to bet money on Intrade as a resident of the United States.

Q:What is HuffPost Pollster, <http://www.huffingtonpost.com/news/pollster/>?

A:HuffPost Pollster, is a site that discusses and aggregates polling data. Polling data is subject to random fluctuations and Pollster's aggregation methods cleanly and transparently aggregate polls over time to provide a more meaningful snapshot of where the polls are at any given moment.

Q:What is PredictIt, <https://www.predictit.org/>?

A:PredictIt is an exchange of real-money prediction markets operated by the Victoria University and Aristotle. Because of the small sums wagered PredictIt has received no-action relief from the United States government, meaning U.S.-based speculators can legally risk upwards of \$850 in any of the markets.

Q:What is BETDAQ, <https://www.betdaq.com/Default.aspx>?

A:BETDAQ, based in Ireland, is an internet betting exchange. Rather than having a bookmaker create odds, the odds for every bet are determined by the market of bettors, working similarly to a stock market. Bettors can either "Back" (buy) or "Lay" (sell) a given bet at certain odds, and the odds move as Backs and Lays are matched. BETDAQ is legal in Ireland and other

countries, but it is illegal to bet money on BETDAQ as a resident of the United States.

Q:What is the Hollywood Stock Exchange (HSX), <http://www.hsx.com/>?

A:The Hollywood Stock Exchange (HSX) is a play-money prediction market in which users can buy or sell shares in movies, actors, directors, and other Hollywood-related topics. For example, users can buy or sell shares of an upcoming film as a means predicting how well that film will do at the box office in its first four weekends of wide release, and then be ranked based on the accuracy of their predictions. Because HSX involves only simulated money, it is legal for all participants.

Q:What is Smarkets, <https://smarkets.com/>?

A:Smarkets, based in the United Kingdom, is an internet betting exchange. Rather than having a bookmaker create odds, the odds for every bet are determined by the market of bettors, working similarly to a stock market. Bettors can either bet "For" (buy) or "Against" (sell) a given bet at certain odds, and the odds move as Fors and Againsts are matched. Smarkets is legal in the UK and other countries, but it is illegal to bet money on Smarkets as a resident of the United States.



# Chapter 14

## Survey-Based Forecast Combination

### 14.1 Survey-Based Point Forecast Combination

A number of groups regularly survey economic and financial forecasters and publish “consensus” forecasts, typically the mean or median – essentially the average! – of the forecasters surveyed. (The median has some desirable robustness to outliers.) The consensus forecasts often perform very well relative to the individual forecasts.

The Survey of Professional Forecasters (SPF) is the leading U.S. consensus macroeconomic forecast. It has been produced each quarter since the late 1960s; currently it is produced by the Federal Reserve Bank of Philadelphia. A similar Survey of Professional Forecasters for Europe has been produced each quarter since the late 1990s; it is produced by the European Central Bank.

Another leading U.S. consensus forecast is the Livingston Survey, which is now also maintained by the Federal Reserve Bank of Philadelphia. It is only bi-annual but has been recorded for more than half a century. There are also many surveys done in the private sector.

### 14.1.1 Surveys and the Wisdom of Crowds

As emphasized in Surowiecki's *Wisdom of Crowds* (Surowiecki (2004)), wise "crowdsourcing" depends on balanced aggregation across disparate information sources. So we need: (1) independent, or at least imperfectly dependent, people, so that there's actually something to aggregate, and (2) a dispassionate aggregation mechanism, so that we avoid "groupthink." Surveys are often good at (1), and certainly they're very good at (2). Other more exotic dispassionate aggregation mechanisms include Google's "pagerank" algorithm and open-source software coding (e.g., Linux).

### 14.1.2 Delphi, Focus Groups, and Related Methods

The "Delphi method" is a forecasting technique that sometimes proves useful in very difficult forecasting situations not amenable to quantification, such as new-technology forecasting. The basic idea is to survey a panel of experts anonymously, reveal the distribution of opinions to the experts so they can revise their opinions, repeat the survey, and so on. Typically the diversity of opinion is reduced as the iterations proceed. However, Delphi may be problematic insofar as it is actually *based* on groupthink. "Focus groups" maybe even worse, as certain individuals may dominate the group. At the same time, it's not clear that we should *dispense* with such techniques; they may be of some value.

### 14.1.3 Cross-Sectional Forecast Dispersion vs. True Uncertainty

The two are very different, and in principle unrelated, even if they are often positively correlated in practice.

In particular, the cross-sectional distribution of survey point forecasts is *not* a density forecast, combined or otherwise. Rather, it's simply the cross-sectional distribution of survey point forecasts. Density forecasts, combined

or otherwise, cannot generally be obtained from surveys of point forecasts. For that, we need a survey density, not point, forecast from each participant, from which a combined survey density forecast may be constructed.<sup>1</sup> (See Chapter \*\*\*).

## 14.2 Survey-Based Density Forecast Combination

### 14.3 Exercises, Problems and Complements

#### 1. Wiki Surveys.

See [Wiki Surveys](#). It's not really a survey; rather, it's idea generation by pairwise comparison. Very interesting and evidently useful, even if naive in its methods for reconstructing preferences from pairwise rankings.

#### 2. Issues in survey design.

##### (a) Time series of cross sections vs. panels.

Both the SPF and the Livingston Survey are time series of cross sections as opposed to true panels, insofar as both the number and composition of underlying forecasters has evolved over time. Other surveys like the Panel Study of Income Dynamics have true panel structure. Panel structure is preferable when possible, but it's not always possible, as with the SPF.

##### (b) Framing survey questions to turn individual responses into combined forecasts.

[Rothchild and Wolfers \(2013\)](#) makes the interesting observation that election surveys are more accurate when the respondents are not asked for whom they intend to vote, but rather whom they expect to win. A natural interpretation is that each response to the latter

---

<sup>1</sup>There are, however, conditions under which the cross-sectional distribution of point forecasts can be interpreted as a density forecast. See \*\*\*.

survey is actually an average response over the respondent’s friends, thereby making the effective sample size much larger than the nominal size  $N$  (more like  $10N$ , say). That is, each response in the latter survey is not a forecast, but rather a combined forecast.

### 3. Using surveys to assess forecastability.

As we saw earlier in Chapter \*\*\*, forecastability involves comparing estimates of “best” forecast accuracy to “naive” forecast accuracy. The question arises as to what to use for the best forecast. A strong case can be made for using a combined forecast from a survey or market. For example, to assess GDP forecastability, we might use MSE (assuming quadratic loss) from a reputable survey of professional forecasters as “best forecast” accuracy, and a historical GDP sample variance as “naive forecast” accuracy.

### 4. Forecastability assessment using surveys.

One could take a *survey-based* approach, based on the predictions of competitive professional forecasters. Conditional upon the assumption that the reported forecasts are optimal, those data can be used for inferences about predictability. The survey-based approach is of interest because the information sets used by actual forecasters are likely much richer than simple univariate histories. They are surely multivariate, for example, and they also contain hard-to-quantify subjective information. The survey-based approach does rely on a crucial and disputable assumption (optimality of reported forecasts), but so too does the model-based approach (adequacy of the fitted model). The key point is that the assumptions made by the two approaches are different, and that the approaches therefore naturally complement one another.

A number of relevant surveys exist, including the former Survey of Professional Forecasters by the Federal Reserve Bank of Philadelphia (see

Croushore 1993). These surveys focus on the major macroeconomic aggregates, such as real GDP growth. It would be interesting to use those forecasts to compute survey-based estimates of predictability, and to compare the survey-based and model-based estimates.

## 5. Adaptive crowdsourcing.

Sometimes forecast combination, particularly when done with surveys or markets, is called “crowdsourcing.” Sometimes “adaptive crowdsourcing” is appealing.

- (a) Traditional forecast combination, but with time-varying combining weights, is a form of adaptive crowdsourcing.
- (b) Another example is the judging of a science fair, with judges re-allocated as various projects are eliminated.

## 14.4 Notes

Useful web sites:

### Macro / Finance Surveys

U.S. Survey of Professional Forecasters (SPF). Quarterly.

European Survey of Professional Forecasters.

Livingston Survey. Goes way back, spanning many business cycles. Biennial.

Blue Chip

### Micro Surveys

Panel Study of Income Dynamics.

Michigan Survey of Consumer Sentiment. Well-known indexes of Consumer Sentiment, Current Economic Conditions, and Consumer Expectations.

### Companies

**Consensus Economics.** Private-sector forecast combination.

**Blue Chip**

Other

**Wiki Surveys**

**HuffPost Pollster** discusses and aggregates polling data. Combined surveys!

**Part V**

**More**



# Chapter 15

## Selection, Shrinkage, and Distillation

We start with more on selection (“hard threshold” – variables are either kept or discarded), and then we introduce shrinkage (“soft threshold” – all variables are kept, but parameter estimates are coaxed in a certain direction), and then lasso, which blends selection and shrinkage.

### 15.1 All-Subsets Model Selection I: Information Criteria

All-subsets model selection means that we examine every possible combination of  $K$  regressors and select the best. Examples include *SIC* and *AIC*.

Let us now discuss *SIC* and *AIC* in greater depth, as they are tremendously important tools for building forecasting models. We often could fit a wide variety of forecasting models, but how do we select among them? What are the consequences, for example, of fitting a number of models and selecting the model with highest  $R^2$ ? Is there a better way? This issue of **model selection** is of tremendous importance in all of forecasting.

It turns out that model-selection strategies such as selecting the model with highest  $R^2$  do *not* produce good out-of-sample forecasting models. Fortunately, however, a number of powerful modern tools exist to assist with model selection. Most model selection criteria attempt to find the model

with the smallest out-of-sample 1-step-ahead mean squared prediction error. The criteria we examine fit this general approach; the differences among criteria amount to different penalties for the number of degrees of freedom used in estimating the model (that is, the number of parameters estimated). Because all of the criteria are effectively estimates of out-of-sample mean square prediction error, they have a negative orientation – the smaller the better.

First consider the **mean squared error**,

$$MSE = \frac{\sum_{t=1}^T e_t^2}{T},$$

where  $T$  is the sample size and  $e_t = y_t - \hat{y}_t$ .  $MSE$  is intimately related to two other diagnostic statistics routinely computed by regression software, the **sum of squared residuals** and  $R^2$ . Looking at the  $MSE$  formula reveals that the model with the smallest  $MSE$  is also the model with smallest sum of squared residuals, because scaling the sum of squared residuals by  $1/T$  doesn't change the ranking. So selecting the model with the smallest  $MSE$  is equivalent to selecting the model with the smallest sum of squared residuals. Similarly, recall the formula for  $R^2$ ,

$$R^2 = 1 - \frac{\sum_{t=1}^T e_t^2}{\sum_{t=1}^T (y_t - \bar{y})^2} = 1 - \frac{MSE}{\frac{1}{T} \sum_{t=1}^T (y_t - \bar{y})^2}.$$

The denominator of the ratio that appears in the formula is just the sum of squared deviations of  $y$  from its sample mean (the so-called “total sum of squares”), which depends only on the data, not on the particular model fit. Thus, selecting the model that minimizes the sum of squared residuals – which as we saw is equivalent to selecting the model that minimizes  $MSE$  – is also equivalent to selecting the model that maximizes  $R^2$ .

Selecting forecasting models on the basis of  $MSE$  or any of the equivalent forms discussed above – that is, using in-sample  $MSE$  to estimate the out-of-sample 1-step-ahead  $MSE$  – turns out to be a bad idea. In-sample

MSE *can't* rise when more variables are added to a model, and typically it will fall continuously as more variables are added, because the estimated parameters are explicitly chosen to *minimize* the sum of squared residuals. Newly-included variables could get estimated coefficients of zero, but that's a probability-zero event, and to the extent that the estimate is anything else, the sum of squared residuals must fall. Thus, the more variables we include in a forecasting model, the lower the sum of squared residuals will be, and therefore the lower  $MSE$  will be, and the higher  $R^2$  will be. Again, the sum of squared residuals can't rise, and due to sampling error it's very unlikely that we'd get a coefficient of exactly zero on a newly-included variable even if the coefficient is zero in population.

The effects described above go under various names, including **in-sample overfitting**, reflecting the idea that including more variables in a forecasting model won't necessarily improve its out-of-sample forecasting performance, although it will improve the model's "fit" on historical data. The upshot is that in-sample  $MSE$  is a downward biased estimator of out-of-sample  $MSE$ , and the size of the bias increases with the number of variables included in the model. In-sample  $MSE$  provides an overly-optimistic (that is, too small) assessment of out-of-sample  $MSE$ .

To reduce the bias associated with  $MSE$  and its relatives, we need to penalize for degrees of freedom used. Thus let's consider the mean squared error corrected for degrees of freedom,

$$s^2 = \frac{\sum_{t=1}^T e_t^2}{T - K},$$

where  $K$  is the number of degrees of freedom used in model fitting.<sup>1</sup>  $s^2$  is just the usual unbiased estimate of the regression disturbance variance. That is, it is the square of the usual standard error of the regression. So selecting the model that minimizes  $s^2$  is equivalent to selecting the model that minimizes

---

<sup>1</sup>The degrees of freedom used in model fitting is simply the number of parameters estimated.

the standard error of the regression.  $s^2$  is also intimately connected to the  $R^2$  adjusted for degrees of freedom (the “**adjusted  $R^2$** ,” or  $\bar{R}^2$ ). Recall that

$$\bar{R}^2 = 1 - \frac{\sum_{t=1}^T e_t^2 / (T - K)}{\sum_{t=1}^T (y_t - \bar{y})^2 / (T - 1)} = 1 - \frac{s^2}{\sum_{t=1}^T (y_t - \bar{y})^2 / (T - 1)}.$$

The denominator of the  $\bar{R}^2$  expression depends only on the data, not the particular model fit, so the model that minimizes  $s^2$  is also the model that maximizes  $\bar{R}^2$ . In short, the strategies of selecting the model that minimizes  $s^2$ , or the model that minimizes the standard error of the regression, or the model that maximizes  $\bar{R}^2$ , are equivalent, and they do penalize for degrees of freedom used.

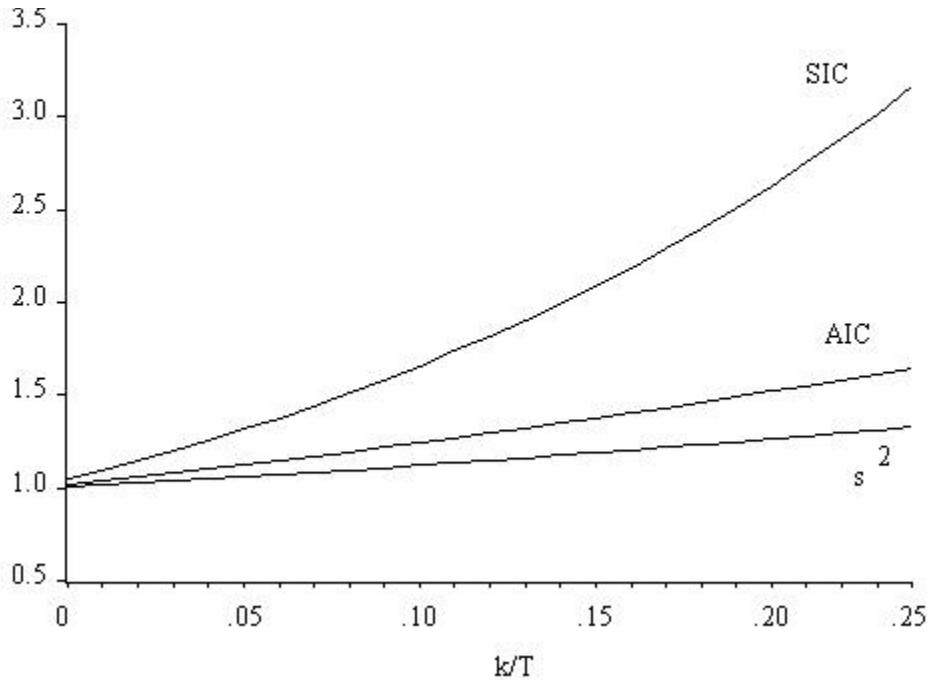
To highlight the degree-of-freedom penalty, let’s rewrite  $s^2$  as a penalty factor times the  $MSE$ ,

$$s^2 = \left( \frac{T}{T - K} \right) \frac{\sum_{t=1}^T e_t^2}{T}.$$

Note in particular that including more variables in a regression will not necessarily lower  $s^2$  or raise  $\bar{R}^2$  – the  $MSE$  will fall, but the degrees-of-freedom penalty will rise, so the product could go either way.

As with  $s^2$ , many of the most important forecast model selection criteria are of the form “penalty factor times  $MSE$ .” The idea is simply that if we want to get an accurate estimate of the 1-step-ahead out-of-sample forecast  $MSE$ , we need to penalize the in-sample residual  $MSE$  to reflect the degrees of freedom used. Two very important such criteria are the **Akaike Information Criterion (AIC)** and the **Schwarz Information Criterion (SIC)**. Their formulas are:

$$AIC = e^{(\frac{2K}{T})} \frac{\sum_{t=1}^T e_t^2}{T}$$



and

$$SIC = T^{(\frac{K}{T})} \frac{\sum_{t=1}^T e_t^2}{T}.$$

How do the penalty factors associated with  $MSE$ ,  $s^2$ ,  $AIC$  and  $SIC$  compare in terms of severity? All of the penalty factors are functions of  $K/T$ , the number of parameters estimated per sample observation, and we can compare the penalty factors graphically as  $K/T$  varies. In Figure \*\*\* we show the penalties as  $K/T$  moves from 0 to .25, for a sample size of  $T = 100$ . The  $s^2$  penalty is small and rises slowly with  $K/T$ ; the  $AIC$  penalty is a bit larger and still rises only slowly with  $K/T$ . The  $SIC$  penalty, on the other hand, is substantially larger and rises much more quickly with  $K/T$ .

It's clear that the different criteria penalize degrees of freedom differently. In addition, we could propose many other criteria by altering the penalty. How, then, do we select among the criteria? More generally, what properties might we expect a "good" model selection criterion to have? Are  $s^2$ ,  $AIC$  and  $SIC$  "good" model selection criteria?

We evaluate model selection criteria in terms of a key property called **consistency**, also known as the **oracle property**. A model selection criterion is consistent if:

- a. when the true model (that is, the **data-generating process, or DGP**) is among a fixed set models considered, the probability of selecting the true DGP approaches one as the sample size gets large, and
- b. when the true model is *not* among a fixed set of models considered, so that it's impossible to select the true DGP, the probability of selecting the best *approximation* to the true DGP approaches one as the sample size gets large.

We must of course define what we mean by “best approximation” above. Most model selection criteria – including all of those discussed here – assess goodness of approximation in terms of out-of-sample mean squared forecast error.

Consistency is of course desirable. If the DGP is among those considered, then we'd hope that as the sample size gets large we'd eventually select it. Of course, all of our models are false – they're intentional simplifications of a much more complex reality. Thus the second notion of consistency is the more compelling.

$MSE$  is inconsistent, because it doesn't penalize for degrees of freedom; that's why it's unattractive.  $s^2$  does penalize for degrees of freedom, but as it turns out, not enough to render it a consistent model selection procedure. The  $AIC$  penalizes degrees of freedom more heavily than  $s^2$ , but it too remains inconsistent; even as the sample size gets large, the  $AIC$  selects models that are too large (“overparameterized”). The  $SIC$ , which penalizes degrees of freedom most heavily, *is* consistent.

The discussion thus far conveys the impression that  $SIC$  is unambiguously superior to  $AIC$  for selecting forecasting models, but such is not the

case. Until now, we've implicitly assumed a fixed set of models. In that case, *SIC* is a superior model selection criterion. However, a potentially more compelling thought experiment for forecasting may be that we may want to expand the set of models we entertain as the sample size grows, to get progressively better approximations to the elusive DGP. We're then led to a different optimality property, called **asymptotic efficiency**. An asymptotically efficient model selection criterion chooses a sequence of models, as the sample size get large, whose out-of-sample forecast MSE approaches the one that would be obtained using the DGP at a rate at least as fast as that of any other model selection criterion. The *AIC*, although inconsistent, is asymptotically efficient, whereas the *SIC* is not.

In practical forecasting we usually report and examine both *AIC* and *SIC*. Most often they select the same model. When they don't, and despite the theoretical asymptotic efficiency property of *AIC*, this author recommends use of the more parsimonious model selected by the *SIC*, other things equal. This accords with the parsimony principle of Chapter 2 and with the results of studies comparing out-of-sample forecasting performance of models selected by various criteria.

The *AIC* and *SIC* have enjoyed widespread popularity, but they are not universally applicable, and we're still learning about their performance in specific situations. However, the general principle that we need somehow to inflate in-sample loss estimates to get good out-of-sample loss estimates *is* universally applicable.

The versions of *AIC* and *SIC* introduced above – and the claimed optimality properties in terms of out-of-sample forecast MSE – are actually specialized to the Gaussian case, which is why they are written in terms of minimized *SSR*'s rather than maximized *lnL*'s.<sup>2</sup> More generally, *AIC* and *SIC* are written not in terms of minimized *SSR*'s, but rather in terms of

---

<sup>2</sup>Recall that in the Gaussian case *SSR* minimization and *lnL* maximization are equivalent.

maximized  $\ln L$ 's. We have:

$$AIC = -2\ln L + 2K$$

and

$$SIC = -2\ln L + K\ln T.$$

These are useful for any model estimated by maximum likelihood, Gaussian or non-Gaussian.

## 15.2 All-Subsets Model Selection II: Cross Validation

Cross validation (CV) proceeds as follows. Consider selecting among  $J$  models. Start with model 1, estimate it using all data observations except the first, use it to predict the first observation, and compute the associated squared prediction error. Then estimate it using all observations except the second, use it to predict the second observation, and compute the associated squared error. Keep doing this – estimating the model with one observation deleted and then using the estimated model to predict the deleted observation – until each observation has been sequentially deleted, and average the squared errors in predicting each of the  $T$  sequentially deleted observations. Repeat the procedure for the other models,  $j = 2, \dots, J$ , and select the model with the smallest average squared prediction error.

Actually this is “ $T$ -fold” CV, because we split the data into  $T$  parts (the  $T$  individual observations) and predict each of them. More generally we can split the data into  $M$  parts ( $M < T$ ) and cross validate on them (“ $M$ -fold” CV). As  $M$  falls,  $M$ -fold CV eventually becomes consistent.  $M = 10$  often works well in practice.

It is instructive to compare SIC and CV, both of which have the oracle property. SIC achieves it by penalizing in-sample residual MSE to obtain an approximately-unbiased estimate of out-of-sample MSE. CV, in contrast,

achieves it by directly obtaining an unbiased estimated out-of-sample MSE.

CV is more general than information criteria insofar as it can be used even when the model degrees of freedom is unclear. In addition, non-quadratic loss can be introduced easily. Generalizations to time-series contexts are available.

## 15.3 Stepwise Selection

All-subsets selection, whether by AIC, SIC or CV, quickly gets hard as there are  $2^K$  subsets of  $K$  regressors. Other procedures, like the stepwise selection procedures that we now introduce, don't explore every possible subset. They are more ad hoc but very useful.

### 15.3.1 Forward

Algorithm:

- Begin regressing only on an intercept
- Move to a one-regressor model by including that variable with the smallest t-stat  $p$ -value
- Move to a two-regressor model by including that variable with the smallest  $p$ -value
- Move to a three-regressor model by including that variable with the smallest  $p$ -value

Often people use information criteria or CV to select from the stepwise sequence of models. This is a “greedy algorithm,” producing an increasing sequence of candidate models. Often people use information criteria or CV to select from the stepwise sequence of models. No guaranteed optimality properties of the selected model.

“forward stepwise regression”

- Often people use information criteria or cross validation to select from the stepwise sequence of models.

### 15.3.2 Backward

Algorithm:

- Start with a regression that includes all  $K$  variables
- Move to a  $K - 1$  variable model by dropping the variable with the largest t-stat  $p$ -value
- Move to a  $K - 2$  variable model by dropping the variable with the largest  $p$ -value

Often people use information criteria or CV to select from the stepwise sequence of models. This is a “greedy algorithm,” producing a decreasing sequence of candidate models. Often people use information criteria or CV to select from the stepwise sequence of models. No guaranteed optimality properties of the selected model.

## 15.4 One-Shot Estimation: Bayesian Shrinkage

Shrinkage is a generic feature of Bayesian estimation. The Bayes rule under quadratic loss is the posterior mean, which is a weighted average of the MLE and the prior mean,

$$\hat{\beta}_{bayes} = \omega_1 \hat{\beta}_{MLE} + \omega_2 \beta_0,$$

where the weights depend on prior precision. Hence the Bayes rule pulls, or “shrinks,” the MLE toward the prior mean.

A classic shrinkage estimator is **ridge regression**,<sup>3</sup>

$$\hat{\beta}_{ridge} = (X'X + \lambda I)^{-1} X'y.$$

$\lambda \rightarrow 0$  produces OLS, whereas  $\lambda \rightarrow \infty$  shrinks completely to 0.  $\lambda$  can be chosen by CV. (Notice that  $\lambda$  can *not* be chosen by information criteria, as  $K$  regressors are included regardless of  $\lambda$ . Hence CV is a more general

---

<sup>3</sup>The ridge regression estimator can be shown to be the posterior mean for a certain prior and likelihood.

selection procedure, useful for selecting various “tuning parameters” (like  $\lambda$ ) as opposed to just numbers of variables in hard-threshold procedures.

## 15.5 One-Shot Estimation: Selection and Shrinkage

### 15.5.1 Penalized Estimation

Consider the penalized estimator,

$$\hat{\beta}_{PEN} = \operatorname{argmin}_{\beta} \left( \sum_{t=1}^T \left( y_t - \sum_i \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^K |\beta_i|^q \right),$$

or equivalently

$$\hat{\beta}_{PEN} = \operatorname{argmin}_{\beta} \sum_{t=1}^T \left( y_t - \sum_i \beta_i x_{it} \right)^2$$

s.t.

$$\sum_{i=1}^K |\beta_i|^q \leq c.$$

Concave penalty functions non-differentiable at the origin produce selection. Smooth convex penalties produce shrinkage. Indeed one can show that taking  $q \rightarrow 0$  produces subset selection, and taking  $q = 2$  produces ridge regression. Hence penalized estimation nests those situations and includes an intermediate case ( $q = 1$ ) that produces the lasso, to which we now turn.

### 15.5.2 The Lasso

The lasso solves the L1-penalized regression problem of finding

$$\hat{\beta}_{LASSO} = \operatorname{argmin}_{\beta} \left( \sum_{t=1}^T \left( y_t - \sum_i \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^K |\beta_i| \right)$$

or equivalently

$$\hat{\beta}_{LASSO} = \operatorname{argmin}_{\beta} \sum_{t=1}^T \left( y_t - \sum_i \beta_i x_{it} \right)^2$$

s.t.

$$\sum_{i=1}^K |\beta_i| \leq c.$$

Ridge shrinks, but the lasso shrinks *and* selects. Figure ?? says it all. Notice that, like ridge and other Bayesian procedures, lasso requires only *one* estimation. And moreover, the lasso uses minimization problem is convex (lasso uses the smallest  $q$  for which it is convex), which renders the single estimation highly tractable computationally.

Lasso also has a very convenient d.f. result. The effective number of parameters is precisely the number of variables selected (number of non-zero  $\beta$ 's). This means that we can use info criteria to select among “lasso models” for various  $\lambda$ . That is, the lasso is another device for producing an “increasing” sequence of candidate models (as  $\lambda$  increases). The “best”  $\lambda$  can then be chosen by information criteria (or cross-validation, of course).

### Elastic Net

$$\hat{\beta}_{EN} = \operatorname{argmin}_{\beta} \left( \sum_{t=1}^T \left( y_t - \sum_i \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^K (\alpha |\beta_i| + (1-\alpha) \beta_i^2) \right)$$

- A mixture of Lasso and Ridge regression; that is, it combines L1 and L2 penalties.
- Unlike Lasso, it moves strongly correlated predictors in or out of the model together, hopefully producing improving prediction accuracy relative to Lasso.
- Unlike Lasso, there are two tuning parameters in the elastic net  $\lambda$  and  $\alpha$ .

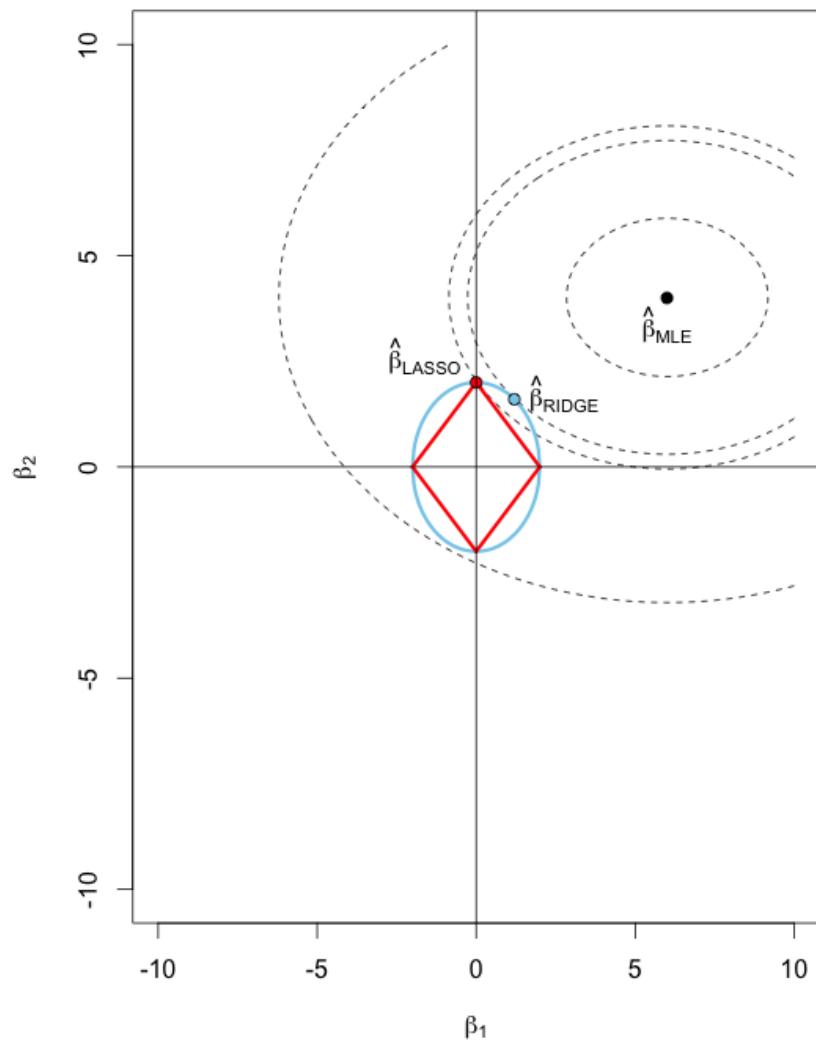


Figure 15.1: Lasso and Ridge Comparison

For  $\alpha = 1$  elastic net turns into a Lasso model, For  $\alpha = 0$  it is equivalent to ridge regression.

### Adaptive Lasso

$$\hat{\beta}_{ALASSO} = \operatorname{argmin}_{\beta} \left( \sum_{t=1}^T \left( y_t - \sum_i \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^K w_i |\beta_i| \right),$$

where  $w_i = 1/\hat{\beta}_i^\nu$ ,  $\hat{\beta}_i$  is the OLS estimate, and  $\nu > 0$ .

- Every parameter in the penalty function is weighted differently, in contrast to the “regular” Lasso.
- The weights are calculated by OLS.
- Oracle property.

### Adaptive Elastic Net

$$\hat{\beta}_{AEN} = \operatorname{argmin}_{\beta} \left( \sum_{t=1}^T \left( y_t - \sum_i \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^K w_i (\alpha |\beta_i| + (1-\alpha) \beta_i^2) \right),$$

where  $w_i = 1/\hat{\beta}_i^\nu$ ,  $\hat{\beta}_i$  is the OLS estimate, and  $\nu > 0$ .

- A combination of elastic net and adaptive Lasso.
- Oracle property.

## 15.6 Distillation: Principal Components

### 15.6.1 Distilling “ $X$ Variables” into Principal Components

Data Summarization. Think of a giant (wide)  $X$  matrix and how to “distill” it.

$X'X$  eigen-decomposition:

$$X'X = VD^2V'$$

The  $j^{th}$  column of  $V$ ,  $v_j$ , is the  $j^{th}$  eigenvector of  $X'X$

Diagonal matrix  $D^2$  contains the descending eigenvalues of  $X'X$

First principal component (PC):

$$z_1 = Xv_1$$

$$\text{var}(z_1) = d_1^2/T$$

(maximal sample variance among all possible l.c.'s of columns of  $X$ )

In general:

$$z_j = Xv_j \perp z_{j'}, j' \neq j$$

$$\text{var}(z_j) \leq d_j^2/T$$

### 15.6.2 Principal Components Regression

The idea is to enforce parsimony with little information loss by regressing not on the full  $X$ , but rather on the first few PC's of  $X$ . We speak of “Principal components regression” (PCR), or “Factor-Augmented Regression”.

Ridge regression and PCR are both shrinkage procedures involving PC's. Ridge effectively includes all PC's and shrinks according to sizes of eigenvalues associated with the PC's. PCR effectively shrinks some PCs completely to zero (those not included) and doesn't shrink others at all (those included).

## 15.7 Exercises, Problems and Complements

1. Information criteria in time-series environments.

This chapter, and hence it's discussion of information criteria, emphasizes cross-section environments. We motivated SIC and AIC in terms of out-of-sample forecast MSE. Everything goes through in time-series

environments, but in time series there is also a horizon issue. SIC and AIC are then linked to *1-step-ahead* out-of-sample forecast MSE. Modifications for multi-step time-series forecasting are also available.

## 15.8 Notes

# Chapter 16

## Multivariate: Vector Autoregression

The regression model is an explicitly multivariate model, in which variables are explained and forecast on the basis of their own history and the histories of other, related, variables. Exploiting such cross-variable linkages may lead to good and intuitive forecasting models, and to better forecasts than those obtained from univariate models.

Regression models are often called causal, or explanatory, models. For example, in the linear regression model,

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2),$$

the presumption is that  $x$  helps determine, or cause,  $y$ , not the other way around. For this reason the left-hand-side variable is sometimes called the “endogenous” variable, and the right-hand side variables are called “exogenous” or “explanatory” variables.

But ultimately regression models, like all statistical models, are models of correlation, not causation. Except in special cases, all variables are endogenous, and it’s best to admit as much from the outset. In this chapter we’ll explicitly do so; we’ll work with systems of regression equations called vector autoregressions (*VARs*).

## 16.1 Distributed Lag Models

An unconditional forecasting model like

$$y_t = \beta_0 + \delta x_{t-1} + \varepsilon_t$$

can be immediately generalized to the **distributed lag model**,

$$y_t = \beta_0 + \sum_{i=1}^{N_x} \delta_i x_{t-i} + \varepsilon_t.$$

We say that  $y$  depends on a distributed lag of past  $x$ 's. The coefficients on the lagged  $x$ 's are called lag weights, and their pattern is called the lag distribution.

One way to estimate a distributed lag model is simply to include all  $N_x$  lags of  $x$  in the regression, which can be estimated by least squares in the usual way. In many situations, however,  $N_x$  might be quite a large number, in which case we'd have to use many degrees of freedom to estimate the model, violating the parsimony principle. Often we can recover many of those degrees of freedom without seriously worsening the model's fit by constraining the lag weights to lie on a low-order polynomial. Such **polynomial distributed lags** promote smoothness in the lag distribution and may lead to sophisticatedly simple models with improved forecasting performance.

Polynomial distributed lag models are estimated by minimizing the sum of squared residuals in the usual way, subject to the constraint that the lag weights follow a low-order polynomial whose degree must be specified. Suppose, for example, that we constrain the lag weights to follow a second-degree polynomial. Then we find the parameter estimates by solving the

problem

$$\min_{\beta_0, \delta_i} \sum_{t=N_x+1}^T \left[ y_t - \beta_0 - \sum_{i=1}^{N_x} \delta_i x_{t-i} \right]^2,$$

subject to

$$\delta_i = P(i) = a + bi + ci^2, \quad i = 1, \dots, N_x.$$

This converts the estimation problem from one of estimating  $1 + N_x$  parameters,  $\beta_0, \delta_1, \dots, \delta_{N_x}$ , to one of estimating four parameters,  $\beta_0$ ,  $a$ ,  $b$  and  $c$ . Sometimes additional constraints are imposed on the shape of the polynomial, such as  $P(N_x) = 0$ , which enforces the idea that the dynamics have been exhausted by lag  $N_x$ .

Polynomial distributed lags produce aesthetically appealing, but basically ad hoc, lag distributions. After all, why should the lag weights necessarily follow a low-order polynomial? An alternative and often preferable approach makes use of the **rational distributed lags** that we introduced in Chapter 7 in the context of univariate *ARMA* modeling. Rational distributed lags promote parsimony, and hence smoothness in the lag distribution, but they do so in a way that's potentially much less restrictive than requiring the lag weights to follow a low-order polynomial. We might, for example, use a model like

$$y_t = \frac{A(L)}{B(L)} x_t + \varepsilon_t,$$

where  $A(L)$  and  $B(L)$  are low-order polynomials in the lag operator. Equivalently, we can write

$$B(L)y_t = A(L)x_t + B(L)\varepsilon_t,$$

which emphasizes that the rational distributed lag of  $x$  actually brings both lags of  $x$  and lags of  $y$  into the model. One way or another, it's crucial to allow for lags of  $y$ , and we now study such models in greater depth.

## 16.2 Regressions with Lagged Dependent Variables, and Regressions with *ARMA* Disturbances

There's something missing in distributed lag models of the form

$$y_t = \beta_0 + \sum_{i=1}^{N_x} \delta_i x_{t-i} + \varepsilon_t.$$

A multivariate model (in this case, a regression model) should relate the current value  $y$  to its own past and to the past of  $x$ . But as presently written, we've left out the past of  $y$ ! Even in distributed lag models, we always want to allow for the presence of the usual univariate dynamics. Put differently, the included regressors may not capture all the dynamics in  $y$ , which we need to model one way or another. Thus, for example, a preferable model includes lags of the dependent variable,

$$y_t = \beta_0 + \sum_{i=1}^{N_y} \alpha_i y_{t-i} + \sum_{j=1}^{N_x} \delta_j x_{t-j} + \varepsilon_t.$$

This model, a **distributed lag regression model with lagged dependent variables**, is closely related to, but not exactly the same as, the rational distributed lag model introduced earlier. (Why?) You can think of it as arising by beginning with a univariate autoregressive model for  $y$ , and then introducing additional explanatory variables. If the lagged  $y$ 's don't play a role, as assessed with the usual tests, we can always delete them, but we never want to eliminate from the outset the possibility that lagged dependent variables play a role. Lagged dependent variables absorb residual serial correlation and can *dramatically* enhance forecasting performance.

Alternatively, we can capture own-variable dynamics in distributed-lag regression models by using a **distributed-lag regression model with *ARMA* disturbances**. Recall that our  $ARMA(p, q)$  models are equivalent to regression

models, with only a constant regressor, and with  $ARMA(p, q)$  disturbances,

$$\begin{aligned} y_t &= \beta_0 + \varepsilon_t \\ \varepsilon_t &= \frac{\Theta(L)}{\Phi(L)} v_t \\ v_t &\sim WN(0, \sigma^2). \end{aligned}$$

We want to begin with the univariate model as a baseline, and then generalize it to allow for multivariate interaction, resulting in models such as

$$\begin{aligned} y_t &= \beta_0 + \sum_{i=1}^{N_x} \delta_i x_{t-i} + \varepsilon_t \\ \varepsilon_t &= \frac{\Theta(L)}{\Phi(L)} v_t \\ v_t &\sim WN(0, \sigma^2). \end{aligned}$$

Regressions with  $ARMA$  disturbances make clear that regression (a statistical and econometric tool with a long tradition) and the  $ARMA$  model of time-series dynamics (a more recent innovation) are not at all competitors; rather, when used appropriately they can be highly complementary.

It turns out that the distributed-lag regression model with autoregressive disturbances – a great workhorse in econometrics – is a special case of the more general model with lags of both  $y$  and  $x$  and white noise disturbances. To see this, let's take the simple example of an unconditional (1-step-ahead) regression forecasting model with  $AR(1)$  disturbances:

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_{t-1} + \varepsilon_t \\ \varepsilon_t &= \phi \varepsilon_{t-1} + v_t \\ v_t &\sim WN(0, \sigma^2). \end{aligned}$$

In lag operator notation, we write the  $AR(1)$  regression disturbance as

$$(1 - \phi L)\varepsilon_t = v_t,$$

or

$$\varepsilon_t = \frac{1}{(1 - \phi L)} v_t.$$

Thus we can rewrite the regression model as

$$y_t = \beta_0 + \beta_1 x_{t-1} + \frac{1}{(1 - \phi L)} v_t.$$

Now multiply both sides by  $(1 - \phi L)$  to get

$$(1 - \phi L)y_t = (1 - \phi)\beta_0 + \beta_1(1 - \phi L)x_{t-1} + v_t,$$

or

$$y_t = \phi y_{t-1} + (1 - \phi)\beta_0 + \beta_1 x_{t-1} - \phi\beta_1 x_{t-2} + v_t.$$

Thus a model with one lag of  $x$  on the right and  $AR(1)$  disturbances is equivalent to a model with  $y_{t-1}$ ,  $x_{t-1}$ , and  $x_{t-2}$  on the right-hand side and white noise errors, *subject to the restriction* that the coefficient on the second lag of  $x_{t-2}$  is the negative of the product of the coefficients on  $y_{t-1}$  and  $x_{t-1}$ . Thus, distributed lag regressions with lagged dependent variables are more general than distributed lag regressions with dynamic disturbances. In practice, the important thing is to allow for own-variable dynamics *somewhat*, in order to account for dynamics in  $y$  not explained by the right-hand-side variables. Whether we do so by including lagged dependent variables or by allowing for  $ARMA$  disturbances can occasionally be important, but usually it's a comparatively minor issue.

## 16.3 Vector Autoregressions

A univariate autoregression involves one variable. In a univariate autoregression of order  $p$ , we regress a variable on  $p$  lags of itself. In contrast, a multivariate autoregression – that is, a vector autoregression, or  $VAR$  – involves  $N$  variables. In an  $N$ -variable **vector autoregression of order  $p$** , or  $VAR(p)$ , we estimate  $N$  different equations. In each equation, we regress the relevant left-hand-side variable on  $p$  lags of itself, *and  $p$  lags of every other variable*.<sup>1</sup> Thus the right-hand-side variables are the same in every equation –  $p$  lags of every variable.

The key point is that, in contrast to the univariate case, vector autoregressions allow for **cross-variable dynamics**. Each variable is related not only to its own past, but also to the past of all the other variables in the system. In a two-variable  $VAR(1)$ , for example, we have two equations, one for each variable ( $y_1$  and  $y_2$ ) . We write

$$y_{1,t} = \phi_{11}y_{1,t-1} + \phi_{12}y_{2,t-1} + \varepsilon_{1,t}$$

$$y_{2,t} = \phi_{21}y_{1,t-1} + \phi_{22}y_{2,t-1} + \varepsilon_{2,t}.$$

Each variable depends on one lag of the other variable in addition to one lag of itself; that's one obvious source of multivariate interaction captured by the  $VAR$  that may be useful for forecasting. In addition, the disturbances may be correlated, so that when one equation is shocked, the other will typically be shocked as well, which is another type of multivariate interaction that univariate models miss. We summarize the disturbance variance-covariance structure as

$$\varepsilon_{1,t} \sim WN(0, \sigma_1^2)$$

$$\varepsilon_{2,t} \sim WN(0, \sigma_2^2)$$

---

<sup>1</sup>Trends, seasonals, and other exogenous variables may also be included, as long as they're all included in every equation.

$$\text{cov}(\varepsilon_{1,t}, \varepsilon_{2,t}) = \sigma_{12}.$$

The innovations *could* be uncorrelated, which occurs when  $\sigma_{12} = 0$ , but they needn't be.

You might guess that *VARs* would be hard to estimate. After all, they're fairly complicated models, with potentially many equations and many right-hand-side variables in each equation. In fact, precisely the opposite is true. *VARs* are very easy to estimate, because we need only run  $N$  linear regressions. That's one reason why *VARs* are so popular – OLS estimation of autoregressive models is simple and stable, in contrast to the numerical estimation required for models with moving-average components.<sup>2</sup> Equation-by-equation OLS estimation also turns out to have very good statistical properties when each equation has the same regressors, as is the case in standard *VARs*. Otherwise, a more complicated estimation procedure called seemingly unrelated regression, which explicitly accounts for correlation across equation disturbances, would be required to obtain estimates with good statistical properties.<sup>3</sup>

When fitting *VARs* to data, we use the Schwarz and Akaike criteria, just as in the univariate case. The formulas differ, however, because we're now working with a multivariate system of equations rather than a single equation. To get an *AIC* or *SIC* value for a *VAR* system, we could add up the equation-by-equation *AICs* or *SICs*, but unfortunately, doing so is appropriate only if the innovations are uncorrelated across equations, which is a very special and unusual situation. Instead, explicitly multivariate versions of the *AIC* and *SIC* – and more advanced formulas – are required that account for cross-equation innovation correlation. It's beyond the scope of this book to derive and present those formulas, because they involve unavoidable use of matrix

---

<sup>2</sup>Estimation of *MA* and *ARMA* models is stable enough in the univariate case but rapidly becomes unwieldy in multivariate situations. Hence multivariate *ARMA* models are used infrequently in practice, in spite of the potential they hold for providing parsimonious approximations to the Wold representation.

<sup>3</sup>For an exposition of seemingly unrelated regression, see Pindyck and Rubinfeld (1997).

algebra, but fortunately we don't need to. They're pre-programmed in many computer packages, and we interpret the *AIC* and *SIC* values computed for *VARs* of various orders in exactly the same way as in the univariate case: we select that order  $p$  such that the *AIC* or *SIC* is minimized.

We construct *VAR* forecasts in a way that precisely parallels the univariate case. We can construct 1-step-ahead point forecasts immediately, because all variables on the right-hand side are lagged by one period. Armed with the 1-step-ahead forecasts, we can construct the 2-step-ahead forecasts, from which we can construct the 3-step-ahead forecasts, and so on in the usual way, following the chain rule of forecasting. We construct interval and density forecasts in ways that also parallel the univariate case. The multivariate nature of *VARs* makes the derivations more tedious, however, so we bypass them. As always, to construct practical forecasts we replace unknown parameters by estimates.

## 16.4 Predictive Causality

There's an important statistical notion of causality that's intimately related to forecasting and naturally introduced in the context of *VARs*. It is based on two key principles: first, cause should occur before effect, and second, a causal series should contain information useful for forecasting that is not available in the other series (including the past history of the variable being forecast). In the unrestricted *VARs* that we've studied thus far, *everything* causes everything else, because lags of every variable appear on the right of every equation. Cause precedes effect because the right-hand-side variables are lagged, and each variable is useful in forecasting every other variable.

We stress from the outset that the notion of **predictive causality** contains little if any information about causality in the philosophical sense. Rather, the statement " $y_i$  causes  $y_j$ " is just shorthand for the more precise, but long-

winded, statement, “ $y_i$  contains useful information for predicting  $y_j$  (in the linear least squares sense), over and above the past histories of the other variables in the system.” To save space, we simply say that  $y_i$  causes  $y_j$ .

To understand what predictive causality means in the context of a  $VAR(p)$ , consider the  $j$ -th equation of the  $N$ -equation system, which has  $y_j$  on the left and  $p$  lags of each of the  $N$  variables on the right. If  $y_i$  causes  $y_j$ , then at least one of the lags of  $y_i$  that appear on the right side of the  $y_j$  equation must have a nonzero coefficient.

It’s also useful to consider the opposite situation, in which  $y_i$  does not cause  $y_j$ . In that case, all of the lags of that  $y_i$  that appear on the right side of the  $y_j$  equation must have zero coefficients.<sup>4</sup> Statistical causality tests are based on this formulation of non-causality. We use an  $F$ -test to assess whether all coefficients on lags of  $y_i$  are jointly zero.

Note that we’ve defined non-causality in terms of 1-step-ahead prediction errors. In the bivariate  $VAR$ , this implies non-causality in terms of  $h$ -step-ahead prediction errors, for all  $h$ . (Why?) In higher dimensional cases, things are trickier; 1-step-ahead noncausality does not necessarily imply noncausality at other horizons. For example, variable  $i$  may 1-step cause variable  $j$ , and variable  $j$  may 1-step cause variable  $k$ . Thus, variable  $i$  2-step causes variable  $k$ , but does not 1-step cause variable  $k$ .

Causality tests are often used when building and assessing forecasting models, because they can inform us about those parts of the workings of complicated multivariate models that are particularly relevant for forecasting. Just staring at the coefficients of an estimated  $VAR$  (and in complicated systems there are *many* coefficients) rarely yields insights into its workings. Thus we need tools that help us to see through to the practical forecasting properties of the model that concern us. And we often have keen interest in the answers to questions such as “Does  $y_i$  contribute toward improving

---

<sup>4</sup>Note that in such a situation the error variance in forecasting  $y_j$  using lags of all variables in the system will be the same as the error variance in forecasting  $y_j$  using lags of all variables in the system *except*  $y_i$ .

forecasts of  $y_j?$ ,” and “Does  $y_j$  contribute toward improving forecasts of  $y_i?$ ” If the results violate intuition or theory, then we might scrutinize the model more closely. In a situation in which we can’t reject a certain noncausality hypothesis, and neither intuition nor theory makes us uncomfortable with it, we might want to *impose* it, by omitting certain lags of certain variables from certain equations.

Various types of causality hypotheses are sometimes entertained. In any equation (the  $j$ -th, say), we’ve already discussed testing the simple noncausality hypothesis that:

- (a) No lags of variable  $i$  aid in one-step-ahead prediction of variable  $j$ .

We can broaden the idea, however. Sometimes we test stronger noncausality hypotheses such as:

- (b) No lags of a *set* of other variables aid in one-step-ahead prediction of variable  $j$ .
- (b) No lags of *any other variables* aid in one-step-ahead prediction of variable  $j$ .

All of hypotheses (a), (b) and (c) amount to assertions that various coefficients are zero. Finally, sometimes we test noncausality hypotheses that involve more than one equation, such as:

- (b) No variable in a set  $A$  causes any variable in a set  $B$ , in which case we say that the variables in  $A$  are block non-causal for those in  $B$ .

This particular noncausality hypothesis corresponds to exclusion restrictions that hold simultaneously in a number of equations. Again, however, standard test procedures are applicable.

## 16.5 Impulse-Response Functions

The **impulse-response function** is another device that helps us to learn about the dynamic properties of vector autoregressions of interest to forecasters. We'll introduce it first in the *univariate* context, and then we'll move to *VARs*. The question of interest is simple and direct: How does a unit innovation to a series affect it, now and in the future? To answer the question, we simply read off the coefficients in the moving average representation of the process.

We're used to normalizing the coefficient on  $\varepsilon_t$  to unity in moving-average representations, but we don't have to do so; more generally, we can write

$$y_t = b_0 \varepsilon_t + b_1 \varepsilon_{t-1} + b_2 \varepsilon_{t-2} + \dots$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

The additional generality introduces ambiguity, however, because we can always multiply and divide every  $\varepsilon_t$  by an arbitrary constant  $m$ , yielding an equivalent model but with different parameters and innovations,

$$y_t = (b_0 m) \left( \frac{1}{m} \varepsilon_t \right) + (b_1 m) \left( \frac{1}{m} \varepsilon_{t-1} \right) + (b_2 m) \left( \frac{1}{m} \varepsilon_{t-2} \right) + \dots$$

$$\varepsilon_t \sim WN(0, \sigma^2)$$

or

$$y_t = b'_0 \varepsilon'_t + b'_1 \varepsilon'_{t-1} + b'_2 \varepsilon'_{t-2} + \dots$$

$$\varepsilon'_t \sim WN(0, \frac{\sigma^2}{m^2}),$$

where  $b'_i = b_i m$  and  $\varepsilon'_t = \frac{\varepsilon_t}{m}$ .

To remove the ambiguity, we must set a value of  $m$ . Typically we set  $m = 1$ , which yields the standard form of the moving average representation. For impulse-response analysis, however, a different normalization turns out

to be particularly convenient; we choose  $m = \sigma$ , which yields

$$y_t = (b_0\sigma) \left( \frac{1}{\sigma} \varepsilon_t \right) + (b_1\sigma) \left( \frac{1}{\sigma} \varepsilon_{t-1} \right) + (b_2\sigma) \left( \frac{1}{\sigma} \varepsilon_{t-2} \right) + \dots$$

$$\varepsilon_t \sim WN(0, \sigma^2),$$

or

$$y_t = b'_0 \varepsilon'_t + b'_1 \varepsilon'_{t-1} + b'_2 \varepsilon'_{t-2} + \dots$$

$$\varepsilon'_t \sim WN(0, 1),$$

where  $b'_i = b_i\sigma$  and  $\varepsilon'_t = \frac{\varepsilon_t}{\sigma}$ . Taking  $m = \sigma$  converts shocks to “standard deviation units,” because a unit shock to  $\varepsilon'_t$  corresponds to a one standard deviation shock to  $\varepsilon_t$ .

To make matters concrete, consider the univariate *AR*(1) process,

$$y_t = \phi y_{t-1} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

The standard moving average form is

$$y_t = \varepsilon_t + \phi \varepsilon_{t-1} + \phi^2 \varepsilon_{t-2} + \dots$$

$$\varepsilon_t \sim WN(0, \sigma^2),$$

and the equivalent representation in standard deviation units is

$$y_t = b_0 \varepsilon'_t + b_1 \varepsilon'_{t-1} + b_2 \varepsilon'_{t-2} + \dots$$

$$\varepsilon'_t \sim WN(0, 1)$$

where  $b_i = \phi^i \sigma$  and  $\varepsilon'_t = \frac{\varepsilon_t}{\sigma}$ . The impulse-response function is  $\{ b_0, b_1, \dots \}$ . The parameter  $b_0$  is the contemporaneous effect of a unit shock to  $\varepsilon'_t$ , or equivalently a one standard deviation shock to  $\varepsilon_t$ ; as must be the case,

then,  $b_0 = \sigma$ . Note well that  $b_0$  gives the immediate effect of the shock at time  $t$ , when it hits. The parameter  $b_1$ , which multiplies  $\varepsilon'_{t-1}$ , gives the effect of the shock one period later, and so on. The full set of impulse-response coefficients,  $\{b_0, b_1, \dots\}$ , tracks the complete dynamic response of  $y$  to the shock.

Now we consider the multivariate case. The idea is the same, but there are more shocks to track. The key question is, “How does a unit shock to  $\varepsilon_i$  affect  $y_j$ , now and in the future, for all the various combinations of  $i$  and  $j$ ?”. Consider, for example, the bivariate  $VAR(1)$ ,

$$y_{1t} = \phi_{11}y_{1,t-1} + \phi_{12}y_{2,t-1} + \varepsilon_{1t}$$

$$y_{2t} = \phi_{21}y_{1,t-1} + \phi_{22}y_{2,t-1} + \varepsilon_{2t}$$

$$\varepsilon_{1,t} \sim WN(0, \sigma_1^2)$$

$$\varepsilon_{2,t} \sim WN(0, \sigma_2^2)$$

$$cov(\varepsilon_1, \varepsilon_2) = \sigma_{12}.$$

The standard moving average representation, obtained by back substitution, is

$$y_{1t} = \varepsilon_{1t} + \phi_{11}\varepsilon_{1,t-1} + \phi_{12}\varepsilon_{2,t-1} + \dots$$

$$y_{2t} = \varepsilon_{2t} + \phi_{21}\varepsilon_{1,t-1} + \phi_{22}\varepsilon_{2,t-1} + \dots$$

$$\varepsilon_{1,t} \sim WN(0, \sigma_1^2)$$

$$\varepsilon_{2,t} \sim WN(0, \sigma_2^2)$$

$$cov(\varepsilon_1, \varepsilon_2) = \sigma_{12}.$$

Just as in the univariate case, it proves fruitful to adopt a different normalization of the moving average representation for impulse-response analysis. The multivariate analog of our univariate normalization by  $\sigma$  is called

normalization by the Cholesky factor.<sup>5</sup> The resulting VAR moving average representation has a number of useful properties that parallel the univariate case precisely. First, the innovations of the transformed system are in standard deviation units. Second, although the current innovations in the standard representation have unit coefficients, the current innovations in the normalized representation have non-unit coefficients. In fact, the first equation has only one current innovation,  $\varepsilon_{1t}$ . (The other has a zero coefficient.) The second equation has both current innovations. Thus, the ordering of the variables can matter.<sup>6</sup>

If  $y_1$  is ordered first, the normalized representation is

$$\begin{aligned} y_{1,t} &= b_{11}^0 \varepsilon'_{1,t} + b_{11}^1 \varepsilon'_{1,t-1} + b_{12}^1 \varepsilon'_{2,t-1} + \dots \\ y_{2,t} &= b_{21}^0 \varepsilon'_{1,t} + b_{22}^0 \varepsilon'_{2,t} + b_{21}^1 \varepsilon'_{1,t-1} + b_{22}^1 \varepsilon'_{2,t-1} + \dots \\ \varepsilon'_{1,t} &\sim WN(0, 1) \\ \varepsilon'_{2,t} &\sim WN(0, 1) \\ cov(\varepsilon'_1, \varepsilon'_2) &= 0. \end{aligned}$$

Alternatively, if  $y_2$  ordered first, the normalized representation is

$$\begin{aligned} y_{2,t} &= b_{22}^0 \varepsilon'_{2,t} + b_{21}^1 \varepsilon'_{1,t-1} + b_{22}^1 \varepsilon'_{2,t-1} + \dots \\ y_{1,t} &= b_{11}^0 \varepsilon'_{1,t} + b_{12}^0 \varepsilon'_{2,t} + b_{11}^1 \varepsilon'_{1,t-1} + b_{12}^1 \varepsilon'_{2,t-1} + \dots \\ \varepsilon'_{1,t} &\sim WN(0, 1) \\ \varepsilon'_{2,t} &\sim WN(0, 1) \\ cov(\varepsilon'_1, \varepsilon'_2) &= 0. \end{aligned}$$

---

<sup>5</sup>For detailed discussion and derivation of this advanced topic, see Hamilton (1994).

<sup>6</sup>In higher-dimensional VAR's, the equation that's first in the ordering has only one current innovation,  $\varepsilon'_{1t}$ . The equation that's second has only current innovations  $\varepsilon'_{1t}$  and  $\varepsilon'_{2t}$ , the equation that's third has only current innovations  $\varepsilon'_{1t}$ ,  $\varepsilon'_{2t}$  and  $\varepsilon'_{3t}$ , and so on.

Finally, the normalization adopted yields a zero covariance between the disturbances of the transformed system. This is crucial, because it lets us perform the experiment of interest – shocking one variable in isolation of the others, which we can do if the innovations are uncorrelated but can't do if they're correlated, as in the original unnormalized representation.

After normalizing the system, for a given ordering, say  $y_1$  first, we compute four sets of impulse-response functions for the bivariate model: response of  $y_1$  to a unit normalized innovation to  $y_1$ ,  $\{ b_{11}^0, b_{11}^1, b_{11}^2, \dots \}$ , response of  $y_1$  to a unit normalized innovation to  $y_2$ ,  $\{ b_{12}^1, b_{12}^2, \dots \}$ , response of  $y_2$  to a unit normalized innovation to  $y_2$ ,  $\{ b_{22}^0, b_{22}^1, b_{22}^2, \dots \}$ , and response of  $y_2$  to a unit normalized innovation to  $y_1$ ,  $\{ b_{21}^0, b_{21}^1, b_{21}^2, \dots \}$ . Typically we examine the set of impulse-response functions graphically. Often it turns out that impulse-response functions aren't sensitive to ordering, but the only way to be sure is to check.<sup>7</sup>

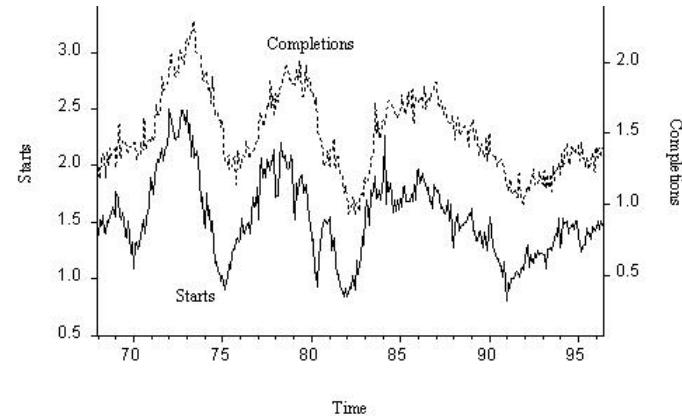
In practical applications of impulse-response analysis, we simply replace unknown parameters by estimates, which immediately yields point estimates of the impulse-response functions. Getting confidence intervals for impulse-response functions is trickier, however, and adequate procedures are still under development.

## 16.6 Variance Decompositions

Another way of characterizing the dynamics associated with *VARs*, closely related to impulse-response functions, is the **variance decomposition**. Variance decompositions have an immediate link to forecasting – they answer the question, “How much of the  $h$ -step-ahead forecast error variance of variable  $i$  is explained by innovations to variable  $j$ , for  $h = 1, 2, \dots$ ” As with impulse-response functions, we typically make a separate graph for every  $(i, j)$  pair.

---

<sup>7</sup>Note well that the issues of normalization and ordering only affect impulse-response analysis; for forecasting we only need the unnormalized model.



Notes to figure: The left scale is starts, and the right scale is completions.

Figure 16.1: Housing Starts and Completions, 1968 - 1996

Impulse-response functions and the variance decompositions present the same information (although they do so in different ways). For that reason it's not strictly necessary to present both, and impulse-response analysis has gained greater popularity. Hence we offer only this brief discussion of variance decomposition. In the application to housing starts and completions that follows, however, we examine both impulse-response functions and variance decompositions. The two are highly complementary, as with information criteria and correlograms for model selection, and the variance decompositions have a nice forecasting motivation.

## 16.7 Application: Housing Starts and Completions

We estimate a bivariate *VAR* for U.S. seasonally-adjusted housing starts and completions, two widely-watched business cycle indicators, 1968.01-1996.06. We use the *VAR* to produce point extrapolation forecasts. We show housing starts and completions in Figure 16.1. Both are highly cyclical, increasing during business-cycle expansions and decreasing during contractions. Moreover, completions tend to lag behind starts, which makes sense because a house takes time to complete.

	Included observations: 288				
	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	0.937	0.937	0.059	255.24	0.000
2	0.907	0.244	0.059	495.53	0.000
3	0.877	0.054	0.059	720.95	0.000
4	0.838	-0.077	0.059	927.39	0.000
5	0.795	-0.096	0.059	1113.7	0.000
6	0.751	-0.058	0.059	1280.9	0.000
7	0.704	-0.067	0.059	1428.2	0.000
8	0.650	-0.098	0.059	1554.4	0.000
9	0.604	0.004	0.059	1663.8	0.000
10	0.544	-0.129	0.059	1752.6	0.000
11	0.496	0.029	0.059	1826.7	0.000
12	0.446	-0.008	0.059	1886.8	0.000
13	0.405	0.076	0.059	1936.8	0.000
14	0.346	-0.144	0.059	1973.3	0.000
15	0.292	-0.079	0.059	1999.4	0.000
16	0.233	-0.111	0.059	2016.1	0.000
17	0.175	-0.050	0.059	2025.6	0.000
18	0.122	-0.018	0.059	2030.2	0.000
19	0.070	0.002	0.059	2031.7	0.000
20	0.019	-0.025	0.059	2031.8	0.000
21	-0.034	-0.032	0.059	2032.2	0.000
22	-0.074	0.036	0.059	2033.9	0.000
23	-0.123	-0.028	0.059	2038.7	0.000
24	-0.167	-0.048	0.059	2047.4	0.000

Figure 16.2: Housing Starts Correlogram

We split the data into an estimation sample, 1968.01-1991.12, and a hold-out sample, 1992.01-1996.06 for forecasting. We therefore perform all model specification analysis and estimation, to which we now turn, on the 1968.01-1991.12 data. We show the starts correlogram in Table 16.2 and Figure 16.3. The sample autocorrelation function decays slowly, whereas the sample partial autocorrelation function appears to cut off at displacement 2. The patterns in the sample autocorrelations and partial autocorrelations are highly statistically significant, as evidenced by both the Bartlett standard errors and the Ljung-Box  $Q$ -statistics. The completions correlogram, in Table 16.4 and Figure 16.5, behaves similarly.

We've not yet introduced the **cross correlation function**. There's been no need, because it's not relevant for univariate modeling. It provides important information, however, in the multivariate environments that now concern us. Recall that the autocorrelation function is the correlation between a variable and lags of itself. The cross-correlation function is a natural multivariate analog; it's simply the correlation between a variable and lags of *another*

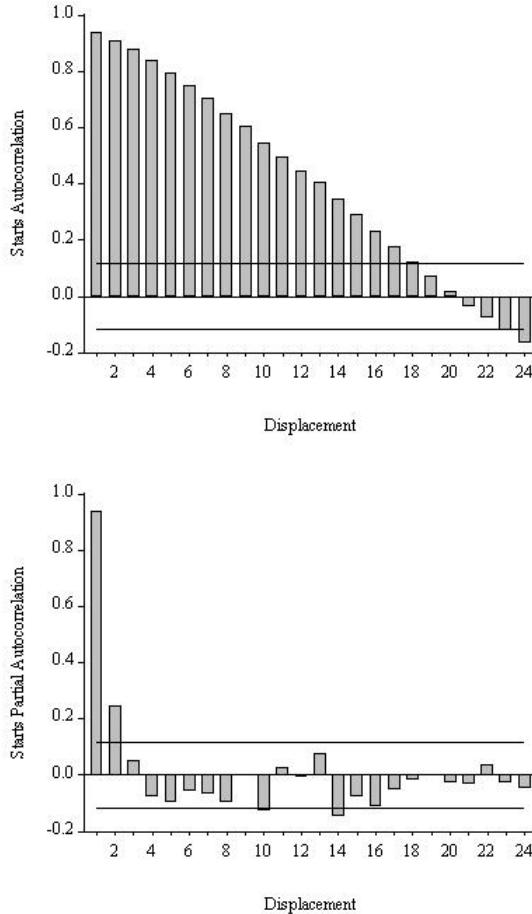


Figure 16.3: Housing Starts Autocorrelations and Partial Autocorrelations

variable. We estimate those correlations using the usual estimator and graph them as a function of displacement along with the Bartlett two- standard-error bands, which apply just as in the univariate case.

The cross-correlation function (Figure 16.6) for housing starts and completions is very revealing. Starts and completions are highly correlated at all displacements, and a clear pattern emerges as well: although the contemporaneous correlation is high (.78), completions are maximally correlated with starts lagged by roughly 6-12 months (around .90). Again, this makes good sense in light of the time it takes to build a house.

Now we proceed to model starts and completions. We need to select the order,  $p$ , of our  $VAR(p)$ . Based on exploration using multivariate versions of  $SIC$  and  $AIC$ , we adopt a  $VAR(4)$ .

Sample: 1968:01 1991:12  
Included observations: 288

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	0.939	0.939	0.059	256.61	0.000
2	0.920	0.328	0.059	504.05	0.000
3	0.896	0.066	0.059	739.19	0.000
4	0.874	0.023	0.059	963.73	0.000
5	0.834	-0.165	0.059	1168.9	0.000
6	0.802	-0.067	0.059	1359.2	0.000
7	0.761	-0.100	0.059	1531.2	0.000
8	0.721	-0.070	0.059	1686.1	0.000
9	0.677	-0.055	0.059	1823.2	0.000
10	0.633	-0.047	0.059	1943.7	0.000
11	0.583	-0.080	0.059	2046.3	0.000
12	0.533	-0.073	0.059	2132.2	0.000
13	0.483	-0.038	0.059	2203.2	0.000
14	0.434	-0.020	0.059	2260.6	0.000
15	0.390	0.041	0.059	2307.0	0.000
16	0.337	-0.057	0.059	2341.9	0.000
17	0.290	-0.008	0.059	2367.9	0.000
18	0.234	-0.109	0.059	2384.8	0.000
19	0.181	-0.082	0.059	2395.0	0.000
20	0.128	-0.047	0.059	2400.1	0.000
21	0.068	-0.133	0.059	2401.6	0.000
22	0.020	0.037	0.059	2401.7	0.000
23	-0.038	-0.092	0.059	2402.2	0.000
24	-0.087	-0.003	0.059	2404.6	0.000

Figure 16.4: Housing Completions Correlogram

First consider the starts equation (Table 16.7a), residual plot (Figure 16.7b), and residual correlogram (Table 16.8, Figure 16.9). The explanatory power of the model is good, as judged by the  $R^2$  as well as the plots of actual and fitted values, and the residuals appear white, as judged by the residual sample autocorrelations, partial autocorrelations, and Ljung-Box statistics. Note as well that no lag of completions has a significant effect on starts, which makes sense – we obviously expect starts to cause completions, but not conversely. The completions equation (Table 16.10a), residual plot (Figure 16.10b), and residual correlogram (Table 16.11, Figure 16.12) appear similarly good. Lagged starts, moreover, most definitely have a significant effect on completions.

Table 16.13 shows the results of formal causality tests. The hypothesis that starts don't cause completions is simply that the coefficients on the four lags of starts in the completions equation are all zero. The  $F$ -statistic is overwhelmingly significant, which is not surprising in light of the previously-

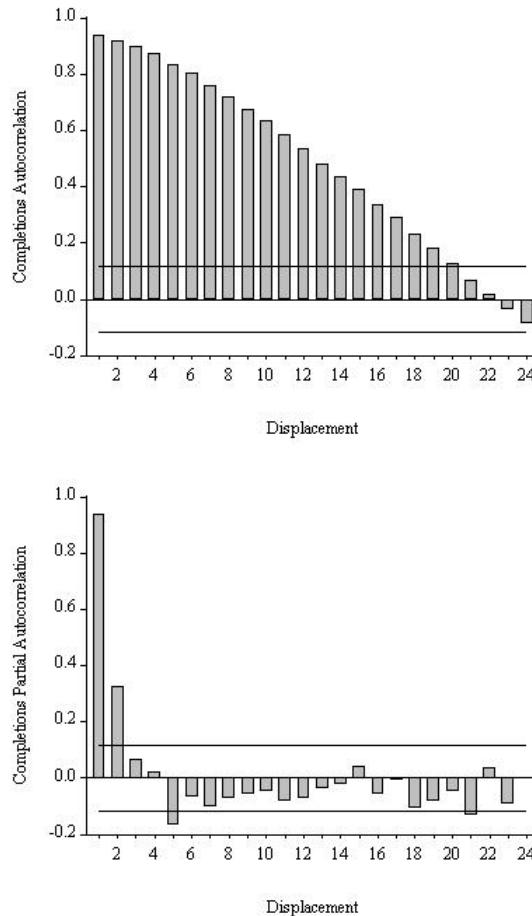
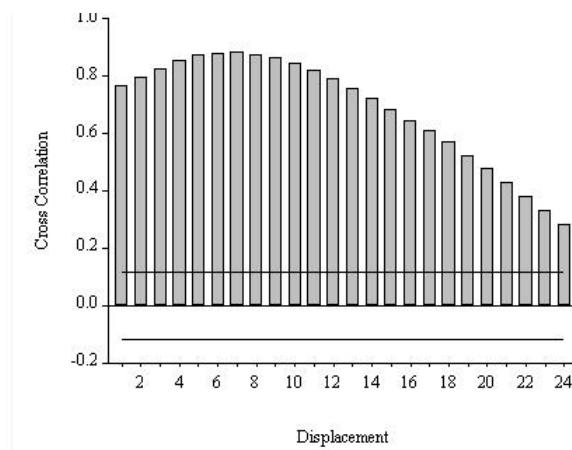


Figure 16.5: Housing Completions Autocorrelations and Partial Autocorrelations

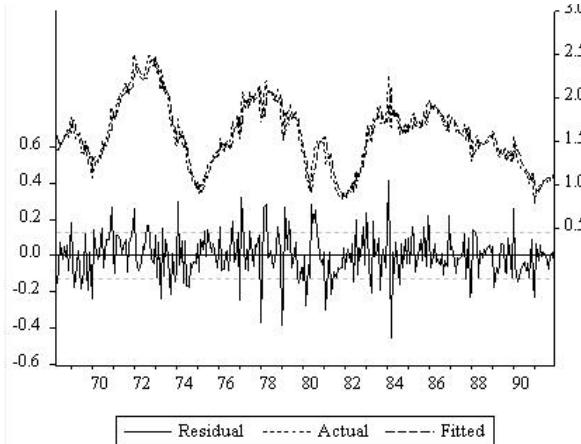


Notes to figure: We graph the sample correlation between completions at time  $t$  and starts at time  $t-i$ ,  $i = 1, 2, \dots, 24$ .

Figure 16.6: Housing Starts and Completions Sample Cross Correlations

Sample(adjusted): 1968:05 1991:12 Included observations: 284 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.146871	0.044235	3.320264	0.0010
STARTS(-1)	0.659939	0.061242	10.77587	0.0000
STARTS(-2)	0.229632	0.072724	3.157587	0.0018
STARTS(-3)	0.142859	0.072655	1.966281	0.0503
STARTS(-4)	0.007806	0.066032	0.118217	0.9060
COMPS(-1)	0.031611	0.102712	0.307759	0.7585
COMPS(-2)	-0.120781	0.103847	-1.163069	0.2458
COMPS(-3)	-0.020601	0.100946	-0.204078	0.8384
COMPS(-4)	-0.027404	0.094569	-0.289779	0.7722
R-squared	0.895566	Mean dependent var		1.574771
Adjusted R-squared	0.892528	S.D. dependent var		0.382362
S.E. of regression	0.125350	Akaike info criterion		-4.122118
Sum squared resid	4.320952	Schwarz criterion		-4.006482
Log likelihood	191.3622	F-statistic		294.7796
Durbin-Watson stat	1.991908	Prob(F-statistic)		0.000000

(a) VAR Starts Equation



(b) VAR Starts Equation - Residual Plot

Figure 16.7: VAR Starts Model

noticed highly-significant t-statistics. Thus we reject noncausality from starts to completions at any reasonable level. Perhaps more surprising, we also reject noncausality from completions to starts at roughly the 5% level. Thus the causality appears bi-directional, in which case we say there is **feedback**.

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	0.001	0.001	0.059	0.0004	0.985
2	0.003	0.003	0.059	0.0029	0.999
3	0.006	0.006	0.059	0.0119	1.000
4	0.023	0.023	0.059	0.1650	0.997
5	-0.013	-0.013	0.059	0.2108	0.999
6	0.022	0.021	0.059	0.3463	0.999
7	0.038	0.038	0.059	0.7646	0.998
8	-0.048	-0.048	0.059	1.4362	0.994
9	0.056	0.056	0.059	2.3528	0.985
10	-0.114	-0.116	0.059	6.1868	0.799
11	-0.038	-0.038	0.059	6.6096	0.830
12	-0.030	-0.028	0.059	6.8763	0.866
13	0.192	0.193	0.059	17.947	0.160
14	0.014	0.021	0.059	18.010	0.206
15	0.063	0.067	0.059	19.199	0.205
16	-0.006	-0.015	0.059	19.208	0.258
17	-0.039	-0.035	0.059	19.664	0.292
18	-0.029	-0.043	0.059	19.927	0.337
19	-0.010	-0.009	0.059	19.959	0.397
20	0.010	-0.014	0.059	19.993	0.458
21	-0.057	-0.047	0.059	21.003	0.459
22	0.045	0.018	0.059	21.644	0.481
23	-0.038	0.011	0.059	22.088	0.515
24	-0.149	-0.141	0.059	29.064	0.218

Figure 16.8: VAR Starts Residual Correlogram

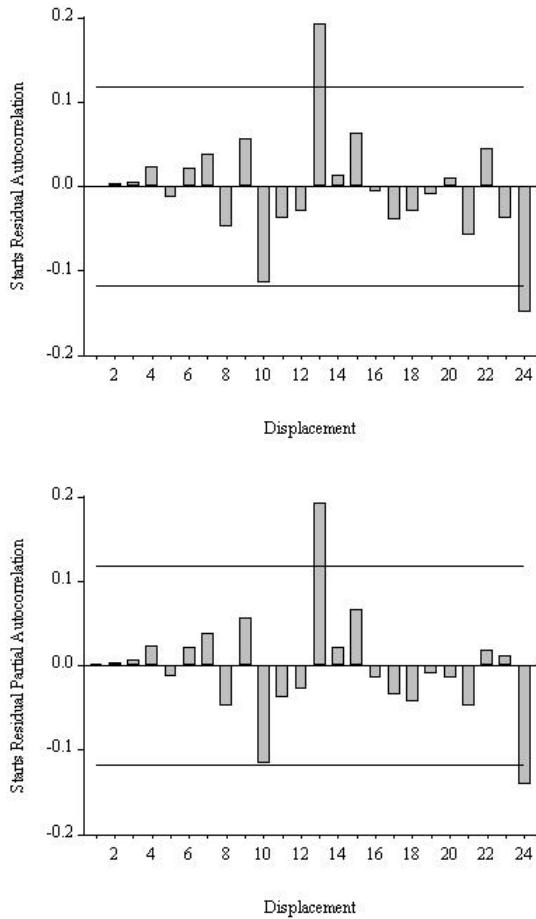
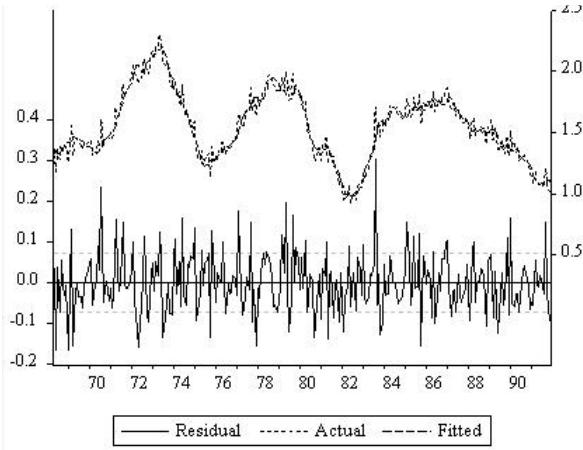


Figure 16.9: VAR Starts Equation - Sample Autocorrelation and Partial Autocorrelation

Sample(adjusted): 1968:05 1991:12 Included observations: 284 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.045347	0.025794	1.758045	0.0799
STARTS(-1)	0.074724	0.035711	2.092461	0.0373
STARTS(-2)	0.040047	0.042406	0.944377	0.3458
STARTS(-3)	0.047145	0.042366	1.112805	0.2668
STARTS(-4)	0.082331	0.038504	2.138238	0.0334
COMPS(-1)	0.236774	0.059893	3.953313	0.0001
COMPS(-2)	0.206172	0.060554	3.404742	0.0008
COMPS(-3)	0.120998	0.058863	2.055593	0.0408
COMPS(-4)	0.156729	0.055144	2.842160	0.0048
R-squared	0.936835	Mean dependent var	1.547958	
Adjusted R-squared	0.934998	S.D. dependent var	0.286689	
S.E. of regression	0.073093	Akaike info criterion	-5.200872	
Sum squared resid	1.469205	Schwarz criterion	-5.085236	
Log likelihood	344.5453	F-statistic	509.8375	
Durbin-Watson stat	2.013370	Prob(F-statistic)	0.000000	

(a) VAR Completions Equation



(b) VAR Completions Equation - Residual Plot

Figure 16.10: VAR Completions Model

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	-0.009	-0.009	0.059	0.0238	0.877
2	-0.035	-0.035	0.059	0.3744	0.829
3	-0.037	-0.037	0.059	0.7640	0.858
4	-0.088	-0.090	0.059	3.0059	0.557
5	-0.105	-0.111	0.059	6.1873	0.288
6	0.012	0.000	0.059	6.2291	0.398
7	-0.024	-0.041	0.059	6.4047	0.493
8	0.041	0.024	0.059	6.9026	0.547
9	0.048	0.029	0.059	7.5927	0.576
10	0.045	0.037	0.059	8.1918	0.610
11	-0.009	-0.005	0.059	8.2160	0.694
12	-0.050	-0.046	0.059	8.9767	0.705
13	-0.038	-0.024	0.059	9.4057	0.742
14	-0.055	-0.049	0.059	10.3118	0.739
15	0.027	0.028	0.059	10.545	0.784
16	-0.005	-0.020	0.059	10.553	0.836
17	0.096	0.082	0.059	13.369	0.711
18	0.011	-0.002	0.059	13.405	0.767
19	0.041	0.040	0.059	13.929	0.788
20	0.046	0.061	0.059	14.569	0.801
21	-0.096	-0.079	0.059	17.402	0.686
22	0.039	0.077	0.059	17.875	0.713
23	-0.113	-0.114	0.059	21.824	0.531
24	-0.136	-0.125	0.059	27.622	0.276

Figure 16.11: VAR Completions Residual Correlogram

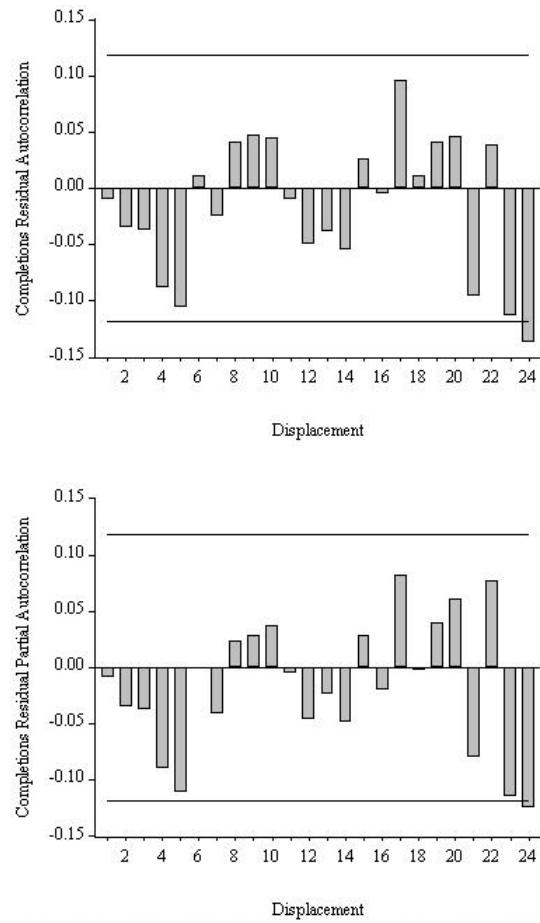


Figure 16.12: VAR Completions Equation - Sample Autocorrelation and Partial Autocorrelation

Sample: 1968:01 1991:12		
Lags: 4		
Obs: 284		
Null Hypothesis:	F-Statistic	Probability
STARTS does not Cause COMPS	26.2658	0.00000
COMPS does not Cause STARTS	2.23876	0.06511

Figure 16.13: Housing Starts and Completions - Causality Tests

In order to get a feel for the dynamics of the estimated *VAR* before producing forecasts, we compute impulse-response functions and variance decompositions. We present results for starts first in the ordering, so that a current innovation to starts affects only current starts, but the results are robust to reversal of the ordering.

In Figure 16.14, we display the impulse-response functions. First let's consider the own-variable impulse responses, that is, the effects of a starts innovation on subsequent starts or a completions innovation on subsequent completions; the effects are similar. In each case, the impulse response is large and decays in a slow, approximately monotonic fashion. In contrast, the cross-variable impulse responses are very different. An innovation to starts produces no movement in completions at first, but the effect gradually builds and becomes large, peaking at about fourteen months. (It takes time to build houses.) An innovation to completions, however, produces little movement in starts at any time. Figure 16.15 shows the variance decompositions. The fraction of the error variance in forecasting starts due to innovations in starts is close to 100 percent at all horizons. In contrast, the fraction of the error variance in forecasting completions due to innovations in starts is near zero at short horizons, but it rises steadily and is near 100 percent at long horizons, again reflecting time-to-build effects.

Finally, we construct forecasts for the out-of-sample period, 1992.01-1996.06. The starts forecast appears in Figure 16.16. Starts begin their recovery before 1992.01, and the *VAR* projects continuation of the recovery. The *VAR* fore-

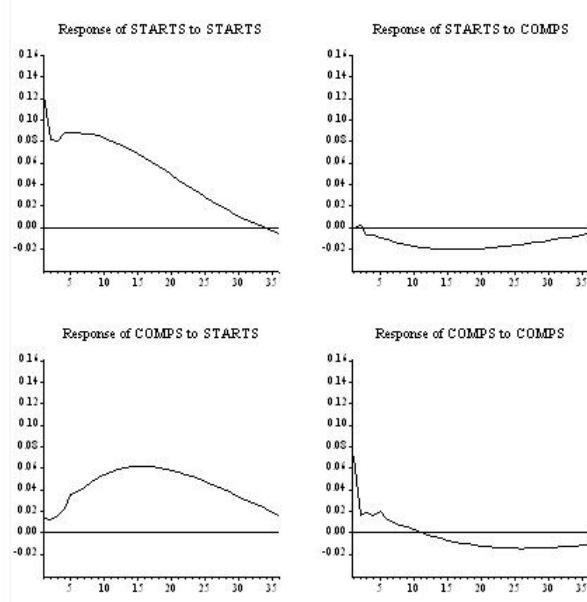


Figure 16.14: Housing Starts and Completions - VAR Impulse Response Functions. Response is to 1 SD innovation.

casts captures the general pattern quite well, but it forecasts quicker mean reversion than actually occurs, as is clear when comparing the forecast and realization in Figure 16.17. The figure also makes clear that the recovery of housing starts from the recession of 1990 was slower than the previous recoveries in the sample, which naturally makes for difficult forecasting. The completions forecast suffers the same fate, as shown in Figures 16.18 and 16.19. Interestingly, however, completions had not yet turned by 1991.12, but the forecast nevertheless correctly predicts the turning point. (Why?)

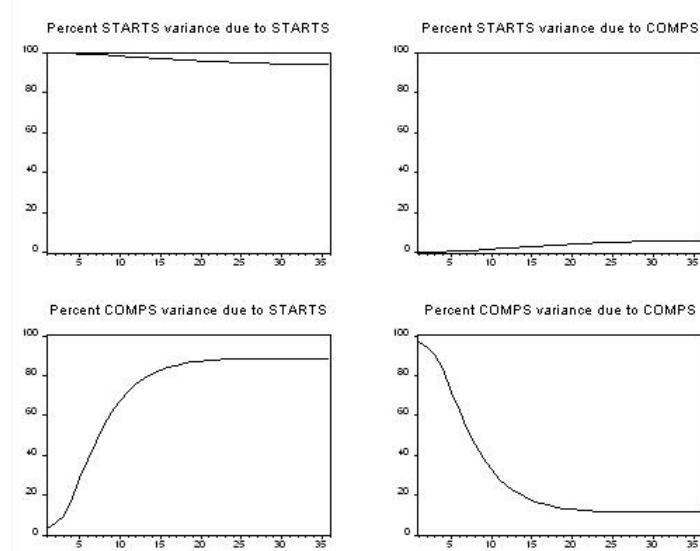


Figure 16.15: Housing Starts and Completions - VAR Variance Decompositions

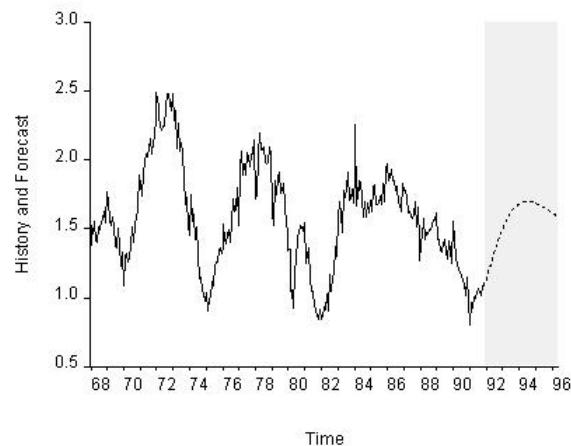


Figure 16.16: Housing Starts Forecast

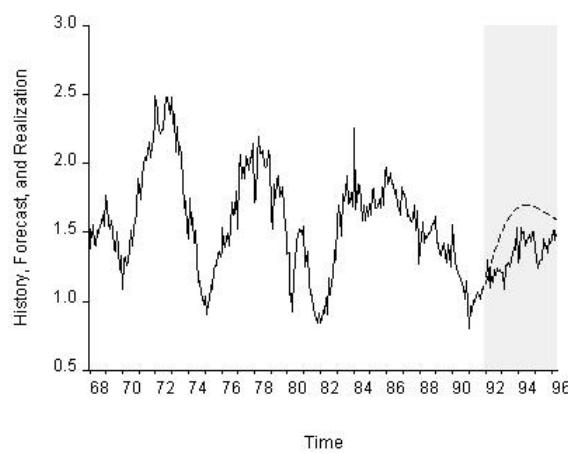


Figure 16.17: Housing Starts Forecast and Realization

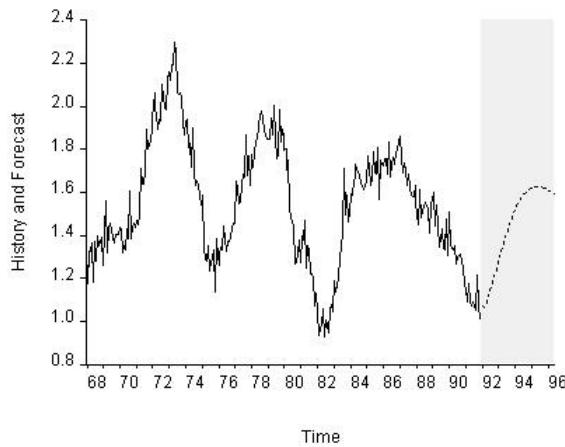


Figure 16.18: Housing Completions Forecast

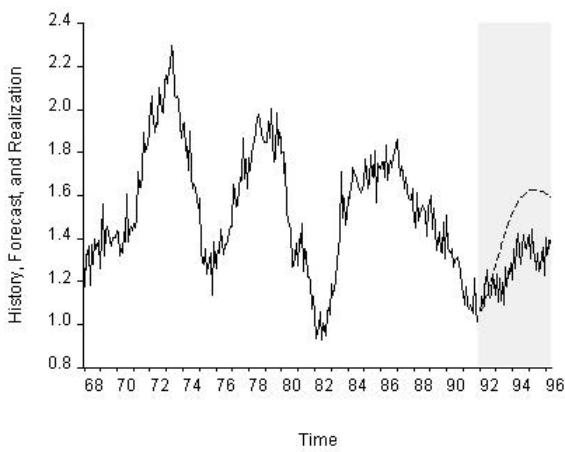


Figure 16.19: Housing Completions Forecast and Realization

## 16.8 Exercises, Problems and Complements

### 1. Housing starts and completions, continued.

Our VAR analysis of housing starts and completions, as always, involved many judgment calls. Using the starts and completions data, assess the adequacy of our models and forecasts. Among other things, you may want to consider the following questions:

- a. Should we allow for a trend in the forecasting model?
  - b. How do the results change if, in light of the results of the causality tests, we exclude lags of completions from the starts equation, re-estimate by seemingly-unrelated regression, and forecast?
  - c. Are the VAR forecasts of starts and completions more accurate than univariate forecasts?
2. Forecasting crop yields.

Consider the following dilemma in agricultural crop yield forecasting:

The possibility of forecasting crop yields several years in advance would, of course, be of great value in the planning of agricultural production. However, the success of long-range crop forecasts is contingent not only on our knowledge of the weather factors determining yield, but also on our ability to predict the weather. Despite an abundant literature in this field, no firm basis for reliable long-range weather forecasts has yet been found. (Sanderson, 1953, p. 3)

- a. How is the situation related to our concerns in this chapter, and specifically, to the issue of conditional vs. unconditional forecasting?
- b. What variables other than weather might be useful for predicting crop yield?
- c. How would you suggest that the forecaster should proceed?

### 3. Econometrics, time series analysis, and forecasting.

As recently as the early 1970s, time series analysis was mostly univariate and made little use of economic theory. Econometrics, in contrast, stressed the cross-variable dynamics associated with economic theory, with equations estimated using multiple regression. Econometrics, moreover, made use of simultaneous systems of such equations, requiring complicated estimation methods. Thus the econometric and time series approaches to forecasting were very different.<sup>8</sup>

As Klein (1981) notes, however, the complicated econometric system estimation methods had little payoff for practical forecasting and were therefore largely abandoned, whereas the rational distributed lag patterns associated with time-series models led to large improvements in practical forecast accuracy.<sup>9</sup> Thus, in more recent times, the distinction between econometrics and time series analysis has largely vanished, with the union incorporating the best of both. In many respects the *VAR* is a modern embodiment of both econometric and time-series traditions. *VARs* use economic considerations to determine which variables to include and which (if any) restrictions should be imposed, allow for rich multivariate dynamics, typically require only simple estimation techniques, and are explicit forecasting models.

### 4. Business cycle analysis and forecasting: expansions, contractions, turning points, and leading indicators<sup>10</sup>.

The use of anticipatory data is linked to business cycle analysis in general, and leading indicators in particular. During the first half of this

<sup>8</sup>Klein and Young (1980) and Klein (1983) provide good discussions of the traditional econometric simultaneous equations paradigm, as well as the link between structural simultaneous equations models and reduced-form time series models. Wallis (1995) provides a good summary of modern large-scale macroeconomic modeling and forecasting, and Pagan and Robertson (2002) provide an intriguing discussion of the variety of macroeconomic forecasting approaches currently employed in central banks around the world.

<sup>9</sup>For an acerbic assessment circa the mid-1970s, see Jenkins (1979).

<sup>10</sup>This complement draws in part upon Diebold and Rudebusch (1996).

century, much research was devoted to obtaining an empirical characterization of the business cycle. The most prominent example of this work was Burns and Mitchell (1946), whose summary empirical definition was:

Business cycles are a type of fluctuation found in the aggregate economic activity of nations that organize their work mainly in business enterprises: a cycle consists of expansions occurring at about the same time in many economic activities, followed by similarly general recessions, contractions, and revivals which merge into the expansion phase of the next cycle. (p. 3)

The comovement among individual economic variables was a key feature of Burns and Mitchell's definition of business cycles. Indeed, the comovement among series, taking into account possible leads and lags in timing, was the centerpiece of Burns and Mitchell's methodology. In their analysis, Burns and Mitchell considered the historical concordance of hundreds of series, including those measuring commodity output, income, prices, interest rates, banking transactions, and transportation services, and they classified series as leading, lagging or coincident. One way to define a leading indicator is to say that a series  $x$  is a leading indicator for a series  $y$  if  $x$  causes  $y$  in the predictive sense. According to that definition, for example, our analysis of housing starts and completions indicates that starts are a leading indicator for completions.

Leading indicators have the potential to be used in forecasting equations in the same way as anticipatory variables. Inclusion of a leading indicator, appropriately lagged, can improve forecasts. Zellner and Hong (1989) and Zellner, Hong and Min (1991), for example, make good use of that idea in their ARLI (autoregressive leading-indicator) models for forecasting aggregate output growth. In those models, Zellner *et al.* build forecasting models by regressing output on lagged output and lagged leading indicators; they also use shrinkage techniques to coax

the forecasted growth rates toward the international average, which improves forecast performance.

Burns and Mitchell used the clusters of turning points in individual series to determine the monthly dates of the turning points in the overall business cycle, and to construct composite indexes of leading, coincident, and lagging indicators. Such indexes have been produced by the National Bureau of Economic Research (a think tank in Cambridge, Mass.), the Department of Commerce (a U.S. government agency in Washington, DC), and the Conference Board (a business membership organization based in New York).<sup>11</sup> Composite indexes of leading indicators are often used to gauge likely future economic developments, but their usefulness is by no means uncontroversial and remains the subject of ongoing research. For example, leading indexes apparently cause aggregate output in analyses of ex post historical data (Auerbach, 1982), but they appear much less useful in real-time forecasting, which is what's relevant (Diebold and Rudebusch, 1991).

## 5. Spurious regression.

Consider two variables  $y$  and  $x$ , both of which are highly serially correlated, as are most series in business, finance and economics. Suppose in addition that  $y$  and  $x$  are completely unrelated, but that we don't know they're unrelated, and we regress  $y$  on  $x$  using ordinary least squares.

- a. If the usual regression diagnostics (e.g.,  $R^2$ , t-statistics,  $F$ -statistic) were reliable, we'd expect to see small values of all of them. Why?
- b. In fact the opposite occurs; we tend to see large  $R^2$ , t-, and  $F$ -statistics, and *a very low Durbin-Watson statistic*. Why the low

---

<sup>11</sup>The indexes build on very early work, such as the Harvard “Index of General Business Conditions.” For a fascinating discussion of the early work, see Hardy (1923), Chapter 7.

Durbin-Watson? Why, given the low Durbin-Watson, might you *expect* misleading  $R^2$ ,  $t$ -, and  $F$ -statistics?

- c. This situation, in which highly persistent series that are in fact unrelated nevertheless appear highly related, is called spurious regression. Study of the phenomenon dates to the early twentieth century, and a key study by Granger and Newbold (1974) drove home the prevalence and potential severity of the problem. How might you insure yourself against the spurious regression problem? (Hint: Consider allowing for lagged dependent variables, or dynamics in the regression disturbances, as we've advocated repeatedly.)
6. Comparative forecasting performance of *VARs* and univariate models.
- Using the housing starts and completions data on the book's website, compare the forecasting performance of the VAR used in this chapter to that of the obvious competitor: univariate autoregressions. Use the same in-sample and out-of-sample periods as in the chapter. Why might the forecasting performance of the *VAR* and univariate methods differ? Why might you expect the *VAR* completions forecast to outperform the univariate autoregression, but the *VAR* starts forecast to be no better than the univariate autoregression? Do your results support your conjectures?
7. *VARs* as Reduced Forms of Simultaneous Equations Models.

*VARs* look restrictive in that only *lagged* values appear on the right. That is, the LHS variables are not contemporaneously affected by other variables – instead they are contemporaneously affected only by shocks. That appearance is deceptive, however, as simultaneous equations systems have *VAR* reduced forms. Consider, for example, the simultaneous system

$$(A_0 + A_1 L + \dots + A_p L^p)y_t = v_t$$

$$v_t \sim iid(0, \Omega).$$

Multiplying through by  $A_0^{-1}$  yields

$$(I + A_0^{-1}A_1L + \dots + A_0^{-1}A_pL^p)y_t = \varepsilon_t$$

$$\varepsilon_t \sim iid(0, A_0^{-1}\Omega A_0^{-1'})$$

or

$$(I + \Phi_1L + \dots + \Phi_pL^p)y_t = \varepsilon_t$$

$$\varepsilon_t \sim iid(0, \Sigma)$$

$$\Sigma = A_0^{-1}\Omega A_0^{-1'},$$

which is a standard *VAR*. The *VAR* structure, moreover, is needed for forecasting, as everything on the RHS is lagged by at least one period, making Wold's chain rule immediately applicable.

## 8. Transfer Function Models.

We saw that distributed lag regressions with lagged dependent variables are more general than distributed lag regressions with dynamic disturbances. **Transfer function models** are more general still, and include both as special cases.<sup>12</sup> The basic idea is to exploit the power and parsimony of rational distributed lags in modeling both own-variable and cross-variable dynamics. Imagine beginning with a univariate *ARMA* model,

$$y_t = \frac{C(L)}{D(L)}\varepsilon_t,$$

which captures own-variable dynamics using a rational distributed lag. Now extend the model to capture cross-variable dynamics using a rational distributed lag of the other variable, which yields the general transfer

---

<sup>12</sup>Table 1 displays a variety of important forecasting models, all of which are special cases of the transfer function model.

function model,

$$y_t = \frac{A(L)}{B(L)}x_t + \frac{C(L)}{D(L)}\varepsilon_t.$$

Distributed lag regression with lagged dependent variables is a potentially restrictive special case, which emerges when  $C(L) = 1$  and  $B(L) = D(L)$ . (Verify this for yourself.) Distributed lag regression with *ARMA* disturbances is also a special case, which emerges when  $B(L) = 1$ . (Verify this too.) In practice, the important thing is to allow for own-variable dynamics *somewhat*, in order to account for dynamics in  $y$  not explained by the RHS variables. Whether we do so by including lagged dependent variables, or by allowing for *ARMA* disturbances, or by estimating general transfer function models, can occasionally be important, but usually it's a comparatively minor issue.

9. Cholesky-Factor Identified *VARs* in Matrix Notation.
10. Inflation Forecasting via “Structural” Phillips-Curve Models vs. Time-Series Models.

The literature started with Atkinson and Ohanian \*\*\*\*. The basic result is that Phillips curve information doesn't improve on univariate time series, which is interesting. Also interesting is thinking about why. For example, the univariate time series used is often *IMA*(0, 1, 1) (i.e., exponential smoothing, or local level), which Hendry, Clements and others have argued is robust to shifts. Maybe that's why exponential smoothing is still so powerful after all these years.

11. Multivariate point forecast evaluation.

All univariate absolute standards continue to hold, appropriately interpreted.

- Zero-mean error vector.
- 1-step-ahead errors are vector white noise.

- $h$ -step-ahead errors are at most vector  $MA(h - 1)$ .
- $h$ -step-ahead error covariance matrices are non-decreasing in  $h$ . That is,  $\Sigma_h - \Sigma_{h-1}$  is p.s.d. for all  $h > 1$ .
- The error vector is orthogonal to all available information.

Relative standards, however, need more thinking, as per Christoffersen and Diebold (1998) and Primiceri, Giannone and Lenza (2014).  $trace(MSE)$ ,  $e'Ie$  is not necessarily adequate, and neither is  $e'De$  for diagonal  $d$ ; rather, we generally want  $e'\Sigma e$ , so as to reflect preferences regarding multivariate interactions.

## 12. Multivariate density forecast evaluation

The principle that governs the univariate techniques in this paper extends to the multivariate case, as shown in Diebold, Hahn and Tay (1998). Suppose that the variable of interest  $y$  is now an  $(N \times 1)$  vector, and that we have on hand  $m$  multivariate forecasts and their corresponding multivariate realizations. Further suppose that we are able to decompose each period's forecasts into their conditionals, i.e., for each period's forecasts we can write

$$p(y_{1t}, y_{2t}, \dots, y_{Nt} | \Phi_{t-1}) = p(y_{Nt} | y_{N-1,t}, \dots, y_{1t}, \Phi_{t-1}) \dots p(y_{2t} | y_{1t}, \Phi_{t-1}) p(y_{1t} | \Phi_{t-1}),$$

where  $\Phi_{t-1}$  now refers to the past history of  $(y_{1t}, y_{2t}, \dots, y_{Nt})$ . Then for each period we can transform each element of the multivariate observation  $(y_{1t}, y_{2t}, \dots, y_{Nt})$  by its corresponding conditional distribution. This procedure will produce a set of  $N$   $z$  series that will be *iid*  $U(0, 1)$  individually, and also when taken as a whole, if the multivariate density forecasts are correct. Note that we will have  $N!$  sets of  $z$  series, depending on how the joint density forecasts are decomposed, giving us a wealth of information with which to evaluate the forecasts. In addition, the univariate formula for the adjustment of forecasts, discussed above,

can be applied to each individual conditional, yielding

$$\begin{aligned} f(y_{1t}, y_{2t}, \dots, y_{Nt} | \Phi_{t-1}) &= \prod_{i=1}^N [p(y_{it} | y_{i-1,t}, \dots, y_{1t}, \Phi_{t-1}) q(P(y_{it} | y_{i-1,t}, \dots, y_{1t}, \Phi_{t-1}))] \\ &= p(y_{1t}, y_{2t}, \dots, y_{Nt} | \Phi_{t-1}) q(z_{1t}, z_{2t}, \dots, z_{Nt} | \Phi_{t-1}) . \end{aligned}$$

## 16.9 Notes

Some software, such as Eviews, automatically accounts for parameter uncertainty when forming conditional regression forecast intervals by using variants of the techniques we introduced in Section \*\*\*. Similar but advanced techniques are sometimes used to produce unconditional forecast intervals for dynamic models, such as autoregressions (see Lütkepohl, 1991), but bootstrap simulation techniques are becoming increasingly popular (Efron and Tibshirani, 1993).

Chatfield (1993) argues that innovation uncertainty and parameter estimation uncertainty are likely of minor importance compared to specification uncertainty. We rarely acknowledge specification uncertainty, because we don't know how to quantify "what we don't know we don't know." Quantifying it is a major challenge for future research, and useful recent work in that direction includes Chatfield (1995).

The idea that regression models with serially correlated disturbances are more restrictive than other sorts of transfer function models has a long history in econometrics and engineering and is highlighted in a memorably-titled paper, "Serial Correlation as a Convenient Simplification, not a Nuisance," by Hendry and Mizon (1978). Engineers have scolded econometricians for not using more general transfer function models, as for example in Jenkins (1979). But the fact is, as we've seen repeatedly, that generality for generality's sake in business and economic forecasting is not necessarily helpful, and can be positively harmful. The shrinkage principle asserts that the imposition of

restrictions – even false restrictions – can be helpful in forecasting.

Sims (1980) is an influential paper arguing the virtues of *VARs*. The idea of predictive causality and associated tests in *VARs* is due to Granger (1969) and Sims (1972), who build on earlier work by the mathematician Norbert Weiner. Lütkepohl (1991) is a good reference on *VAR* analysis and forecasting.

Gershenfeld and Weigend (1993) provide a perspective on time series forecasting from the computer-science/engineering/nonlinear/neural-net perspective, and Swanson and White (1995) compare and contrast a variety of linear and nonlinear forecasting methods.

Some slides that might be usefully incorporated:

Univariate  $AR(p)$ :

$$y_t = \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t$$

$$y_t = \phi_1 L y_t + \dots + \phi_p L^p y_t + \varepsilon_t$$

$$(I - \phi_1 L - \dots - \phi_p L^p) y_t = \varepsilon_t$$

$$\phi(L) y_t = \varepsilon_t$$

$$\varepsilon_t \sim iid(0, \sigma^2)$$

But what if we have more than 1 “ $y$ ” variable?

Cross-variable interactions? Leads? Lags? Causality?

$N$ -Variable  $VAR(p)$

$$y_{1t} = \phi_{11}^1 y_{1,t-1} + \dots + \phi_{1N}^1 y_{N,t-1} + \dots + \phi_{11}^p y_{1,t-p} + \dots + \phi_{1N}^p y_{N,t-p} + \varepsilon_{1t}$$

⋮

$$y_{Nt} = \phi_{N1}^1 y_{1,t-1} + \dots + \phi_{NN}^1 y_{N,t-1} + \dots + \phi_{N1}^p y_{1,t-p} + \dots + \phi_{NN}^p y_{N,t-p} + \varepsilon_{Nt}$$

$$\begin{pmatrix} y_{1t} \\ \vdots \\ y_{Nt} \end{pmatrix} = \begin{pmatrix} \phi_{11}^1 & \dots & \phi_{1N}^1 \\ \vdots & & \vdots \\ \phi_{N1}^1 & \dots & \phi_{NN}^1 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ \vdots \\ y_{N,t-1} \end{pmatrix} + \dots + \begin{pmatrix} \phi_{11}^p & \dots & \phi_{1N}^p \\ \vdots & & \vdots \\ \phi_{N1}^p & \dots & \phi_{NN}^p \end{pmatrix} \begin{pmatrix} y_{1,t-p} \\ \vdots \\ y_{N,t-p} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{Nt} \end{pmatrix}$$

$$y_t = \Phi_1 y_{t-1} + \dots + \Phi_p y_{t-p} + \varepsilon_t$$

$$y_t = \Phi_1 L y_t + \dots + \Phi_p L^p y_t + \varepsilon_t$$

$$(I - \Phi_1 L - \dots - \Phi_p L^p) y_t = \varepsilon_t$$

$$\Phi(L)y_t = \varepsilon_t$$

$$\varepsilon_t \sim iid(0, \Sigma)$$

Estimation and Selection

Estimation: Equation-by-equation OLS

Selection: AIC, SIC

$$AIC = \frac{-2\ln L}{T} + \frac{2K}{T}$$

$$SIC = \frac{-2\ln L}{T} + \frac{K\ln T}{T}$$

The Cross-Correlation Function

Recall the univariate autocorrelation function:

$$\rho_y(\tau) = corr(y_t, y_{t-\tau})$$

In multivariate environments we also have

the cross-correlation function:

$$\rho_{yx}(\tau) = corr(y_t, x_{t-\tau})$$

Granger-Sims Causality

Bivariate case:

$y_i$  Granger-Sims causes  $y_j$  if  
 $y_i$  has predictive content for  $y_j$ ,  
*over and above the past history of  $y_j$ .*

Testing:

Are lags of  $y_i$  significant in the  $y_j$  equation?

Impulse-Response Functions in  $AR(1)$  Case

$$y_t = \phi y_{t-1} + \varepsilon_t, \varepsilon_t \sim iid(0, \sigma^2)$$

$$\begin{aligned} \implies y_t &= B(L)\varepsilon_t = \varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \dots \\ &= \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \dots \end{aligned}$$

IRF is  $\{1, \phi, \phi^2, \dots\}$  “dynamic response to a unit shock in  $\varepsilon$ ”

Alternatively write  $\varepsilon_t = \sigma v_t, v_t \sim iid(0, 1)$

$$\implies y_t = \sigma v_t + (\phi\sigma)v_{t-1} + (\phi^2\sigma)v_{t-2} + \dots$$

IRF is  $\{\sigma, \phi\sigma, \phi^2\sigma, \dots\}$  “dynamic response to a one- $\sigma$  shock in  $\varepsilon$ ”

Impulse-Response Functions in  $VAR(p)$  Case

$$y_t = \Phi y_{t-1} + \varepsilon_t, \varepsilon_t \sim iid(0, \Sigma)$$

$$\begin{aligned} \implies y_t &= B(L)\varepsilon_t = \varepsilon_t + B_1\varepsilon_{t-1} + B_2\varepsilon_{t-2} + \dots \\ &= \varepsilon_t + \Phi\varepsilon_{t-1} + \Phi^2\varepsilon_{t-2} + \dots \end{aligned}$$

But we need orthogonal shocks. Why?

So write  $\varepsilon_t = Pv_t, v_t \sim iid(0, I)$ , where  $P$  is Cholesky factor of  $\Sigma$

$$\implies y_t = Pv_t + (\Phi P)v_{t-1} + (\Phi^2 P)v_{t-2} + \dots$$

$ij$ 'th IRF is the sequence of  $ij$ 'th elements of  $\{P, \Phi P, \Phi^2 P, \dots\}$  “Dynamic response of  $y_i$  to a one- $\sigma$  shock in  $\varepsilon_j$ ”



# **Part VI**

# **Appendices**



# Appendix A

## Elements of Probability and Statistics

You've already studied some probability and statistics, but chances are that you could use a bit of review, so we supply it here, with emphasis on ideas that we will use repeatedly. Be warned, however: this section is no substitute for a full introduction to probability and statistics, which you should have had already.

### A.1 Populations: Random Variables, Distributions and Moments

#### A.1.1 Univariate

Consider an experiment with a set  $O$  of possible outcomes. A random variable  $Y$  is simply a mapping from  $O$  to the real numbers. For example, the experiment might be flipping a coin twice, in which case  $O = \{(Heads, Heads), (Tails, Tails), (Heads, Tails), (Tails, Heads)\}$ . We might define a random variable  $Y$  to be the number of heads observed in the two flips, in which case  $Y$  could assume three values,  $y = 0$ ,  $y = 1$  or  $y = 2$ .<sup>1</sup>

**Discrete random variables**, that is, random variables with **discrete probability distributions**, can assume only a countable number of values

---

<sup>1</sup>Note that, in principle, we use capitals for random variables ( $Y$ ) and small letters for their realizations ( $y$ ). We will often neglect this formalism, however, as the meaning will be clear from context.

$y_i$ ,  $i = 1, 2, \dots$ , each with positive probability  $p_i$  such that  $\sum_i p_i = 1$ . The probability distribution  $f(y)$  assigns a probability  $p_i$  to each such value  $y_i$ . In the example at hand,  $Y$  is a discrete random variable, and  $f(y) = 0.25$  for  $y = 0$ ,  $f(y) = 0.50$  for  $y = 1$ ,  $f(y) = 0.25$  for  $y = 2$ , and  $f(y) = 0$  otherwise.

In contrast, **continuous random variables** can assume a continuous range of values, and the **probability density function**  $f(y)$  is a non-negative continuous function such that the area under  $f(y)$  between any points  $a$  and  $b$  is the probability that  $Y$  assumes a value between  $a$  and  $b$ .<sup>2</sup>

In what follows we will simply speak of a “distribution,”  $f(y)$ . It will be clear from context whether we are in fact speaking of a discrete random variable with probability distribution  $f(y)$  or a continuous random variable with probability density  $f(y)$ .

**Moments** provide important summaries of various aspects of distributions. Roughly speaking, moments are simply expectations of powers of random variables, and expectations of different powers convey different sorts of information. You are already familiar with two crucially important moments, the mean and variance. In what follows we’ll consider the first four moments: mean, variance, skewness and kurtosis.<sup>3</sup>

The **mean**, or **expected value**, of a discrete random variable is a probability-weighted average of the values it can assume,<sup>4</sup>

$$E(y) = \sum_i p_i y_i.$$

Often we use the Greek letter  $\mu$  to denote the mean, which measures the **location**, or **central tendency**, of  $y$ .

<sup>2</sup>In addition, the total area under  $f(y)$  must be 1.

<sup>3</sup>In principle, we could of course consider moments beyond the fourth, but in practice only the first four are typically examined.

<sup>4</sup>A similar formula holds in the continuous case.

The **variance** of  $y$  is its expected squared deviation from its mean,

$$\text{var}(y) = E(y - \mu)^2.$$

We use  $\sigma^2$  to denote the variance, which measures the **dispersion, or scale**, of  $y$  around its mean.

Often we assess dispersion using the square root of the variance, which is called the **standard deviation**,

$$\sigma = \text{std}(y) = \sqrt{E(y - \mu)^2}.$$

The standard deviation is more easily interpreted than the variance, because it has the same units of measurement as  $y$ . That is, if  $y$  is measured in dollars (say), then so too is  $\text{std}(y)$ .  $\text{Var}(y)$ , in contrast, would be measured in rather hard-to-grasp units of “dollars squared”.

The **skewness** of  $y$  is its expected cubed deviation from its mean (scaled by  $\sigma^3$  for technical reasons),

$$S = \frac{E(y - \mu)^3}{\sigma^3}.$$

Skewness measures the amount of **asymmetry** in a distribution. The larger the absolute size of the skewness, the more asymmetric is the distribution. A large positive value indicates a long right tail, and a large negative value indicates a long left tail. A zero value indicates symmetry around the mean.

The **kurtosis** of  $y$  is the expected fourth power of the deviation of  $y$  from its mean (scaled by  $\sigma^4$ , again for technical reasons),

$$K = \frac{E(y - \mu)^4}{\sigma^4}.$$

Kurtosis measures the thickness of the tails of a distribution. A kurtosis above three indicates “fat tails” or **leptokurtosis**, relative to the **normal, or Gaussian distribution** that you studied earlier. Hence a kurtosis above

three indicates that extreme events (“tail events”) are more likely to occur than would be the case under normality.

### A.1.2 Multivariate

Suppose now that instead of a single random variable  $Y$ , we have two random variables  $Y$  and  $X$ .<sup>5</sup> We can examine the distributions of  $Y$  or  $X$  in isolation, which are called **marginal distributions**. This is effectively what we’ve already studied. But now there’s more:  $Y$  and  $X$  may be related and therefore move together in various ways, characterization of which requires a **joint distribution**. In the discrete case the joint distribution  $f(y, x)$  gives the probability associated with each possible pair of  $y$  and  $x$  values, and in the continuous case the joint density  $f(y, x)$  is such that the area in any region under it gives the probability of  $(y, x)$  falling in that region.

We can examine the moments of  $y$  or  $x$  in isolation, such as mean, variance, skewness and kurtosis. But again, now there’s more: to help assess the dependence between  $y$  and  $x$ , we often examine a key moment of relevance in multivariate environments, the **covariance**. The covariance between  $y$  and  $x$  is simply the expected product of the deviations of  $y$  and  $x$  from their respective means,

$$\text{cov}(y, x) = E[(y_t - \mu_y)(x_t - \mu_x)].$$

A positive covariance means that  $y$  and  $x$  are positively related; that is, when  $y$  is above its mean  $x$  tends to be above its mean, and when  $y$  is below its mean  $x$  tends to be below its mean. Conversely, a negative covariance means that  $y$  and  $x$  are inversely related; that is, when  $y$  is below its mean  $x$  tends to be above its mean, and vice versa. The covariance can take any value in the real numbers.

---

<sup>5</sup>We could of course consider more than two variables, but for pedagogical reasons we presently limit ourselves to two.

Frequently we convert the covariance to a **correlation** by standardizing by the product of  $\sigma_y$  and  $\sigma_x$ ,

$$\text{corr}(y, x) = \frac{\text{cov}(y, x)}{\sigma_y \sigma_x}.$$

The correlation takes values in  $[-1, 1]$ . Note that covariance depends on units of measurement (e.g., dollars, cents, billions of dollars), but correlation does not. Hence correlation is more immediately interpretable, which is the reason for its popularity.

Note also that covariance and correlation measure only *linear* dependence; in particular, a zero covariance or correlation between  $y$  and  $x$  does not necessarily imply that  $y$  and  $x$  are independent. That is, they may be *non-linearly* related. If, however, two random variables are jointly *normally* distributed with zero covariance, then they are independent.

Our multivariate discussion has focused on the joint distribution  $f(y, x)$ . In various chapters we will also make heavy use of the **conditional distribution**  $f(y|x)$ , that is, the distribution of the random variable  $Y$  *conditional* upon  $X = x$ . **Conditional moments** are similarly important. In particular, the **conditional mean** and **conditional variance** play key roles in econometrics, in which attention often centers on the mean or variance of a series conditional upon the past.

## A.2 Samples: Sample Moments

### A.2.1 Univariate

Thus far we've reviewed aspects of known distributions of random variables, in **population**. Often, however, we have a **sample** of data drawn from an unknown population distribution  $f$ ,

$$\{y_i\}_{i=1}^N \sim f(y),$$

and we want to learn from the sample about various aspects of  $f$ , such as its moments. To do so we use various **estimators**.<sup>6</sup> We can obtain estimators by replacing population expectations with sample averages, because the arithmetic average is the sample analog of the population expectation. Such “analog estimators” turn out to have good properties quite generally. The **sample mean** is simply the arithmetic average,

$$\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i.$$

It provides an empirical measure of the location of  $y$ .

The **sample variance** is the average squared deviation from the sample mean,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^N (y_i - \bar{y})^2}{N}.$$

It provides an empirical measure of the dispersion of  $y$  around its mean.

We commonly use a slightly different version of  $\hat{\sigma}^2$ , which corrects for the one degree of freedom used in the estimation of  $\bar{y}$ , thereby producing an unbiased estimator of  $\sigma^2$ ,

$$s^2 = \frac{\sum_{i=1}^N (y_i - \bar{y})^2}{N - 1}.$$

Similarly, the **sample standard deviation** is defined either as

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{\sum_{i=1}^N (y_i - \bar{y})^2}{N}}$$

or

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^N (y_i - \bar{y})^2}{N - 1}}.$$

It provides an empirical measure of dispersion in the same units as  $y$ .

---

<sup>6</sup>An estimator is an example of a **statistic**, or **sample statistic**, which is simply a function of the sample observations.

The **sample skewness** is

$$\hat{S} = \frac{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^3}{\hat{\sigma}^3}.$$

It provides an empirical measure of the amount of asymmetry in the distribution of  $y$ .

The **sample kurtosis** is

$$\hat{K} = \frac{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^4}{\hat{\sigma}^4}.$$

It provides an empirical measure of the fatness of the tails of the distribution of  $y$  relative to a normal distribution.

Many of the most famous and important statistical sampling distributions arise in the context of sample moments, and the normal distribution is the father of them all. In particular, the celebrated central limit theorem establishes that under quite general conditions the sample mean  $\bar{y}$  will have a normal distribution as the sample size gets large. The  $\chi^2$  **distribution** arises from squared normal random variables, the  $t$  **distribution** arises from ratios of normal and  $\chi^2$  variables, and the  $F$  **distribution** arises from ratios of  $\chi^2$  variables. Because of the fundamental nature of the normal distribution as established by the central limit theorem, it has been studied intensively, a great deal is known about it, and a variety of powerful tools have been developed for use in conjunction with it.

Because of the fundamental nature of the normal distribution as established by the central limit theorem, it has been studied intensively, a great deal is known about it, and a variety of powerful tools have been developed for use in conjunction with it. Hence it is often of interest to assess whether the normal distribution governs a given sample of data. A simple strategy is to check various implications of normality, such as  $S = 0$  and  $K = 3$ , via informal examination of  $\hat{S}$  and  $\hat{K}$ . Alternatively and more formally, the

**Jarque-Bera test** (JB) effectively aggregates the information in the data about both skewness and kurtosis to produce an overall test of the hypothesis that  $S = 0$  and  $K = 3$ , based upon  $\hat{S}$  and  $\hat{K}$ . The test statistic is

$$JB = \frac{T}{6} \left( \hat{S}^2 + \frac{1}{4}(\hat{K} - 3)^2 \right),$$

where  $T$  is the number of observations. Under the null hypothesis of *iid* Gaussian observations, the Jarque-Bera statistic is distributed in large samples as a  $\chi^2$  random variable with two degrees of freedom.<sup>7</sup>

### A.2.2 Multivariate

We also have sample versions of moments of multivariate distributions. In particular, the **sample covariance** is

$$\widehat{\text{cov}}(y, x) = \frac{1}{N} \sum_{i=1}^N [(y_i - \bar{y})(x_i - \bar{x})],$$

and the **sample correlation** is

$$\widehat{\text{corr}}(y, x) = \frac{\widehat{\text{cov}}(y, x)}{\hat{\sigma}_y \hat{\sigma}_x}.$$

## A.3 Finite-Sample and Asymptotic Sampling Distributions of the Sample Mean

Here we refresh your memory on the sampling distribution of the most important sample moment, the sample mean.

---

<sup>7</sup>Other tests of conformity to the normal distribution exist and may of course be used, such as the Kolmogorov-Smirnov test. The Jarque-Bera test, however, has the convenient and intuitive decomposition into skewness and kurtosis components.

### A.3.1 Exact Finite-Sample Results

In your earlier studies you learned about *statistical inference*, such as how to form confidence intervals for the population mean based on the sample mean, how to test hypotheses about the population mean, and so on. Here we partially refresh your memory.

Consider the benchmark case of Gaussian **simple random sampling**,

$$y_i \sim iid N(\mu, \sigma^2), i = 1, \dots, N,$$

which corresponds to a special case of what we will later call the “full ideal conditions” for regression modeling. The sample mean  $\bar{y}$  is the natural estimator of the population mean  $\mu$ . In this case, as you learned earlier,  $\bar{y}$  is unbiased, consistent, normally distributed with variance  $\sigma^2/N$ , and indeed the minimum variance unbiased (MVUE) estimator. We write

$$\bar{y} \sim N\left(\mu, \frac{\sigma^2}{N}\right),$$

or equivalently

$$\sqrt{N}(\bar{y} - \mu) \sim N(0, \sigma^2).$$

We construct exact finite-sample confidence intervals for  $\mu$  as

$$\mu \in \left[ \bar{y} \pm t_{1-\frac{\alpha}{2}}(N-1) \frac{s}{\sqrt{N}} \right] \text{ w.p. } \alpha,$$

where  $t_{1-\frac{\alpha}{2}}(N-1)$  is the  $1 - \frac{\alpha}{2}$  percentile of a  $t$  distribution with  $N-1$  degrees of freedom. Similarly, we construct exact finite-sample (likelihood ratio) hypothesis tests of  $H_0 : \mu = \mu_0$  against the two-sided alternative  $H_0 : \mu \neq \mu_0$  using

$$\frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{N}}} \sim t_{1-\frac{\alpha}{2}}(N-1).$$

### A.3.2 Approximate Asymptotic Results (Under Weaker Assumptions)

Much of statistical inference is linked to large-sample considerations, such as the law of large numbers and the central limit theorem, which you also studied earlier. Here we again refresh your memory.

Consider again a simple random sample, but without the normality assumption,

$$y_i \sim iid(\mu, \sigma^2), i = 1, \dots, N.$$

Despite our dropping the normality assumption we still have that  $\bar{y}$  is unbiased, consistent, **asymptotically** normally distributed with variance  $\sigma^2/N$ , and best linear unbiased (BLUE). We write,

$$\bar{y} \xrightarrow{a} N\left(\mu, \frac{\sigma^2}{N}\right).$$

More precisely, as  $T \rightarrow \infty$ ,

$$\sqrt{N}(\bar{y} - \mu) \rightarrow_d N(0, \sigma^2).$$

This result forms the basis for asymptotic inference. It is a Gaussian central limit theorem, and it also has a law of large numbers ( $\bar{y} \rightarrow_p \mu$ ) imbedded within it.

We construct asymptotically-valid confidence intervals for  $\mu$  as

$$\mu \in \left[ \bar{y} \pm z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{N}} \right] \text{ w.p. } \alpha,$$

where  $z_{1-\frac{\alpha}{2}}$  is the  $1 - \frac{\alpha}{2}$  percentile of a  $N(0, 1)$  distribution. Similarly, we construct asymptotically-valid hypothesis tests of  $H_0 : \mu = \mu_0$  against the

two-sided alternative  $H_0 : \mu \neq \mu_0$  using

$$\frac{\bar{y} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{N}}} \sim N(0, 1).$$

## A.4 Exercises, Problems and Complements

### 1. (Interpreting distributions and densities)

The Sharpe Pencil Company has a strict quality control monitoring program. As part of that program, it has determined that the distribution of the amount of graphite in each batch of one hundred pencil leads produced is continuous and uniform between one and two grams. That is,  $f(y) = 1$  for  $y$  in  $[1, 2]$ , and zero otherwise, where  $y$  is the graphite content per batch of one hundred leads.

- a. Is  $y$  a discrete or continuous random variable?
  - b. Is  $f(y)$  a probability distribution or a density?
  - c. What is the probability that  $y$  is between 1 and 2? Between 1 and 1.3? Exactly equal to 1.67?
  - d. For high-quality pencils, the desired graphite content per batch is 1.8 grams, with low variation across batches. With that in mind, discuss the nature of the density  $f(y)$ .
2. (Covariance and correlation)

Suppose that the annual revenues of world's two top oil producers have a covariance of 1,735,492.

- a. Based on the covariance, the claim is made that the revenues are “very strongly positively related.” Evaluate the claim.
- b. Suppose instead that, again based on the covariance, the claim is made that the revenues are “positively related.” Evaluate the claim.

- c. Suppose you learn that the revenues have a *correlation* of 0.93. In light of that new information, re-evaluate the claims in parts a and b above.

3. (Simulation)

You will often need to simulate data from various models. The simplest model is the  $iidN(\mu, \sigma^2)$  (Gaussian simple random sampling) model.

- a. Using a random number generator, simulate a sample of size 30 for  $y$ , where  $y \sim iidN(0, 1)$ .
- b. What is the sample mean? Sample standard deviation? Sample skewness? Sample kurtosis? Discuss.
- c. Form an appropriate 95 percent confidence interval for  $E(y)$ .
- d. Perform a  $t$  test of the hypothesis that  $E(y) = 0$ .
- e. Perform a  $t$  test of the hypothesis that  $E(y) = 1$ .

4. (Sample moments of the wage data)

Use the 1995 wage dataset.

- a. Calculate the sample mean wage and test the hypothesis that it equals \$9/hour.
- b. Calculate sample skewness.
- c. Calculate and discuss the sample correlation between wage and years of education.

5. Notation.

We have used standard cross-section notation:  $i = 1, \dots, N$ . The standard time-series notation is  $t = 1, \dots, T$ . Much of our discussion will be valid in *both* cross-section and time-series environments, but still we have to pick a notation. Without loss of generality, henceforth we will typically use  $t = 1, \dots, T$ .

## A.5 Notes

Numerous good introductory probability and statistics books exist. [Wonnacott and Wonnacott \(1990\)](#) remains a time-honored classic, which you may wish to consult to refresh your memory on statistical distributions, estimation and hypothesis testing. [Anderson et al. \(2008\)](#) is a well-written recent text.



# Appendix B

## Elements of Nonparametrics

### B.1 Density Estimation

#### B.1.1 The Basic Problem

$$\begin{aligned} & \text{iid} \\ \{x_i\}_{i=1}^N & \sim f(x) \end{aligned}$$

$f$  smooth in  $[x_0 - h, x_0 + h]$

Goal: Estimate  $f(x)$  at arbitrary point  $x = x_0$

By the mean-value theorem,

$$f(x_0) \approx \frac{1}{2h} \int_{x_0-h}^{x_0+h} f(u) du = \frac{1}{2h} P(x \in [x_0 - h, x_0 + h])$$

Estimate  $P(x \in [x_0 - h, x_0 + h])$  by  $\frac{\#x_i \in [x_0 - h, x_0 + h]}{N}$

$$\begin{aligned} \hat{f}_h(x_0) &= \frac{1}{2h} \frac{\#x_i \in [x_0 - h, x_0 + h]}{N} \\ &= \frac{1}{Nh} \sum_{i=1}^N \frac{1}{2} I\left(\left|\frac{x_0 - x_i}{h}\right| \leq 1\right) \end{aligned}$$

“Rosenblatt estimator”

Kernel density estimator with

kernel:  $K(u) = \frac{1}{2}I(|u| \leq 1)$

bandwidth:  $h$

### B.1.2 Kernel Density Estimation

Issues with uniform kernels:

1. Why weight distant observations as heavily as nearby ones?
2. Why use a discontinuous kernel if we think that  $f$  is smooth?

Obvious solution: Choose *smooth* kernel

Standard conditions:

$$\int K(u)du = 1$$

$$K(u) = K(-u)$$

Common Kernel Choices

Standard normal:  $K(u) = \frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$

Triangular  $K(u) = (1 - |u|)I(|u| \leq 1)$

Epinechnikov:  $K(u) = \frac{3}{4}(1 - u^2)I(|u| \leq 1)$

General Form of the Kernel Density Estimator

$$\hat{f}_h(x_0) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x_0 - x_i}{h}\right)$$

“Rosenblatt-Parzen estimator”

Figure B.1: Bandwidth Choice – from Silverman (1986)

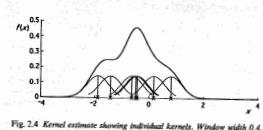


Fig. 2.4 Kernel estimate showing individual kernels. Window width 0.4.

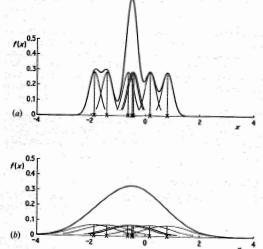


Fig. 2.5 Kernel estimates showing individual kernels. Window widths: (a) 0.2; (b) 0.8;

### B.1.3 Bias-Variance Tradeoffs

Inescapable Bias-Variance Tradeoff (in Practice, Fixed  $N$ )

Escapable Bias-Variance Tradeoff (in Theory,  $N \rightarrow \infty$ )

$$E(\hat{f}_h(x_0)) \approx f(x_0) + \frac{h^2}{2} \cdot O_p(1)$$

$$(So h \rightarrow 0 \implies bias \rightarrow 0)$$

$$var(\hat{f}_h(x_0)) \approx \frac{1}{Nh} \cdot O_p(1)$$

$$(So Nh \rightarrow \infty \implies var \rightarrow 0)$$

Thus,

$$\left. \begin{array}{l} h \rightarrow 0 \\ Nh \rightarrow \infty \end{array} \right\} \implies \hat{f}_h(x_0) \xrightarrow{p} f(x_0)$$

### Convergence Rate

$$\sqrt{Nh}(\hat{f}_h(x_0) - f(x_0)) \xrightarrow{d} D$$

Effects of  $K$  minor; effects of  $h$  major.

### B.1.4 Optimal Bandwidth Choice

$$MSE(\hat{f}_h(x_0)) = E(\hat{f}_h(x_0) - f(x_o))^2$$

$$IMSE = \int MSE(\hat{f}_h(x_0)) f(x) dx$$

Choose bandwidth to minimize IMSE:

$$h^* = \gamma^* N^{-1/5}$$

Corresponding Optimal Convergence Rate

Recall:

$$\sqrt{Nh} (\hat{f}_h(x_0) - f(x_0)) \xrightarrow{d} D$$

$$h^* \propto N^{-1/5}$$

Substituting yields the best obtainable rate:

$$\sqrt{N^{4/5}} (\hat{f}_h(x_0) - f(x_0)) \xrightarrow{d} D$$

“Stone optimal rate”

Silverman’s Rule

For the Gaussian case,

$$h^* = 1.06\sigma N^{-1/5}$$

So use:

$$\hat{h}^* = 1.06\hat{\sigma} N^{-1/5}$$

Better to err on the side of too little smoothing:

$$\hat{h}^* = \hat{\sigma} N^{-1/5}$$

## B.2 Multivariate

Earlier univariate kernel density estimator:

$$\hat{f}_h(x_0) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x_0 - x_i}{h}\right)$$

Can be written as:

$$\hat{f}_h(x_0) = \frac{1}{N} \sum_{i=1}^N K_h(x_0 - x_i)$$

where  $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$

or  $K_h(\cdot) = h^{-1}K(h^{-1}\cdot)$

Multivariate Version ( $d$ -Dimensional)

Precisely follows equation (B.2):

$$\hat{f}_H(x_0) = \frac{1}{N} \sum_{i=1}^N K_H(x_0 - x_i),$$

where  $K_H(\cdot) = |H|^{-1}K(H^{-1}\cdot)$ , and  $H$  ( $d \times d$ ) is psd.

Common choice:  $K(u) = N(0, I)$ ,  $H = hI$

$$\implies K_H(\cdot) = \frac{1}{h^d}K\left(\frac{1}{h}\cdot\right) = \frac{1}{h^d}K\left(\frac{x_0 - x_i}{h}\right)$$

$$\implies \hat{f}_h(x_0) = \frac{1}{Nh^d} \sum_{i=1}^N K\left(\frac{x_0 - x_i}{h}\right)$$

Bias-Variance Tradeoff, Convergence Rate, Optimal Bandwidth, Corresponding Optimal Convergence Rate

$$\left. \begin{array}{l} h \rightarrow 0 \\ Nh^d \rightarrow \infty \end{array} \right\} \implies \hat{f}_h(x_0) \xrightarrow{p} f(x_0)$$

$$\sqrt{Nh^d} \left( \hat{f}_h(x_0) - f(x_0) \right) \xrightarrow{d} D$$

$$h^* \propto N^{-\frac{1}{d+4}}$$

$$\sqrt{N^{1-\frac{d}{d+4}}} \left( \hat{f}_h(x_0) - f(x_0) \right) \xrightarrow{d} D$$

Stone-optimal rate drops with  $d$

“Curse of dimensionality”

Silverman’s Rule

$$\hat{h}^* = \left( \frac{4}{d+2} \right)^{\frac{1}{d+4}} \hat{\sigma} N^{-\frac{1}{d+4}}$$

where

$$\hat{\sigma}^2 = \frac{1}{d} \sum_{i=1}^d \hat{\sigma}_i^2$$

(average sample variance)

## B.3 Functional Estimation

Conditional Mean (Regression)

$$E(y|x) = M(x) = \int y \frac{f(y,x)}{f(x)} dy$$

Regression Slope

$$\beta(x) = \frac{\partial M(x)}{\partial x_j} = \lim_{h \rightarrow 0} \frac{(M(x + \frac{h}{2}) - M(x - \frac{h}{2}))}{h}$$

Regression Disturbance Density

$$f(u), \quad u = y - M(x)$$

Conditional Variance

$$var(y|x) = V(x) = \int y^2 \frac{f(y,x)}{f(x)} dy - M(x)^2$$

Hazard Function

$$\lambda(t) = \frac{f(t)}{1 - F(t)}$$

Curvature (Higher-Order Derivative Estimation)

$$C(x) = \frac{\partial}{\partial x_j} \beta(x) = \left( \frac{\partial^2}{\partial x_j^2} \right) M(x) = \lim_{h \rightarrow 0} \frac{\beta(x + \frac{h}{2}) - \beta(x - \frac{h}{2})}{h}$$

The curse of dimensionality is much worse for curvature...

$d$ -vector:  $r = (r_1, \dots, r_d)$ ,  $|r| = \sum_{i=1}^d r_i$

Define  $M^{(r)}(x) \equiv \partial^{\frac{|r|}{\partial^{r_1} x_1, \dots, \partial^{r_d} x_d}} M(x)$

Then  $\sqrt{Nh}^{2|r|+d} [\hat{M}^{(r)}(x_0) - M^{(r)}(x_0)] \rightarrow_d D$

## B.4 Local Nonparametric Regression

### B.4.1 Kernel Regression

$$M(x_0) = \int y f(y|x_0) dy = \int y \frac{f(x_0, y)}{f(x_0)} dy$$

Using multivariate kernel density estimates and manipulating gives the “Nadaraya-Watson” estimator:

$$\hat{M}_h(x_0) = \sum_{i=1}^N \left[ \frac{K\left(\frac{x_0-x_i}{h}\right)}{\sum_{i=1}^N K\left(\frac{x_0-x_i}{h}\right)} \right] y_i$$

$$h \rightarrow 0, \quad Nh \rightarrow \infty \implies$$

$$\sqrt{Nh^d} (\hat{M}_h(x_0) - M(x_0)) \xrightarrow{d} N(0, V)$$

### B.4.2 Nearest-Neighbor Regression

#### Basic Nearest-Neighbor Regression

$$\hat{M}_k(x_0) = \frac{1}{k} \sum_{i \in n(x_0)} y_i \text{ (Locally Constant, uniform weighting)}$$

$$k \rightarrow \infty, \quad \frac{k}{N} \rightarrow 0 \quad \Rightarrow \quad \hat{M}_k(x_0) \xrightarrow{P} M(x_0)$$

$$\sqrt{k} (\hat{M}_k(x_0) - M(x_0)) \xrightarrow{d} D$$

Equivalent to Nadaraya-Watson kernel regression with:

$$K(u) = \frac{1}{2} I(|u| \leq 1) \text{ (uniform)}$$

and  $h = R(k)$  (distance from  $x_0$  to  $k^{th}$  nearest neighbor)  
 $\Rightarrow$  Variable bandwidth!

### Locally-Weighted Nearest-Neighbor Regression (Locally Polynomial, Non-Uniform Weighting)

$$y_t = g(x_t) + \varepsilon_t$$

Computation of  $\hat{g}(x^*)$ :

$$0 < \xi \leq 1$$

$$k_T = \text{int}(\xi \cdot T)$$

Find  $K_T$  nearest neighbors using norm:

$$\lambda(x^*, x_{k_T}^*) = [\sum_{j=1}^P (x_{k_T j}^* - x_j^*)^2]^{\frac{1}{2}}$$

Neighborhood weight function:

$$v_t(x_t, x^*, x_{k_T}^*) = C\left(\frac{\lambda(x_t, x^*)}{\lambda(x^*, x_{k_T}^*)}\right)$$

$$C(u) = \begin{cases} (1-u^3)^3 & \text{for } u < 1 \\ 0 & \text{otherwise} \end{cases}$$

## B.5 Global Nonparametric Regression

### B.5.1 Series (Sieve, Projection, ...)

$$M(x_0) = \sum_{j=0}^{\infty} \beta_j \phi_j(x_0)$$

(the  $\phi_j$  are orthogonal basis functions)

$$\hat{M}_J(x_0) = \sum_{j=0}^J \hat{\beta}_j \phi_j(x_0)$$

$$J \rightarrow \infty, \frac{J}{N} \rightarrow 0 \Rightarrow \hat{M}_J(x_0) \xrightarrow{P} M(x_0)$$

Stone-optimal convergence rate, for suitable choice of  $J$ .

### B.5.2 Neural Networks

Run linear combinations of inputs through “squashing functions”  $i = 1, \dots, R$  inputs,  $j = 1, \dots, S$  neurons

$$h_{jt} = \Psi(\gamma_{jo} + \sum_{i=1}^R \gamma_{ij} x_{it}), \quad j = 1, \dots, S \quad (\text{Neuron } j)$$

e.g.  $\Psi(\cdot)$  can be logistic (regression), 0-1 (classification)

$$O_t = \Phi(\beta_0 + \sum_{j=1}^S \beta_j h_{jt})$$

e.g.  $\Phi(\cdot)$  can be the identity function

$$\text{Compactly: } O_t = \Phi(\beta_0 + \sum_{j=1}^S \beta_j \Psi(\gamma_{jo} + \sum_{i=1}^R \gamma_{ij} x_{it})) \equiv f(x_t; \theta)$$

Universal Approximator:  $S \rightarrow \infty, \frac{S}{N} \rightarrow 0 \Rightarrow \hat{O}(x_0) \rightarrow_p O(x_0)$

Same as other nonparametric methods.

### B.5.3 More

Ace, projection pursuit, regression splines, smoothing splines, CART,

## B.6 Time Series Aspects

1. Many results go through under mixing or Markov conditions.
2. Recursive kernel regression.

Use recursive kernel estimator:

$$\hat{f}_N(x_0) = \left(\frac{N-1}{N}\right) f_{N-1}(x_0) + \frac{1}{Nh^d} K\left(\frac{x_0 - x_N}{h}\right)$$

to get:

$$\hat{M}_N(x_0) = \frac{(N-1)h^d \hat{f}_{N-1}(x_0) \hat{M}_{N-1}(x_0) + Y_N K\left(\frac{x_0 - x_N}{h}\right)}{(N-1)h^d \hat{f}_{N-1}(x_0) + K\left(\frac{x_0 - x_N}{h}\right)}$$

3. Bandwidth selection via recursive prediction.

4. Nonparametric nonlinear autoregression.

$$y_t = g(y_{t-1}, \dots, y_{t-p}) + \varepsilon_t$$

$$\begin{aligned} E(y_{t+1} | y_t, \dots, y_{t-p+1}) &= \int y_{t+1} f(y_{t+1} | y_t, \dots, y_{t-p+1}) dy \\ &= \int y_{t+1} \frac{f(y_{t+1}, \dots, y_{t-p+1})}{f(y_t, \dots, y_{t-p+1})} dy \end{aligned}$$

Implementation: Kernel, Series, NN, LWR

5. Recurrent neural nets.

$$h_{jt} = \Psi(\gamma_{j0} + \sum_{i=1}^R \gamma_{ij} x_{it} + \sum_{l=1}^S \delta_{jl} h_{l, t-1}), \quad j = 1, \dots, S$$

$$O_t = \Phi(\beta_0 + \sum_{j=1}^S \beta_j h_{jt})$$

Compactly:  $O_t = \Phi(\beta_0 + \sum_{j=1}^S \beta_j \Psi(\gamma_{j0} + \sum_{i=1}^R \gamma_{ij} x_{it} + \sum_{l=1}^S \delta_{jl} h_{l, t-1}))$

Back substitution:

$$O_t = g(x_t, x_{t-1}, \dots, x_1; \theta)$$

## B.7 Exercises, Problems and Complements

1. Tightly parametric models are often best for time-series prediction.

Generality isn't so great; restrictions often help!

2. Semiparametric and related approaches.

$\sqrt{N}$  consistent estimation. Adaptive estimation.

## B.8 Notes



# Appendix C

## “Problems and Complements” Data

Here we provide data for the in-chapter examples as well as end-of-chapter EPC’s. The data are also available on the web.

### C.1 Liquor Sales

480 467 514 505 534 546 539 541 551 537 584 854 522 506 558 538 605 583  
607 624 570 609 675 861 605 537 575 588 656 623 661 668 603 639 669 915  
643 563 616 645 703 684 731 722 678 713 725 989 687 629 687 706 754 774  
825 755 751 783 804 1139 711 693 790 754 799 824 854 810 798 807 832 1142  
740 713 791 768 846 884 886 878 813 840 884 1245 796 750 834 838 902 895  
962 990 882 936 997 1305 866 805 905 873 1024 985 1049 1034 951 1010 1016  
1378 915 854 922 965 1014 1040 1137 1026 992 1052 1056 1469 916 934 987  
1018 1048 1086 1144 1077 1036 1076 1114 1595 949 930 1045 1015 1091 1142  
1182 1161 1145 1119 1189 1662 1048 1019 1129 1092 1176 1297 1322 1330  
1263 1250 1341 1927 1271 1238 1283 1283 1413 1371 1425 1453 1311 1387  
1454 1993 1328 1250 1308 1350 1455 1442 1530 1505 1421 1485 1465 2163  
1361 1284 1392 1442 1504 1488 1606 1488 1442 1495 1509 2135 1369 1320  
1448 1495 1522 1575 1666 1617 1567 1551 1624 2367 1377 1294 1401 1362  
1466 1559 1569 1575 1456 1487 1549 2178 1423 1312 1465 1488 1577 1591  
1669 1697 1659 1597 1728 2326 1529 1395 1567 1536 1682 1675 1758 1708

1561 1643 1635 2240 1485 1376 1459 1526 1659 1623 1731 1662 1589 1683  
1672 2361 1480 1385 1505 1576 1649 1684 1748 1642 1571 1567 1637 2397  
1483 1390 1562 1573 1718 1752 1809 1759 1698 1643 1718 2399 1551 1497  
1697 1672 1805 1903 1928 1963 1807 1843 1950 2736 1798 1700 1901 1820  
1982 1957 2076 2107 1799 1854 1968 2364 1662 1681 1725 1796 1938 1871  
2001 1934 1825 1930 1867 2553 1624 1533 1676 1706 1781 1772 1922 1743  
1669 1713 1733 2369 1491 1445 1643 1683 1751 1774 1893 1776 1743 1728  
1769 2431

## C.2 Housing Starts and Completions

”OBS” ”STARTS” ”COMPS”

”1968M01” 1.38 1.257  
”1968M02” 1.52 1.174  
”1968M03” 1.466 1.323  
”1968M04” 1.554 1.328  
”1968M05” 1.408 1.367  
”1968M06” 1.405 1.184  
”1968M07” 1.512 1.37  
”1968M08” 1.495 1.279  
”1968M09” 1.556 1.397  
”1968M10” 1.569 1.348  
”1968M11” 1.63 1.367  
”1968M12” 1.548 1.39  
”1969M01” 1.769 1.257  
”1969M02” 1.705 1.414  
”1969M03” 1.561 1.558  
”1969M04” 1.524 1.318  
”1969M05” 1.583 1.43

”1969M06” 1.528 1.455  
”1969M07” 1.368 1.432  
”1969M08” 1.358 1.393  
”1969M09” 1.507 1.367  
”1969M10” 1.381 1.406  
”1969M11” 1.229 1.404  
”1969M12” 1.327 1.402  
”1970M01” 1.085 1.434  
”1970M02” 1.305 1.43  
”1970M03” 1.319 1.317  
”1970M04” 1.264 1.354  
”1970M05” 1.29 1.334  
”1970M06” 1.385 1.431  
”1970M07” 1.517 1.384  
”1970M08” 1.399 1.609  
”1970M09” 1.534 1.383  
”1970M10” 1.58 1.437  
”1970M11” 1.647 1.457  
”1970M12” 1.893 1.437  
”1971M01” 1.828 1.471  
”1971M02” 1.741 1.448  
”1971M03” 1.91 1.489  
”1971M04” 1.986 1.709  
”1971M05” 2.049 1.637  
”1971M06” 2.026 1.637  
”1971M07” 2.083 1.699  
”1971M08” 2.158 1.896  
”1971M09” 2.041 1.804  
”1971M10” 2.128 1.815

”1971M11” 2.182 1.844  
”1971M12” 2.295 1.895  
”1972M01” 2.494 1.942  
”1972M02” 2.39 2.061  
”1972M03” 2.334 1.981  
”1972M04” 2.249 1.97  
”1972M05” 2.221 1.896  
”1972M06” 2.254 1.936  
”1972M07” 2.252 1.93  
”1972M08” 2.382 2.102  
”1972M09” 2.481 2.053  
”1972M10” 2.485 1.995  
”1972M11” 2.421 1.985  
”1972M12” 2.366 2.121  
”1973M01” 2.481 2.162  
”1973M02” 2.289 2.124  
”1973M03” 2.365 2.196  
”1973M04” 2.084 2.195  
”1973M05” 2.266 2.299  
”1973M06” 2.067 2.258  
”1973M07” 2.123 2.066  
”1973M08” 2.051 2.056  
”1973M09” 1.874 2.061  
”1973M10” 1.677 2.052  
”1973M11” 1.724 1.925  
”1973M12” 1.526 1.869  
”1974M01” 1.451 1.932  
”1974M02” 1.752 1.938  
”1974M03” 1.555 1.806

”1974M04” 1.607 1.83  
”1974M05” 1.426 1.715  
”1974M06” 1.513 1.897  
”1974M07” 1.316 1.695  
”1974M08” 1.142 1.634  
”1974M09” 1.15 1.651  
”1974M10” 1.07 1.63  
”1974M11” 1.026 1.59  
”1974M12” 0.975 1.54  
”1975M01” 1.032 1.588  
”1975M02” 0.904 1.346  
”1975M03” 0.993 1.293  
”1975M04” 1.005 1.278  
”1975M05” 1.121 1.349  
”1975M06” 1.087 1.234  
”1975M07” 1.226 1.276  
”1975M08” 1.26 1.29  
”1975M09” 1.264 1.333  
”1975M10” 1.344 1.134  
”1975M11” 1.36 1.383  
”1975M12” 1.321 1.306  
”1976M01” 1.367 1.258  
”1976M02” 1.538 1.311  
”1976M03” 1.421 1.347  
”1976M04” 1.395 1.332  
”1976M05” 1.459 1.44  
”1976M06” 1.495 1.39  
”1976M07” 1.401 1.322  
”1976M08” 1.55 1.374

”1976M09” 1.72 1.371  
”1976M10” 1.629 1.388  
”1976M11” 1.641 1.428  
”1976M12” 1.804 1.457  
”1977M01” 1.527 1.457  
”1977M02” 1.943 1.655  
”1977M03” 2.063 1.619  
”1977M04” 1.892 1.548  
”1977M05” 1.971 1.555  
”1977M06” 1.893 1.636  
”1977M07” 2.058 1.687  
”1977M08” 2.02 1.673  
”1977M09” 1.949 1.865  
”1977M10” 2.042 1.675  
”1977M11” 2.042 1.77  
”1977M12” 2.142 1.634  
”1978M01” 1.718 1.777  
”1978M02” 1.738 1.719  
”1978M03” 2.032 1.785  
”1978M04” 2.197 1.843  
”1978M05” 2.075 1.85  
”1978M06” 2.07 1.905  
”1978M07” 2.092 1.957  
”1978M08” 1.996 1.976  
”1978M09” 1.97 1.944  
”1978M10” 1.981 1.885  
”1978M11” 2.094 1.877  
”1978M12” 2.044 1.844  
”1979M01” 1.63 1.85

”1979M02” 1.52 1.845  
”1979M03” 1.847 1.946  
”1979M04” 1.748 1.866  
”1979M05” 1.876 2.007  
”1979M06” 1.913 1.853  
”1979M07” 1.76 1.759  
”1979M08” 1.778 1.779  
”1979M09” 1.832 1.983  
”1979M10” 1.681 1.832  
”1979M11” 1.524 1.892  
”1979M12” 1.498 1.863  
”1980M01” 1.341 1.794  
”1980M02” 1.35 1.803  
”1980M03” 1.047 1.701  
”1980M04” 1.051 1.751  
”1980M05” 0.927 1.532  
”1980M06” 1.196 1.48  
”1980M07” 1.269 1.472  
”1980M08” 1.436 1.44  
”1980M09” 1.471 1.267  
”1980M10” 1.523 1.272  
”1980M11” 1.51 1.313  
”1980M12” 1.482 1.378  
”1981M01” 1.547 1.27  
”1981M02” 1.246 1.395  
”1981M03” 1.306 1.377  
”1981M04” 1.36 1.469  
”1981M05” 1.14 1.246  
”1981M06” 1.045 1.35

”1981M07” 1.041 1.337  
”1981M08” 0.94 1.222  
”1981M09” 0.911 1.221  
”1981M10” 0.873 1.206  
”1981M11” 0.837 1.074  
”1981M12” 0.91 1.129  
”1982M01” 0.843 1.052  
”1982M02” 0.866 0.935  
”1982M03” 0.931 0.965  
”1982M04” 0.917 0.979  
”1982M05” 1.025 1.06  
”1982M06” 0.902 0.93  
”1982M07” 1.166 1.006  
”1982M08” 1.046 0.985  
”1982M09” 1.144 0.947  
”1982M10” 1.173 1.059  
”1982M11” 1.372 1.079  
”1982M12” 1.303 1.047  
”1983M01” 1.586 1.187  
”1983M02” 1.699 1.135  
”1983M03” 1.606 1.168  
”1983M04” 1.472 1.197  
”1983M05” 1.776 1.3  
”1983M06” 1.733 1.344  
”1983M07” 1.785 1.41  
”1983M08” 1.91 1.711  
”1983M09” 1.71 1.493  
”1983M10” 1.715 1.586  
”1983M11” 1.785 1.462

”1983M12” 1.688 1.509  
”1984M01” 1.897 1.595  
”1984M02” 2.26 1.562  
”1984M03” 1.663 1.6  
”1984M04” 1.851 1.683  
”1984M05” 1.774 1.732  
”1984M06” 1.843 1.714  
”1984M07” 1.732 1.692  
”1984M08” 1.586 1.685  
”1984M09” 1.698 1.642  
”1984M10” 1.59 1.633  
”1984M11” 1.689 1.611  
”1984M12” 1.612 1.629  
”1985M01” 1.711 1.646  
”1985M02” 1.632 1.772  
”1985M03” 1.8 1.715  
”1985M04” 1.821 1.63  
”1985M05” 1.68 1.665  
”1985M06” 1.676 1.791  
”1985M07” 1.684 1.693  
”1985M08” 1.743 1.685  
”1985M09” 1.676 1.806  
”1985M10” 1.834 1.565  
”1985M11” 1.698 1.749  
”1985M12” 1.942 1.732  
”1986M01” 1.972 1.723  
”1986M02” 1.848 1.753  
”1986M03” 1.876 1.756  
”1986M04” 1.933 1.685

”1986M05” 1.854 1.833  
”1986M06” 1.847 1.672  
”1986M07” 1.782 1.722  
”1986M08” 1.807 1.763  
”1986M09” 1.687 1.732  
”1986M10” 1.681 1.782  
”1986M11” 1.623 1.793  
”1986M12” 1.833 1.84  
”1987M01” 1.774 1.862  
”1987M02” 1.784 1.771  
”1987M03” 1.726 1.694  
”1987M04” 1.614 1.735  
”1987M05” 1.628 1.713  
”1987M06” 1.594 1.635  
”1987M07” 1.575 1.685  
”1987M08” 1.605 1.624  
”1987M09” 1.695 1.587  
”1987M10” 1.515 1.577  
”1987M11” 1.656 1.578  
”1987M12” 1.4 1.632  
”1988M01” 1.271 1.554  
”1988M02” 1.473 1.45  
”1988M03” 1.532 1.6  
”1988M04” 1.573 1.615  
”1988M05” 1.421 1.483  
”1988M06” 1.478 1.512  
”1988M07” 1.467 1.527  
”1988M08” 1.493 1.551  
”1988M09” 1.492 1.531

”1988M10” 1.522 1.529  
”1988M11” 1.569 1.407  
”1988M12” 1.563 1.547  
”1989M01” 1.621 1.561  
”1989M02” 1.425 1.597  
”1989M03” 1.422 1.442  
”1989M04” 1.339 1.542  
”1989M05” 1.331 1.449  
”1989M06” 1.397 1.346  
”1989M07” 1.427 1.386  
”1989M08” 1.332 1.429  
”1989M09” 1.279 1.338  
”1989M10” 1.41 1.333  
”1989M11” 1.351 1.475  
”1989M12” 1.251 1.304  
”1990M01” 1.551 1.508  
”1990M02” 1.437 1.352  
”1990M03” 1.289 1.345  
”1990M04” 1.248 1.332  
”1990M05” 1.212 1.351  
”1990M06” 1.177 1.263  
”1990M07” 1.171 1.295  
”1990M08” 1.115 1.307  
”1990M09” 1.11 1.312  
”1990M10” 1.014 1.282  
”1990M11” 1.145 1.248  
”1990M12” 0.969 1.173  
”1991M01” 0.798 1.149  
”1991M02” 0.965 1.09

”1991M03” 0.921 1.176  
”1991M04” 1.001 1.093  
”1991M05” 0.996 1.07  
”1991M06” 1.036 1.093  
”1991M07” 1.063 1.076  
”1991M08” 1.049 1.05  
”1991M09” 1.015 1.216  
”1991M10” 1.079 1.076  
”1991M11” 1.103 1.013  
”1991M12” 1.079 1.002  
”1992M01” 1.176 1.061  
”1992M02” 1.25 1.098  
”1992M03” 1.297 1.128  
”1992M04” 1.099 1.083  
”1992M05” 1.214 1.187  
”1992M06” 1.145 1.189  
”1992M07” 1.139 1.251  
”1992M08” 1.226 1.14  
”1992M09” 1.186 1.123  
”1992M10” 1.244 1.139  
”1992M11” 1.214 1.224  
”1992M12” 1.227 1.199  
”1993M01” 1.21 1.135  
”1993M02” 1.21 1.236  
”1993M03” 1.083 1.105  
”1993M04” 1.258 1.216  
”1993M05” 1.26 1.111  
”1993M06” 1.28 1.193  
”1993M07” 1.254 1.09

”1993M08” 1.3 1.264  
”1993M09” 1.343 1.172  
”1993M10” 1.392 1.246  
”1993M11” 1.376 1.235  
”1993M12” 1.533 1.289  
”1994M01” 1.277 1.21  
”1994M02” 1.333 1.354  
”1994M03” 1.531 1.261  
”1994M04” 1.491 1.369  
”1994M05” 1.507 1.423  
”1994M06” 1.401 1.337  
”1994M07” 1.431 1.278  
”1994M08” 1.454 1.353  
”1994M09” 1.483 1.419  
”1994M10” 1.437 1.363  
”1994M11” 1.504 1.354  
”1994M12” 1.505 1.4  
”1995M01” 1.37 1.415  
”1995M02” 1.322 1.302  
”1995M03” 1.241 1.442  
”1995M04” 1.278 1.331  
”1995M05” 1.3 1.324  
”1995M06” 1.301 1.256  
”1995M07” 1.45 1.332  
”1995M08” 1.401 1.247  
”1995M09” 1.401 1.267  
”1995M10” 1.351 1.32  
”1995M11” 1.458 1.36  
”1995M12” 1.425 1.225

”1996M01” 1.453 1.403  
”1996M02” 1.514 1.328  
”1996M03” 1.439 1.391  
”1996M04” 1.511 1.35  
”1996M05” 1.478 1.392  
”1996M06” 1.474 1.398

### C.3 Shipping Volume

”VOL” ”VOLJ” ”VOLQ”

19.2717057789 17.459748181 18.7609251809  
19.5739427053 17.0051823285 18.971430836  
20.2496352454 20.0632864458 21.5160397659  
18.7581267693 19.0300416396 22.511024212  
18.9623879164 19.27406249 23.6257746082  
18.7082264913 17.7225435923 18.9527794473  
17.5583325839 16.3996071649 16.3155079075  
16.200570162 16.0688532171 16.4200268795  
17.5672224715 15.9365700733 16.7922075809  
18.3506389645 15.174656225 17.2723634089  
19.6108588322 17.399682921 18.8658616044  
19.0548224273 18.3899433918 17.5524349924  
17.8562732579 17.3099553279 17.59936768  
17.3026348251 15.7391009507 17.3483112881  
16.992232973 16.2263804308 16.1946474378  
16.6783874199 15.0494683232 15.8035069624  
17.440059836 14.8752473335 16.715966412  
16.6618026428 17.0961955995 17.5161819485  
16.384313619 16.7257725533 17.0938652092

16.050331812 13.739513488 15.0166120316  
15.7184746166 13.4520789836 14.2574702548  
15.849494067 15.2452098373 16.1171312944  
15.2144285697 13.7662941367 14.7161769243  
15.5820599759 13.2800857116 13.5803587289  
16.504876926 14.5873238183 15.4941943822  
16.3726266283 14.752659297 15.6499708269  
17.441672857 16.3451653899 17.4300040172  
18.1500104973 17.4648816444 18.1790757434  
18.874365366 18.1946450784 17.9539480087  
18.1098021573 16.5289387474 17.0578268847  
18.6816660898 16.9546621479 17.4720840136  
18.246280095 16.8771274629 17.5725869706  
18.0012782954 17.1811808451 17.7155012812  
19.587794813 17.494746818 20.0703911986  
19.5981770221 17.217759479 21.6789128429  
19.2223298359 17.8218538169 19.8239398072  
20.0634140058 18.5751902922 18.3789688332  
20.0809239368 20.6290468968 20.7880586038  
19.1786299632 19.0280437604 21.0085727009  
19.1588286054 17.0601921473 19.933106815  
19.3928968784 18.122574268 19.2421396264  
18.9646978349 17.3292945255 18.511792914  
20.435792902 19.6739965193 20.1347179524  
20.4202833337 20.4439466979 20.4893669879  
20.9052188136 19.7398566084 20.1878793077  
19.6652673577 18.6546627952 18.249475265  
20.0951191985 19.4590133306 20.6924024332  
20.7800095041 19.0475471902 20.5414206421

21.2965069366 18.9291269898 20.0406867042  
20.8192028548 20.769493656 20.7077992512  
21.0614028758 20.9171131275 19.9587141065  
20.9765403357 18.7027400646 19.2220417591  
19.1698867079 17.6112329857 17.6748160012  
18.9439652669 17.8949795526 18.0725569547  
19.8093280201 19.4476632819 19.1307345563  
19.5802661175 20.7283882343 19.7635124256  
20.188836761 22.031446338 21.3858459329  
18.1792990372 19.8472107672 19.1444759791  
17.7548507547 18.72925784 18.6182439677  
17.2289318147 16.8165314785 17.7292985173  
16.4121068243 15.5654085445 15.9577922676  
16.2045884936 16.3836619494 15.7260238369  
15.0130253194 15.3710289573 16.7921314762  
13.8266097099 13.052907295 14.9949652982  
13.1843688204 11.6840841891 12.0034276402  
12.870213406 11.6770775998 12.9172510431  
13.2334442709 11.5479612775 14.9634674363  
12.9173329865 11.5577104018 13.6367828064  
12.7117169428 11.3733253536 11.787848607  
14.1808084522 13.3665634393 12.8296343543  
12.9484589444 12.878562356 12.4906568078  
13.6214661551 14.483489407 16.3731037573  
14.7098312316 15.4407091256 17.3526214658  
14.4560809397 13.8277648613 14.4358762589  
14.8227523736 13.6335269982 12.5340732824  
16.5007559885 15.7312209525 15.0489415034  
17.5106649474 17.840872251 17.7325640302

17.6557029729 18.2473296622 16.7282353463  
17.8485823627 17.6195057968 16.7266919875  
19.67633947 18.3353228515 19.0679266201  
20.277734492 19.3934423648 20.0492502848  
18.8260717628 18.7528520307 19.2689093322  
17.9179529725 17.8954628599 19.0376518667  
19.2171790962 19.1967894497 20.7927473245  
18.914769394 17.6541858873 20.4452750898  
20.2323399576 17.7928865573 20.5384209949  
19.5391206912 16.8753676725 17.9956280129  
19.8538946266 18.3103862958 20.1637225044  
20.7581219007 20.6901786614 22.1197827216  
20.9316987843 19.3631664403 19.8699426046  
20.7462892596 19.3640030428 20.5257133466  
19.4225143403 19.1001722665 19.4388955707  
18.2520026658 16.9920619381 17.0712611214  
19.608942608 17.1943404778 17.2762205259  
20.4870375324 19.0729051414 20.8129575344  
20.9231428276 20.2297824525 22.2245091066  
21.02105968 21.1690728089 21.9992265702  
22.7732010085 23.0189060197 23.6707827542  
22.514446987 21.0231950769 23.1500426815  
22.3465504392 21.1552536397 24.20522105  
22.6577539724 21.6521714803 22.9788447684  
21.5418439884 19.8456301907 20.3661042955  
22.3394036118 20.2014282014 21.979090813  
23.0377384332 22.5350106791 25.557668128  
24.4548555232 25.4417185746 26.3802290462  
25.4492262625 26.4301907851 26.7259049531

26.0800222942 26.1117321025 26.6803797332  
25.7911761748 23.9400932834 24.4245899475  
25.5847992209 21.93907776 24.6919257021  
26.0231773653 23.9091204958 26.7800488092  
24.4973264721 23.1144196118 24.9828317369  
24.3927828027 23.4339056057 25.5422039392  
22.8601533285 23.9932847305 25.2274388821  
21.9722786038 20.9681427918 21.3364275906  
22.340510202 20.9392190077 21.3514886656  
23.178110338 22.8665030015 24.3604684646  
24.6484941185 22.7150279438 24.0351232014  
24.7817200659 22.5761641267 23.9361608628  
23.6865622916 23.1766610537 24.9860708056  
25.2079488338 24.6010715071 24.7884467738  
27.3211087537 27.0249055141 26.7444706647  
27.7235428258 26.5757318032 25.0603692831  
28.309548854 25.6624963062 25.5292199701  
28.0578284722 26.4611757937 29.0803741879  
26.1581604322 25.2816128422 27.3139243645  
24.190133874 23.4134398476 24.1925016467  
24.4548767286 23.2861317007 24.5498712482  
25.7278951519 25.8504378763 26.7140269071  
25.5131783927 25.5063764553 25.7889361141  
24.9237228352 23.6404561965 25.8633352475  
25.409962031 24.5375853257 26.7049047215  
25.1984221295 23.8659325277 25.3081948561  
26.6551560148 25.7888997326 27.2553131992  
26.0002259958 27.5926843712 27.6890420101  
24.0036455729 26.4846757191 25.73693066

24.7777899876 24.9877932221 25.2164930005  
23.4921955228 22.7141645557 23.8826562386  
22.2793511249 21.9125168987 23.8191504286  
21.0976561552 21.1247412228 23.894114282  
20.8288163817 19.44188337 23.1970348431  
20.8638228094 18.3573265486 21.5891808012  
20.8980558018 19.8206384805 20.4934938956  
21.0619706551 20.6085901372 21.8742971165  
18.6701274063 18.1244279374 20.3411534843  
20.6007088077 18.6105112695 19.9682448467  
20.1389619959 17.0835208855 18.7665465342  
21.0123229261 19.6156353981 19.3833124362  
20.667671816 19.756398576 19.9818255248  
19.3152257144 17.9826814757 19.2152138844  
18.679959164 19.0937240221 20.1580164641  
19.445442753 19.0690271957 20.3613634188  
18.3950216375 16.6347947986 17.2943772219  
18.0754720839 16.7480007967 18.9798968194  
17.949099935 15.3911546702 18.3499016254  
17.2085569825 14.0417133154 16.8890516998  
17.0534788489 16.6259563941 17.0382068092  
17.1429907854 18.8552913007 16.4384938674  
18.5021723746 19.7240189504 19.1165368056  
17.1809981532 18.0932390067 19.6651639173  
16.8715575665 16.1294506738 17.746526401  
17.4542411461 15.2359072701 16.2825178305  
17.7473900782 16.1718781121 17.1518078679  
17.829701024 17.8086415346 17.7970579525  
17.3685099902 16.4949978314 17.2808975505

17.3949434238 16.0550286765 16.1101830324  
18.6989747141 18.9542659565 18.4792272786  
18.6846465065 18.7648160468 20.8419383248  
18.9306198844 17.7578610147 20.2409134333  
17.5729229166 16.5438598725 17.9606830042  
18.0739332244 17.8880302871 18.6181490647  
17.2507295452 17.3171442986 17.5916131555  
17.3792585334 17.7651379488 17.6504299241  
16.4320763485 16.2982752451 17.8254445424  
16.7933478109 15.6933362068 17.9872436465  
16.5190627996 15.3485530218 16.3673260868  
17.7564694411 15.4315761325 15.9449256264  
18.8370114707 16.9575325807 17.611194868  
16.928292708 15.2967887037 17.1400231536  
18.657251668 18.4232770204 19.824527657  
18.2216460571 19.1539754996 21.3437806632  
18.1765329722 18.7223040949 20.3646186335  
18.1231211241 17.7984509251 17.2622289511  
18.0267230907 16.3030742753 16.1332113107  
17.2057201056 15.3882137759 15.5255516343  
19.3695759924 16.7182098318 18.3802550567  
17.1969637212 16.2442837156 17.2353515985  
17.7241137916 17.2317792517 17.1203437917  
17.2944190208 15.8843349068 16.6669091401  
17.470491388 15.3641123977 16.8551813982  
16.1498378983 13.1472904612 15.6012614531  
16.7740602842 14.4230964821 15.5086448779  
18.5417153526 17.2153398311 18.5540378889  
20.0944359175 19.0291421836 19.8158585998

19.5618854131 18.3320013774 17.7131366713  
19.2589074566 18.4277187063 20.1977164553  
18.4509674474 16.6629259136 17.8855835828  
20.4582337378 18.3891106399 19.6143190348  
18.8991549148 18.1939683622 19.9814583911  
16.8327435075 16.990464713 18.7612193209  
16.3872484992 16.1880929528 17.8850416195  
17.0434071889 17.4056840078 18.7971980838  
16.1873353526 14.4414953077 16.8184456409  
15.6693799332 11.833032659 13.0781945376  
16.9720675111 15.1806169456 16.0668334749  
19.1266563136 17.2570035319 19.4533781776  
19.3585341754 16.9028050758 18.5335424625  
21.5516285496 19.4802597811 20.1442177594  
21.7587725871 20.6168240706 20.9903258975  
21.8200171304 20.6700043623 21.6358988412  
19.9900423085 18.6721268744 19.6380808657  
18.1223443626 17.2656949423 18.4544019026  
17.0073882459 14.8938017204 17.2494692953  
17.6764622101 16.7706966062 17.4938713105  
16.5471831125 17.3608550016 17.5129378141  
15.1336065036 14.6423622788 18.8696919892  
13.5933196417 11.9614203809 15.0914451671  
11.8922993507 10.8989818839 12.3719969593  
12.9792532877 11.3093549053 14.874765085  
13.2525228835 10.0072486302 13.4098884284  
14.2012719789 12.7478370304 14.754422718  
13.6390791171 14.4037521116 14.0637227305  
13.4863734289 12.5546065488 11.944629052

13.3058904489 9.8681948227 11.1977633977  
13.8880929566 12.3213850067 13.8194375469  
14.5353434757 13.6643526293 14.9182149824  
14.2270370612 11.9699489905 13.6694506986  
13.9279040761 13.8643051684 14.5464302004  
15.5458094922 15.8959039789 15.4938533744  
17.1028315286 17.7287858103 18.6680621028  
16.3770032224 16.4426582273 17.8485416373  
15.9922044218 14.0712909853 14.7548208802  
17.1087386592 15.0603369214 15.9055684402  
17.6512137478 15.6631332073 17.0522838584  
17.6538328172 18.1289527652 18.2660043146  
16.6832498412 17.8887371637 18.6728109572  
15.305531634 14.9279931726 17.5729848736  
14.3195009831 12.7877164232 13.0553330776  
14.4053019566 12.107440994 13.1649766486  
14.4842521561 13.5984482771 15.9549228982  
14.0636282626 12.0090024876 14.3949805786  
12.8584560571 10.9501498295 13.9343271549  
13.0860805015 14.4201934438 16.3693368261  
12.0045389739 13.5897899096 15.1533367919  
10.1860556759 11.2654176894 11.2305040843  
11.6172375556 11.7897245077 11.8953155341  
13.6887086068 11.9049896003 13.6439356193  
13.4339843303 11.6735261414 14.1251999412  
14.297034008 14.518163804 15.5499969183  
14.5631142042 13.6598936396 14.249444216  
14.462636831 12.31610567 12.8322807377  
14.9756193359 13.1153849658 14.8811074394

13.2281979392 11.8856206242 12.7763373734  
11.7223121602 11.6820077126 11.700749822  
10.988282058 10.6323597054 11.9456987648  
12.6780557077 12.0987220557 13.0086293827  
13.1879117244 13.4444662146 15.8594941185  
13.1708257104 12.6744878153 15.176984815  
15.1370107561 14.5434709767 15.0800799684  
16.5389693478 16.6532639951 15.3514687031  
17.5778333136 17.049090127 17.523751158  
17.101844839 15.821528298 17.2895009478  
17.1815774639 15.4422459511 16.353068953  
16.9303511141 16.2555602655 17.1338698692  
17.6041430019 17.1916134874 18.166043312  
17.3397375674 15.3263700856 17.4882543394  
17.397222218 16.3586599867 17.8351820647  
18.1118957978 17.6659505236 18.1721698844  
20.4871702842 19.5667184188 19.3851257499  
20.0732084475 18.5900566876 19.1596501799  
20.6831061094 19.2984072947 21.3427199232  
21.9311965536 20.1818451959 21.4651039503  
19.8731129437 17.8698724488 18.6347104915  
18.8870897878 18.4831658733 18.8844647261  
18.9373467112 18.7657221868 21.073085258  
19.1222161684 18.205919517 22.2839612374  
18.4419961166 16.8004779789 18.2486488717  
19.0919098255 17.012484616 17.6643303519  
20.4842902986 19.1957898359 19.9359566247  
20.1493131966 19.4869774878 20.1466759111  
19.1493887473 19.8103031926 20.2410268015

19.0803921474 19.6331294076 20.1894038473  
18.9774237315 18.4838204515 18.5255073377  
18.5951204971 17.2380193453 17.0952695772  
18.1202398412 15.4738698799 16.1724956515  
18.9493083777 17.1012271058 17.2513459355  
19.4330846221 17.9582462404 17.1747880028  
18.9388177502 17.186551866 15.925098325  
17.1994059115 15.3040242876 14.0525424022  
18.2425007177 16.9683657256 16.4553290302  
19.1864508026 16.9975200548 16.4254802264  
20.3883877206 17.1688724494 17.7150312537  
20.9237898566 19.8632664667 21.0055981934  
20.4857838628 20.2765320654 19.5923552627  
20.9344768673 20.2555920022 19.765716827  
22.6286362817 22.5406746977 24.2559186341  
23.3835467107 23.9766267744 26.0840689496  
22.9080141853 21.5620943949 24.1434845471  
21.9087452385 20.1485252991 22.8028759064  
23.3625975594 22.876000135 24.533062705  
23.7507675386 22.8000020619 24.7802677328  
22.9991666475 21.0444272266 23.349061915  
23.9101097677 21.7736734484 24.1372974453  
25.0481965385 23.2442106217 25.1928892128  
24.378337561 22.9443109328 23.9085559835  
24.0729702983 24.1641013804 22.734968973  
24.03330376 22.3932053631 22.2289221234  
25.8573819761 22.7792350251 24.8463087148  
25.247985532 24.5454737364 27.1846385734  
26.6352566668 26.2679469323 28.5772601514

26.5706393237 25.0674303642 27.1124521629  
27.1020003043 23.8338381778 25.9495915889  
25.407451359 22.1617576711 21.7258099887  
24.9217137001 22.3908187053 22.9860986468  
24.7336224961 21.8527485073 25.6282992548  
23.733044213 20.8828150041 24.1866788436  
23.7002788761 21.7153498813 24.4057981764  
22.0169551843 22.2532215733 23.6781326117  
22.3872561117 21.8207079904 22.665252291  
23.2468318591 22.0037164455 23.3720385988  
22.3031641246 20.59708956 22.8464936989  
22.2882789426 18.5334714503 21.4659071838  
22.0142923185 18.8125614082 20.8689316677  
21.147403764 20.1339433991 20.5697742398  
23.1582607822 22.6016035839 22.5155118781  
24.5494166858 23.516222821 24.3412964107  
25.4438237759 24.2016784812 25.43146408  
26.2419657113 23.4613715669 25.4476293145  
28.0251788813 25.1381195175 26.426280573  
27.1103832179 24.4821257753 26.0235121131  
26.0636161013 21.9629934657 25.5023633851  
24.045421333 20.345249999 23.3470138301  
24.6230735748 24.2465087457 26.0368943072  
24.8632535937 25.2916452395 25.6742120389  
25.1107771897 24.2685449126 23.6806571676  
24.1460380115 23.43034969 22.5463709003  
24.6554541615 24.1090924498 24.5693824503  
25.803343129 24.2028796484 26.5239434661  
25.3210480052 24.5483346202 27.1710669468

24.3512790657 24.8491588249 26.2951505306  
23.2398097757 22.7840152024 24.5075273728  
23.7241444609 21.8148420646 24.3200068831  
26.0353232687 24.2522043325 25.4361786164  
25.6097408758 26.4400100575 26.7910897862  
25.063915726 25.0954334514 25.6814399254  
24.8958915922 22.8567963975 25.5866891795  
24.3703393673 23.1188081545 25.53691195  
23.8271292384 23.33094276 22.8929259453  
24.4649332039 23.2794283617 22.3399911894  
23.687565252 22.8605971108 22.6686592849  
22.2380979023 20.4431944678 21.8716033803  
21.2517328293 18.6501250939 19.5984581859  
23.4427655694 21.453116372 21.744249143  
23.0893939784 21.8778477559 22.0716295269  
22.618359591 22.7316141988 22.2482273796  
23.3007725567 22.5400084265 24.117049645  
22.6457677274 21.4637610354 23.6131170032  
20.9280736606 19.6999045161 20.7981175777  
23.32502062 21.6019012651 23.2329287692  
22.9606839238 21.175852282 23.0163238851  
21.4593501176 20.5116668993 23.1472203339  
20.2915284518 18.7965466782 21.8750713427  
21.6026410977 19.4070771868 22.0692832518  
21.2713155413 20.8821145097 22.1446262998  
21.753485771 22.8736738225 23.4722854425  
22.0449829683 22.3135898403 23.0481856245  
21.9606947779 20.1276998265 20.8756023949  
20.6831695911 19.0538943225 19.5515318425

20.4705890117 19.5938982172 20.9078733403  
19.2668812869 18.480906785 21.0805823524  
18.9119477041 17.8215688645 20.5010875444  
18.874718744 18.1218202426 19.8423708377  
19.2447445598 19.0777652508 20.6073522484  
20.0046662833 20.2414263753 20.9381706722  
19.9169518794 20.7151340149 21.0468436088  
19.0840575418 18.9154928962 20.1140672061  
19.1831364912 16.7070405998 19.5959314411  
18.4047333053 16.4088827162 18.4363570721  
18.5150431465 18.8862179884 18.2222225423  
21.1052367891 21.3631058603 21.5641207421  
20.3224822562 20.3357962296 20.1752533424  
21.6262169581 20.8966817131 19.2322121201  
21.6628109831 20.3742045012 19.7437892188  
20.8894316342 19.9240120458 20.0617332717  
19.7357968942 19.603144811 21.2685459271  
18.6649522425 18.3498091278 19.8446702664  
18.9272714656 18.3335516281 17.8497048108  
18.6408302137 17.3005167375 17.817158861  
19.0348509046 15.5603227378 16.9413415764  
18.8620380441 16.6206855703 18.0618215214  
17.7582018138 17.6833355663 19.7603378309  
17.5865912068 17.5007424849 19.9151530734  
17.4569461928 16.8756312559 18.8736735748  
18.6928705763 17.5490627771 17.406025111  
20.8554312267 20.518518501 20.2285407441  
21.0431871878 20.7553777845 21.4478749251  
21.7605112375 21.5179825665 22.6879389307

22.2635859944 22.2947487461 23.0692442556  
23.4593270587 23.8728061029 22.2072964945  
24.3643538925 24.1161240546 23.7067318271  
22.8511668654 23.4799577383 23.7518257326  
22.2158881701 24.1399220831 22.4762667724  
21.506969127 22.4898191792 21.0257920594  
21.3643153637 22.0539723757 21.6471236819  
20.6308301889 19.8958877919 20.9097801487  
22.3502803992 20.6295465845 21.4109599994  
22.4509312433 20.5845742334 21.8772735351  
22.9625097649 21.3040703952 24.9445821146  
22.2889775934 22.1084431949 25.1297478098  
21.8275142561 20.7767966791 21.7927009695  
22.767427257 20.2312336356 20.8178045959  
23.2639014579 21.8411265196 21.6162171178  
23.5056807811 22.3836713841 22.2470168841  
22.2531214982 21.0609077247 23.2046391724  
22.5109651384 21.2180327032 22.4055994067  
21.1315412798 20.2621906819 20.3601866072  
22.7979670363 20.7973540572 22.6120355119  
22.4597998002 19.4819583415 20.2589984649  
22.5685806919 21.4656028243 20.8244965736  
23.1117992222 22.7704556329 23.5759658262  
23.1148278442 21.6541335313 23.814912354  
22.7080404892 20.5543261522 21.3280078065  
20.6138812777 20.8340311797 20.7436982306  
19.8175184359 20.627340169 22.0233781607  
21.8397837472 21.5781456957 23.2415075421  
22.4863739235 21.7160170737 22.3953071808

22.5862446044 19.9504899514 21.1590792018  
23.5966887071 21.123650223 22.3135410898  
22.389376973 21.646393941 22.2621277802  
22.7772227728 21.6994261979 22.3895548005  
22.4970718077 20.961671766 21.25748542  
23.5577574081 22.5463723266 24.1901600257  
23.9486486406 22.9994169231 25.0256848536  
23.8292055026 23.0176469965 23.780490369  
24.0573682683 24.1477877229 24.0234715129  
24.5508444004 23.5898086646 23.7772076304  
23.8058600714 23.1746657293 23.894093882  
21.8489699241 22.2416986601 22.5296947225  
21.73433214 22.2611011035 21.8584446115  
20.5412739422 21.6018474042 20.9543264569  
20.1933247515 19.3293856417 20.4071418934  
19.9090955897 18.5109230835 20.0829073771  
21.6728167003 22.1893561572 22.8038479162  
21.4656896916 20.6219149552 21.500495681  
21.1880209471 18.535793256 20.2834927491  
21.4120801886 19.6960497939 21.0566737511  
21.0518283049 20.2304960393 21.3641478721  
21.4639082222 20.3030557583 22.9946190166  
20.5700062517 18.8692110634 20.8827151928  
19.8200976155 18.1861361645 19.7724005212  
19.1911291962 17.2131394148 18.6597562381  
18.9614604716 17.6232807408 17.5566910945  
18.7299738725 17.5731644718 17.8560461459  
18.5187869958 17.2163096472 17.2747463023  
18.5545182341 17.1931258096 17.0383757727

19.1291528441 18.2630897704 19.3162915175  
19.4902985602 18.1336875229 20.4157543726  
20.6587642681 19.2375993093 21.8807963886  
20.4673930836 19.1057691056 20.8438942223  
20.0897427646 18.3323598047 19.3826887806  
20.5063698251 19.7758929097 20.8951597138  
20.505497928 19.9000689072 20.4197723677  
18.8076524367 18.3810439884 19.6625495626  
19.8634849023 19.4156944955 22.0339293394  
21.1948630392 20.649772101 21.6611309692  
20.8044111905 18.6091384893 20.4370543985  
20.6277443624 17.7730402107 19.65093184  
19.0335146395 17.495896754 16.4858391269  
18.6942260448 16.7361848048 15.3694826033  
17.6886993803 15.8142204458 16.0435568897  
18.3988659 16.3183613043 17.6679485898  
18.4186108459 17.449570972 20.5543209185  
18.1913308174 18.3289437767 21.4661893014  
19.9161736068 19.7393175151 21.28545009  
19.4905202791 19.2702556335 20.7353554538  
19.5372256525 18.2167986815 19.5896205263  
17.8331262582 15.9013072733 17.451699795  
16.8628809052 15.3021180272 16.1689763559  
19.1377174263 17.7481579931 18.752046662  
20.5528472205 18.350169557 20.8465209868  
20.044417392 18.9260857695 19.8991045339  
21.90705532 21.9526527667 21.0379488008  
18.9636093739 16.7995112978 16.7299529102  
18.2798826723 15.5327962381 15.8078716203

17.8625515725 16.5998163031 16.4388265869  
 15.9985922602 14.3669864476 14.4182778797  
 15.4755380256 13.8280178767 13.8770849082  
 14.9382346447 13.9390460484 12.7130002608  
 17.9565443851 17.6240705632 16.4280246338  
 16.3803314531 15.4006963992 16.3058512474  
 16.8321951211 14.8465643447 17.1307350074  
 18.071538169 17.0518079596 18.7561666438  
 17.98005764 17.0521784417 18.1471518528  
 19.1752983323 16.736676832 18.4692580172  
 18.3863809923 15.6003634767 17.0590983557  
 19.8586744085 18.5544195266 19.5944455766  
 18.585237438 17.9512109824 19.330102047  
 18.3261409431 16.5016242493 18.9295382482  
 18.8030741444 18.5584192154 19.5012890623  
 18.3620479436 18.3916749438 18.2236209754

## C.4 Hungarian Exchange Rate

1 1.6268391 1.6984192 2.0079907 1.4345695 2.7828483 2.8362358 4.3383264  
 4.5941219 5.3779608 4.0980233 3.4269932 4.5741974 3.9609699 4.4903911 4.1765334  
 4.0659293 3.0434249 2.0164477 2.8522073 2.8140498 2.1848722 1.5950817 2.2429898  
 2.2012101 2.5564244 2.8183936 3.2920329 3.5386639 2.7520406 2.9887184 3.6628315  
 4.1155835 2.6670804 2.4475717 2.205739 2.4292855 2.0911023 2.0898105 3.043442  
 3.6113511 3.7893799 3.2121155 3.1678467 3.2550351 2.9450505 2.7632934 2.9777748  
 3.7541152 2.3789054 1.5524019 2.4166115 2.8760458 2.6712716 2.9638433 2.3101149  
 1.6210284 1.8385815 2.8168296 3.3515586 2.9978249 3.5861905 3.4218998 2.9695071  
 2.6977919 2.340162 2.2215253 2.5238235 1.9671895 2.1577204 2.7455625 2.8270665  
 3.1897584 3.1630046 4.1443688 4.6993679 3.6025463 3.6273713 2.4304996 3.2260433

3.5346954 4.0054737 4.6256033 5.8589386 5.5990677 5.4946565 5.9304322 6.596674  
5.8305304 5.4417317 5.4687066 3.8988953 4.8830323 3.9859455 5.0013413 4.2901215  
4.8488491 5.5400411 5.394801 5.8261948 5.732879 6.111303 5.3929717 4.9007317  
5.8244318 5.382873 5.5454446 4.5243989 4.2348796 3.7097975 3.5342468 4.1482148  
4.7702349 5.522976 5.6296711 5.9432146 5.308443 5.0303299 5.7792977 6.3424265  
6.6176091 5.8713597 6.0768544 6.3105203 6.0791903 6.2389322 7.5763895 7.2482205  
6.1525888 4.112468 3.3052322 2.612247 1.4597108 3.5152237 4.7022798 5.5172526  
6.0048037 2.8930202 5.5298636 5.5789776 4.1278874 0.89193745 1.0621893  
4.7105699 5.4896383 7.1584 6.805053 8.6144752 5.9383055 7.7796817 8.7711985  
8.445656 8.7674898 11.132449 11.185289 12.520995 10.611369 12.42819 15.286988  
14.225634 12.496366 10.861144 12.023192 10.807324 10.657917 8.6713615 10.223299  
9.0802962 10.345198 11.421047 11.195249 11.571653 11.198371 10.802763 11.950971  
11.993388 10.957325 12.460033 11.349358 11.800016 10.95823 10.65431 11.015266  
11.907817 11.614755 11.885188 11.718403 11.730121 10.947176 10.856941 11.810782  
11.220396 10.313982 11.477275 12.436179 13.103131 11.894569 13.290609 12.698543  
11.558128 10.872649 11.20708 11.778828 11.960049 11.75378 13.07026 12.523631  
12.61295 12.068711 12.377789 11.036417 11.58504 10.704319 10.620286 9.8174616  
8.8637119 7.3421925 6.415701 6.616977 7.050929 8.3362776 8.9276029 7.0421763  
5.918664 5.7636259 3.2131937 2.7884873 2.0434108 2.4381397 3.0539853 3.991256  
3.7851832 3.5634831 4.8391543 6.5874414 6.1625992 6.3229257 5.4022381 5.4390715  
6.1107061 5.6039065 5.7098516 5.7363062 5.2972892 4.9316486 6.1513195 5.3786194  
4.9928725 3.8859135 3.8087715 4.1064588 3.7335037 3.9662801 3.5048923 4.2965473  
4.2758837 4.7575813 6.0889414 6.4267421 7.3985392 7.5934401 7.368304 6.739546  
6.477317 7.2241545 7.8019595 7.1136077 6.9831564 6.1580276 6.3652111 6.9822191  
6.5015883 4.7377317 5.3674897 8.1587713 8.8813851 9.2597047 9.5927926 10.634377  
12.883255 14.895499 14.614359 14.637208 14.936354 13.978189 15.247189 15.428835  
17.039209 16.818807 18.553565 16.317014 18.843618 16.408123 14.57398 14.661209  
12.336184 13.137634 14.011296 15.997919 15.359496 15.972913 16.310244 17.195519  
17.956373 16.820556 18.835553 18.336075 17.228906 16.038311 17.461696 17.266336

20.315677 20.003691 19.799512 20.344618 18.889694 18.538185 16.689197 18.044006  
18.436305 16.9659 17.052194 16.600613 18.075054 19.619692 20.40992 20.9723  
21.832826 21.031899 21.40006 21.025204 21.787375 21.633286 23.142391 23.305911  
24.422373 24.389852 23.113381 23.453081 23.280491 22.159267 22.357717 22.575927  
22.300664 23.27424 22.668373 21.971108 21.969299 22.243379 21.910315 21.742244  
21.102066 22.330443 23.778787 22.421849 24.318354 23.276223 22.928615 23.516817  
23.192031 24.439776 25.646361 24.013173 22.800817 21.930868 22.396108 21.948636  
22.63152 22.843553 21.965477 22.086954 21.484588 20.992311 22.040551 20.708479  
18.874245 19.415677 17.806978 17.646874 17.903768 17.644062 17.129987 16.969529  
17.931248 17.402765 17.883352 17.408811 17.826599 17.320226 18.132334 18.264919  
19.324093 19.563313 20.169274 21.064913 21.383854 21.115485 21.041537 20.447841  
19.167268 19.282245 19.669213 19.842788 19.645848 19.393213 20.050257 20.285704  
21.206103 21.741322 22.684601 21.714409 21.517625 22.221479 23.531033 22.094984  
22.177942 23.901701 24.071843 23.46034 22.124996 22.058321 22.403095 22.590732  
22.446712 22.465884 22.912077 23.531783 22.486634 22.982667 22.542868 22.934954  
22.528691 23.318882 23.990015 23.4417 23.044388 24.138724 24.843834 23.414284  
23.867763 24.197097 23.485826 22.491991 22.776258 21.857711 22.390119 22.338  
21.097323 21.759801 21.024496 20.614141 22.051117 21.823722 21.678008 22.437265  
23.441052 23.00545 22.465148 21.816945 22.210985 22.049267 22.299764 23.068759  
23.795316 23.215257 22.605824 23.075941 23.392012 23.89957 24.102533 23.171489  
22.563685 22.157582 21.943941 21.992486 21.965444 21.635254 21.761487 21.824159  
21.717114 21.537092 20.917835 21.449659 22.103253 22.371956 21.862982 22.107934  
21.452881 21.595871 21.819217 22.351561 22.990218 23.426957 22.771936 23.356011  
22.746644 22.104567 23.903777 23.660177 23.246803 23.582306 23.02254 23.442928  
23.342003 23.239028 22.862249 22.922168 22.296407 22.855469 22.513571 22.223672  
21.680807 21.434534 22.321791 22.649076 21.389639 20.758869 20.44828 19.735225  
19.489814 19.043549 19.68086 19.304515 18.50201 18.689426 17.842429 17.905011  
18.781568 18.886111 19.397855 19.525678 19.275408 19.967104 18.91759 20.965193  
21.155894 21.787756 21.397364 20.419689 18.779237 20.229039 18.601482 17.634975

17.406744 16.201074 16.397084 16.427208 15.768109 16.176317 15.270209 14.685959  
 15.355334 15.465751 15.768159 16.678746 17.015378 16.930714 17.128888 17.529663  
 16.454602 16.312587 15.496236 15.89618 16.518156 17.274122 16.648604 16.293775  
 17.070625 17.208007 16.974044 16.865031 16.961967 17.3944 17.143105 16.718039  
 16.753252 16.844759 16.818118 17.125553 18.288319 19.027563 18.176753 17.891172  
 16.673228 15.974902 16.879096 18.164183 20.231471 19.295101 20.248109 19.621391  
 21.030168 22.29068 21.865809 19.379118 19.267991 17.739786 16.340927 15.767129  
 13.917159 16.121872 15.741244 14.635771 14.533094 12.590452 14.162835 13.254253  
 13.759789 13.531156 11.736316 10.540666 10.701331 11.714493 12.102604 12.890286  
 11.972669 13.447575 15.760853 17.270281 15.322816

## C.5 Eurostar

99.1999969482 86.9000015259 108.5 119 121.099998474 117.800003052 111.199996948  
 102.800003052 93.0999984741 94.1999969482 81.4000015259 57.4000015259  
 52.5 59.0999984741 73.8000030518 99.6999969482 97.6999969482 103.400001526  
 103.5 94.6999969482 86.5 101.800003052 75.5999984741 65.5999984741

## C.6 BankWire Transfers

11.5538495 13.6827883 12.483232 10.8330683 10.8457835 11.6694254 11.546721  
 11.7410884 10.8265671 10.2322593 10.074095 11.1264895 11.2652772 10.2842486  
 9.1769437 9.3005372 8.9790619 10.510669 12.1111369 12.8633695 12.9791453  
 13.3202588 14.9058295 13.3445574 13.60132 13.9392483 13.8055779 14.7512005  
 15.7884112 14.425972 12.1438859 12.1084447 11.6292785 9.7112687 8.8009283  
 8.5336967 7.4968967 6.1815601 6.3582354 5.0254212 5.8837991 7.6623125 8.086742  
 8.2718261 7.6887475 6.556665 6.8305189 8.3272832 10.3902244 11.1315264  
 11.8735433 14.2927949 17.5727407 18.2033083 20.0942024 20.7989315 19.7259136  
 19.7014543 19.6978237 20.601377 21.5619129 20.0131328 15.9583137 14.9364171  
 14.200887 14.6443906 16.698498 16.4256365 16.8727126 17.4415194 16.811762

18.001792 18.1220941 18.9354647 20.9364672 19.3426313 21.0305699 23.5599389  
 24.8600723 24.1249326 23.0274481 21.1254041 18.8655556 17.1435016 17.1779413  
 19.206218 21.265189 21.1346605 18.4047332 17.3827121 16.646665 14.4172067  
 14.0046669 14.8596628 16.013894 16.1252398 15.5183761 15.0779161 14.7967942  
 14.9373701 16.7335848 17.6085812 18.1885837 20.4345491 20.8636272 21.1309462  
 21.4834243 20.4571287 17.2258595 13.4686958 12.0095145 11.3749635 10.842555  
 9.4307203 9.2285011 9.8386113 9.4809494 10.1548046 9.1328098 7.5477886  
 7.1403309 6.1090428 7.1376253 9.6419962 10.893147 9.8120998 10.0281064  
 8.0494831 6.5837567 9.9396207 12.2330996 11.4411421 11.930899 12.2443499  
 13.2757754 13.5147769 14.7485865 16.4410226 17.0872817 16.7905909 16.8072786  
 16.6540136 16.3968396 16.9873186 16.4413381 15.1697176 14.7529914 14.0347321  
 12.2286886 12.4921379 12.7870233 11.1981392 10.210248 10.1085626 7.9469598  
 5.7749489 3.741743 4.505234 5.4420682 6.9616641 8.4792232 9.5632638 9.3949009  
 9.6204303 8.6989569 7.6235613 6.9667928 7.2350961 5.9276945 4.096203 4.145325  
 6.2562708 7.7946143 7.8953547 7.9806367 8.3850321 10.712545 11.8330249  
 12.8937596 13.3888684 12.6286618 12.6561732 13.979903 12.8926098 11.0750817  
 10.5342009 11.1250268 10.6263291 9.7332819 10.2946824 9.5062875 10.1611082  
 11.7902245 11.7399657 12.2417721 14.0038044 15.0152511 16.0313511 16.53824  
 16.5422309

## C.7 Nile.com Hits

10527 11510 14982 11609 13962 14829 11811 15315 13702 14136 12513 13447  
 15791 11032 11552 15616 10698 13013 11990 18108 13341 19639 14734 10308  
 20065 15601 12745 14778 14227 16321 11750 12596 11046 8203 21149 15019  
 13109 15456 17693 16824 13117 11156 15489 18109 17760 20384 11889 12650  
 18174 13942 16485 16015 15010 11684 16182 9811 18900 16397 20547 21057  
 14467 9365 19399 19388 14776 12164 10494 16762 12231 17009 16362 23383  
 17742 18326 16453 15082 13735 13893 11698 13851 15218 14424 17427 15253

15230 20236 14149 18682 18458 20022 15808 20427 19109 14244 17348 19860  
 17013 16165 11351 16602 17804 19386 14606 15158 20604 15041 21182 14643  
 21980 15930 13342 18783 18262 20398 16426 18919 16314 15636 11820 38742  
 55050 45774 22393 16737 21300 13452 15563 17914 22325 19595 20574 18968  
 23808 23364 21628 18773 16804 15599 18642 20220 22484 18273 14450 23979  
 18250 21754 18832 19441 18701 21359 18468 22955 21661 19033 18164 22093  
 19848 20138 18353 20090 16290 18583 25099 21317 20996 20529 19307 19044  
 20879 17008 23621 15661 23771 24859 17587 14257 13307 21755 26337 11135  
 11589 14550 23208 19635 19707 22167 21662 16799 16689 21876 17366 22486  
 24097 23285 21429 22065 18864 23088 16801 24548 14481 18147 21900 18779  
 15816 21044 23272 24930 19943 22989 16038 24357 22432 24922 22110 25009  
 26188 21825 22849 25099 19081 19485 24610 24377 24091 23478 23282 24906  
 19448 17377 23815 23144 24378 19399 17009 25104 24468 17035 22536 21211  
 23178 24648 27645 20447 19200 23276 23192 27933 23872 25774 25783 25449  
 27509 21806 23073 18541 18427 30563 20843 17985 19585 25337 24555 25131  
 22736 27476 22782 20129 24485 27028 23214

## C.8 Thompson Energy Investors

"EHAT1Q" "EHAT2Q" "EHAT3Q" "EHAT4Q"  
 1.23425460192 1.03743302041 0.660664778877 -1.48341192714  
 -0.0758299215719 -0.343467088302 -2.38911063376 -3.10166390297  
 -0.275070631486 -2.32741895311 -3.04601974189 -2.83841393733  
 -2.07931297888 -2.82223513021 -2.63656651073 -1.37640273576  
 -0.946753296415 -0.944934637792 0.149401642603 -1.45467026859  
 -0.0909897884365 0.919635882612 -0.759940644155 -0.338374762846  
 1.00170611302 -0.685915604252 -0.27160625854 0.61491325413  
 -1.58942634884 -1.08654754219 -0.120140866444 -0.721249088095  
 0.347070327601 1.17294213679 0.445075417539 -0.159337273813

0.85989446139 0.162715217663 -0.414018233742 0.457091124262  
-0.61288540509 -1.1135881845 -0.173901298451 -1.50805412218  
-0.560782784297 0.324713626883 -1.05831748127 -1.01120432348  
0.830523928332 -0.602090813443 -0.599700698829 -0.0746939061698  
-1.35120004132 -1.27537615639 -0.684134167958 -0.880011601021  
-0.0566317167294 0.415139043834 0.111501922933 1.02111702514  
0.466219259761 0.157574841254 1.06267350183 0.39468591984  
-0.262941820321 0.683379298596 0.0525732148032 -1.45164388882  
0.920545426238 0.266490392497 -1.25869661019 -2.08804434312  
-0.56381569496 -2.00760935172 -2.76354257552 -2.40213541057  
-1.49906345014 -2.30484847331 -1.98840622949 -2.54229656128  
-0.952735397019 -0.768838257964 -1.44228054689 -0.113721647002  
0.090502277911 -0.667179549378 0.585397655846 -0.82885790713  
-0.748810059004 0.511769231726 -0.895268675153 -2.08774224413  
1.18717484737 -0.286071803178 -1.53826379707 -1.18015865035  
-1.35687013051 -2.50409377599 -2.05131013086 -0.94479783328  
-1.28023507534 -0.947423999468 0.0508764144353 0.0596754486756  
0.207312036506 1.09241584611 0.999114726518 0.561258008437  
0.905426219083 0.830455341346 0.409131985947 0.405931835908  
0.0137863535701 -0.327480475166 -0.258471889456 -1.15819822745  
-0.339915378395 -0.269687822482 -1.16831468363 -2.43336593472  
0.0369062899403 -0.891775416078 -2.18393529205 -3.61473160519  
-0.925063851748 -2.21396052491 -3.64181352088 -2.71248624896  
-1.37957894471 -2.8892247995 -2.03367238611 -1.56799475676  
-1.64488338704 -0.91131142467 -0.555656731881 -2.67399157867  
0.572327134044 0.782543526113 -1.466972578 -0.734539525669  
0.266320545835 -1.93259118931 -1.15451441923 -1.55061038079  
-2.17280483235 -1.37118037011 -1.74603697555 -2.27724677795  
0.588628486395 0.0216555004906 -0.682837874525 -0.257681578979

-0.509270841191 -1.16171850434 -0.689618440287 1.76748713777  
-0.702370526984 -0.275299482213 2.14119118673 1.14470097459  
0.358218981497 2.71260702558 1.66010199223 1.20479133305  
2.38950357788 1.36867176853 0.941929466687 0.433708107423  
-0.786593262579 -1.00205898234 -1.31971480596 -1.89943543039  
-0.292573979582 -0.679779285169 -1.32223158651 -0.447422692441  
-0.415885783171 -1.08420706875 -0.232731271007 1.74700413174  
-0.709089787618 0.105613967763 2.05218202238 1.89508644959  
0.745193015609 2.62906433769 2.41541807527 0.800458391031  
1.95692119347 1.80916386078 0.253634138264 1.44335976922  
0.0440759755961 -1.33842549746 0.00736674871561 0.876655530317  
-1.37818078797 -0.028491405351 0.84431248373 1.78405563277  
1.21458890905 1.96553597018 2.79536768719 2.32362747072  
0.870010930139 1.80723490875 1.43235949223 1.6530858896  
1.02250951568 0.724559255237 1.01466998015 1.09365028566  
-0.197715573522 0.182804020252 0.343330583055 -0.283858088565  
0.361137907534 0.504182734388 -0.138773971447 -0.411945133711  
0.17844649749 -0.432578896977 -0.676948914338 -0.763144588712  
-0.59353261955 -0.822124645858 -0.894089016001 -1.64050003309  
-0.286774913982 -0.411218612622 -1.20496450823 -0.660625243623  
-0.152555704718 -0.971657839256 -0.450189189117 0.27836917217  
-0.834056883678 -0.326076996714 0.390314879511 2.27538029395  
0.426218857531 1.06886458444 2.88741304604 3.5111061776  
0.684427160811 2.54066130083 3.19834585457 4.55697089648  
1.92332724882 2.64152789994 4.05473675989 4.0698868119  
0.90674080812 2.49000758768 2.65854510092 4.42108164086  
1.67215287756 1.92086315159 3.75571326721 2.60168516355  
0.41262827928 2.39532782603 1.37465575152 0.0486309249447  
2.0231487218 1.0389606657 -0.254156602262 -1.66975231483

-0.785862587624 -1.90009586021 -3.15434323941 -1.12467617648  
-1.19126990571 -2.51500216162 -0.548008503469 0.532984902951  
-1.44051020683 0.421153023558 1.40714134617 0.175182960536  
1.72045272098 2.57907313514 1.23223247785 3.18047331836  
1.02727316232 -0.167447458873 1.91800137637 3.65065233056  
-1.09401896295 1.08225993617 2.8968370537 3.39162789891  
2.06903427542 3.78687975825 4.19442138918 3.58538673215  
1.92066902592 2.51115178905 2.06712488869 1.28631735606  
0.778762339531 0.504558322416 -0.12307374519 0.423315560572  
-0.197863407968 -0.756638392926 -0.148141934973 -0.215127909402  
-0.578171163132 0.0128304875567 -0.0699353110479 -1.6581235686  
0.534324617729 0.400437729232 -1.23386031381 -0.970434789562  
-0.081508050013 -1.6685618537 -1.36252335414 -1.36854847243



## Appendix D

### Some Pop and “Cross-Over” Books and Sites Worth Examining

Lewis (2003) [Michael Lewis, *Moneyball*]. Appearances may lie, but the numbers don’t, so pay attention to the numbers.

Silver (2012) [Nate Silver, *The Signal and the Noise*]. Entertaining general investigation of forecasting’s successes and failures in a variety of disciplines (including in baseball, speaking of *Moneyball*), with an eye toward extracting general principles for what makes a good forecaster.

Tetlock and Gardner (2015) [Philip E. Tetlock and Dan Gardner, *Superforecasting: The Art and Science of Prediction*]. More (*much* more) extraction of general principles for what makes a good forecaster – indeed a “Superforecaster” – based on Tetlock’s huge IARPA-sponsored “good judgment project.”

[www.ForecastingPrinciples.com](http://www.ForecastingPrinciples.com). Still more on what makes a good forecaster.

Tetlock (2006) [Philip Tetlock, *Expert Political Judgment: How Good Is It? How Can We Know?*]. It’s lousy. Forecasts and “hopecasts” are not the same.

Gladwell (2000) [Malcolm Gladwell, *The Tipping Point*]. Hard-to-predict nonlinear phenomena are everywhere.

**Taleb (2007)** [Nasim Taleb, *The Black Swan*]. Why, if you’re careless, you’ll find that events you assess as likely to happen only “once-a-century” wind up happening every five years.

**Taleb (2008)** [Nasim Taleb, *Fooled by Randomness*]. Why it’s so easy to confuse luck with skill, with good lessons for model selection (i.e., avoiding in-sample overfitting) and forecast evaluation.

**Surowiecki (2004)** [James Surowiecki, *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*.] Often the whole is much greater than the sum of the parts, providing a foundation for forecast combination.

**Koopmans (1957)** [Tjalling Koopmans, *Three Essays on the State of Economic Science*]. Why markets work. The classic statement of how decentralized markets work to aggregate information efficiently. Warning: This is *not* a pop book!

**Kindleberger and Aliber (2011)** [Charles Kindleberger and Robert Aliber, *Manias, Panics and Crashes*]. Why markets sometimes fail. In bubbles, for example, groupthink usurps control of the group.

**Shiller (2005)** [Robert Shiller, *Irrational Exuberance*]. A great account of a particular bubble, in the midst of its growing.

**Olson (1971)** [Mancur Olson, *The Logic of Collective Action: Public Goods and the Theory of Groups*]. More on why markets can sometimes fail, as people free-ride and don’t contribute to the group, which is therefore much smaller than it appears.

**Schelling (1980)** [Thomas Schelling, *The Strategy of Conflict*]. Why market outcomes are complicated, but interesting.

## Appendix E

# Construction of the Wage Datasets

We construct our datasets from randomly sampling the much-larger Current Population Survey (CPS) datasets.<sup>1</sup>

We extract the data from the March CPS for 1995, 2004 and 2012 respectively, using the National Bureau of Economic Research (NBER) front end (<http://www.nber.org/data/cps.html>) and NBER SAS, SPSS, and Stata data definition file statements ([http://www.nber.org/data/cps\\_progs.html](http://www.nber.org/data/cps_progs.html)). We use both personal and family records.

We summarize certain of our selection criteria in Table ???. As indicated, the variable names change slightly in 2004 and 2012 relative to 1995. We focus our discussion on 1995.

### CPS Personal Data Selection Criteria

---

<sup>1</sup>See <http://aspe.hhs.gov/hsp/06/catalog-ai-an-na/cps.htm> for a brief and clear introduction to the CPS datasets.

Variable	Name (95)	Name (04,12)	Selection Criteria
Age	PEAGE	A_AGE	18-65
Labor force status		A_LFSR	1 working (we exclude armed forces)
Class of worker		A_CLSWKR	1,2,3,4 (we exclude self-employed and pro bono)

There are many CPS observations for which earnings data are completely missing. We drop those observations, as well as those that are not in the universe for the eligible CPS earning items ( $A\_ERNEL=0$ ), leaving 14363 observations. From those, we draw a random unweighted subsample with ten percent selection probability. This weighting combined with the selection criteria described above results in 1348 observations.

As summarized in the Table ??, we keep seven CPS variables. From the CPS data, we create additional variables AGE (age), FEMALE (1 if female, 0 otherwise), NONWHITE (1 if nonwhite, 0 otherwise), UNION (1 if union member, 0 otherwise). We also create EDUC (years of schooling) based on CPS variable PEEDUCA (educational attainment), as described in Table ?? . Because the CPS does not ask about years of experience, we construct the variable EXPER (potential working experience) as AGE (age) minus EDUC (year of schooling) minus 6.

## Variable List

The variable WAGE equals PRERNHLY (earnings per hour) in dollars for those paid hourly. For those not paid hourly ( $PRERNHLY=0$ ), we use PRERNWA (gross earnings last week) divided by PEHRUSL1 (usual working hours per week). That sometimes produces missing values, which we treat as missing earnings and drop from the sample. The final dataset contains 1323 observations with AGE, FEMALE, NONWHITE, UNION, EDUC, EXPER and WAGE.

Variable	Description
PEAGE (A_AGE)	Age
A_LFSR	Labor force status
A_CLSWKR	Class of worker
PEEDUCA (A_HGA)	Educational attainment
PERACE (PRDTRACE)	RACE
PESEX (A_SEX)	SEX
PEERNLAB (A_UNMEM)	UNION
PRERNWA (A_GRSWK)	Usual earnings per week
PEHRUSL1 (A_USLHRS)	Usual hours worked weekly
PEHRACTT (A_HRS1)	Hours worked last week
PRERNHLY (A_HRSPAY)	Earnings per hour
AGE	Equals PEAGE
FEMALE	Equals 1 if PESEX=2, 0 otherwise
NONWHITE	Equals 0 if PERACE=1, 0 otherwise
UNION	Equals 1 if PEERNLAB=1, 0 otherwise
EDUC	Refers to the Table
EXPER	Equals AGE-EDUC-6
WAGE	Equals PRERNHLY or PRERNWA/ PEHRUSL1
NOTE: Variable names in parentheses are for 2004 and 2012.	

## Definition of EDUC

mn3—l—Definition of EDUC			
EDUC	PEEDUCA (A_HGA)	Description	
0	31	Less than first grade	
1	32	Frist, second, third or four grade	
5	33	Fifth or sixth grade	
7	34	Seventh or eighth grade	
9	35	Ninth grade	
10	36	Tenth grade	
11	37	Eleventh grade	
12	38	Twelfth grade no diploma	
12	39	High school graduate	
12	40	Some college but no degree	
14	41	Associate degree-occupational/vocational	
14	42	Associate degree-academic program	
16	43	Bachelor' degree (B.A., A.B., B.S.)	
18	44	Master' degree (M.A., M.S., M.Eng., M.Ed., M.S.W., M.B.A.)	
20	45	Professional school degree (M.D., D.D.S., D.V.M., L.L.B., J.D.)	
20	46	Doctorate degree (Ph.D., Ed.D.)	

# Bibliography

- Aldrich, E.M., F. Fernndez-Villaverde, A.R. Gallant, and J.F. Rubio-Ramrez (2011), “Tapping the Supercomputer Under Your Desk: Solving Dynamic Equilibrium Models with Graphics Processors,” *Journal of Economic Dynamics and Control*, 35, 386–393.
- Anderson, D.R., D.J. Sweeney, and T.A. Williams (2008), *Statistics for Business and Economics*, South-Western.
- Box, G.E.P. (1979), “Robustness in the Strategy of Scientific Model Building,” In R.L. Launer and G.N. Wilkinson (eds.), *Robustness in Statistics: Proceedings of a Workshop*, Academic Press.
- Diebold, F.X. and J.A. Lopez (1996), “Forecast Evaluation and Combination,” In G.S. Maddala and C.R. Rao (eds.) *Handbook of Statistics (Statistical Methods in Finance)*, North- Holland, 241-268.
- Giacomini, R. and H. White (2006), “Tests of Conditional Predictive Ability,” *Econometrica*, 74, 1545–1578.
- Gladwell, M. (2000), *The Tipping Point*, Little, Brown and Company.
- Hamilton, J.D. (1989), “A New Approach to the Economic Analysis of Non-stationary Time Series and the Business Cycle,” *Econometrica*, 57, 357–384.
- Kindleberger, C.P. and R.Z. Aliber (2011), *Manias, Panics and Crashes*, Palgrave MacMillan.

- Koopmans, T.C. (1957), *Three Essays on the State of Economic Science*, McGraw-Hill.
- Lewis, M. (2003), *Moneyball*, Norton.
- Olson, M. (1971), *The Logic of Collective Action: Public Goods and the Theory of Groups*, Revised Edition, Harvard.
- Pesaran, M.H. and M. Weale (2006), “Survey Expectations,” In G. Elliot, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting*, 715-776.
- Plott, C. (2000), “Markets as Information Gathering Tools,” *Southern Economic Journal*, 67, 2–15.
- Rothchild, D. and J. Wolfers (2013), “Forecasting Elections: Voter Intentions versus Expectations ,” Manuscript, University of Michigan.
- Schelling, T.C. (1980), *The Strategy of Conflict*, Revised Edition, Harvard.
- Shiller, R.J. (2005), *Irrational Exuberance*, Second Edition, Princeton University Press.
- Silver, N.. (2012), *The Signal and the Noise*, Penguin Press.
- Snowberg, E., J. Wolfers, and E. Zitzewitz (2013), “Prediction Markets for Economic Forecasting,” In G. Elliott and A. Timmermann (eds), *Handbook of Economic Forecasting*, Volume 2, Elsevier.
- Surowiecki, J. (2004), *The Wisdom of Crowds: Why the Many Are Smarter Than the Few, and How Collective Wisdom Shapes Business, Economies, Societies and Nations*, Random House.
- Taleb, N.N. (2007), *The Black Swan*, Random House.
- Taleb, N.N. (2008), *Fooled by Randomness*, Revised Edition, Random House.

- Tetlock, P.E. (2006), *Expert Political Judgment: How Good Is It? How Can We Know?*, Princeton University Press.
- Tetlock, P.E. and D. Gardner (2015), *Superforecasting: The Art and Science of Prediction*, Crown.
- Timmermann, A. (2006), “Forecast Combinations,” In G. Elliot, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting*, 135-196.
- Wonnacott, T.H. and R.J. Wonnacott (1990), *Introductory Statistics*. New York: John Wiley and Sons, Fifth Edition.

# Index

- $R^2$ , 290  
 $h$ -step-ahead forecast, 29  
 $s^2$ , 291  
Absolute error loss, 28  
Absolute loss, 28  
Adjusted  $R^2$ , 291  
Aggregation, 36, 130  
Akaike Information Criterion ( $AIC$ ),  
    291  
Analog principle, 72  
Argmin, 45  
Asymmetric loss, 26, 34  
Autocorrelation function, 64  
Autocovariance function, 62  
Autoregression, 65  
Bartlett bands, 99  
Bias, 330  
Bias correction, 52  
Bias-variance tradeoff, 52  
binomial logit, 322  
Box-Pierce Q-Statistic, 73  
Breusch-Godfrey test, 98  
Calendar effects, 53  
Conditional mean and variance, 69  
Covariance stationary, 62  
CUSUM, 438  
CUSUM plot, 438  
Cycles, 61  
Data mining, 290  
Data-generating process (DGP), 292  
Decision environment, 20  
Density forecast, 23  
Deterministic seasonality, 42  
Deterministic trend, 42  
Detrending, 193  
Diebold-Mariano statistic, 336  
Direction-of-change forecast, 28  
Disaggregation, 36, 130  
Distributed lag, 71  
Dummy left-hand-side variable, 317  
Dummy right-hand-side variable, 317  
Durbin's h test, 98  
Error variance, 330  
Event outcome, 20  
Event timing, 20  
Ex post smoothing, 193  
Expert Opinion, 20

- Exponential smoothing, 185  
Exponential trend, 43  
Exponentially-weighted moving average, 185  
First-order serial correlation, 98  
Forecast accuracy comparison, 329  
Forecast error, 27  
Forecast evaluation, 329  
Forecast horizon, 20  
Forecast object, 19  
Forecast statement, 20  
Gaussian white noise, 67  
Generalized linear model, 323  
GLM, 323  
h-step-ahead extrapolation forecast, 29  
Holiday variation, 53  
Holt-Winters smoothing, 188  
Holt-Winters smoothing with seasonality, 188  
In-sample overfitting, 290  
Independent white noise, 66  
Indicator variable, 317  
Information set, 19  
Intercept, 42  
Interval forecast, 23  
Lag operator, 70  
Limited dependent variable, 317  
Linear probability model, 318  
Linear trend, 42  
Linex loss, 35  
Link function, 323  
Linlin loss, 35  
Ljung-Box Q-Statistic, 73  
Log-linear trend, 43  
Logistic function, 318  
Logistic trend, 51  
Logit model, 318  
Loss function, 20  
Mean absolute error, 331  
Mean error, 330  
Mean squared error, 290, 330  
Measurement error, 36  
Missing observations, 36  
Model Complexity, 20  
Model improvement, 19  
Model selection consistency, 292  
Model selection efficiency, 293  
Model uncertainty, 19  
Moments, 69  
Multinomial logit, 324  
Multivariate information set, 22  
Noise, 186  
Nonlinear least squares, 45  
Nonseasonal fluctuations, 52  
Normal white noise, 67

- Odds, 323
- Off-line smoothing, 193
- On-line smoothing, 193
- Optimal forecast, 28
- Ordered logit, 320
- Ordered outcomes, 319
- Ordinary least squares regression, 44
- Out-of-sample 1-step-ahead prediction error variance, 290
- Outliers, 36
- Parameter instability, 437
- Parsimony principle, 20
- Partial autocorrelation function, 65
- Periodic models, 53
- Peso problem, 344
- Phase shift, 185
- Point forecast, 23
- Polynomial in the lag operator, 70
- Population regression, 65
- Probability forecast, 25
- Probability forecasts, 317
- Probit model, 323
- Proportional odds, 320
- Quadratic loss, 28
- Quadratic trend, 42
- Ragged edges, 36
- Random number generator, 96
- Random walk, 438
- Real-time smoothing, 193
- Realization, 62
- Recursive residuals, 437
- Recursive structure, 187
- Regression on seasonal dummies, 44
- Root mean squared error, 331
- Sample autocorrelation function, 72
- Sample mean, 72
- Sample partial autocorrelation, 74
- Sample path, 62
- Schwarz Information Criterion (*SIC*), 291
- Seasonal adjustment, 52
- Seasonal dummy variables, 44
- Seasonally-adjusted series, 52
- Seasonals, 41
- Second-order stationarity, 64
- Serial correlation, 66
- Signal, 186
- Simple exponential smoothing, 185
- Single exponential smoothing, 185
- Slope, 42
- Standardized recursive residuals, 437
- Stochastic seasonality, 42
- Stochastic trend, 42
- Strong white noise, 66
- Sum of squared residuals, 290
- Symmetric loss, 26, 34
- Time dummy, 42

- Time series, 20, 62
- Time series process, 66
- Time-varying parameters, 440
- Tobit model, 324
- Top-down vs. bottom-up models, 130
- Trading-day variation, 53
- Trends, 41
- Unconditional mean and variance, 69
- Unforecastability principle, 333
- Univariate information set, 22
- Unobserved components, 20
- Weak stationarity, 64
- Weak white noise, 66
- White noise, 66
- Zero-mean white noise, 66