

Econometric Data Science: *A Predictive Modeling Approach*



Francis X. Diebold
University of Pennsylvania

Version 2019.01.14

Econometric Data Science:
A Predictive Modeling Approach

Econometric Data Science

Francis X. Diebold

Copyright © 2013-2019
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. (2019), *Econometric Data Science: A Predictive Modeling Approach*, Department of Economics, University of Pennsylvania, <http://www.ssc.upenn.edu/~fdiebold/Textbooks.html>.”

To my undergraduates,
who continually surprise and inspire me

Brief Table of Contents

About the Author	xix
About the Cover	xxi
Preface	xxix
I Beginnings	1
1 Introduction to Econometrics	3
2 Graphics and Graphical Style	13
II Cross Sections	29
3 Regression Under Ideal Conditions	31
4 Misspecification and Model Selection	73
5 Non-Normality	91
6 Group Heterogeneity and Indicator Variables	113
7 Nonlinearity	121
8 Heteroskedasticity	139
9 Limited Dependent Variables	151
10 Causal Estimation	161

III Time Series	177
11 Trend and Seasonality	179
12 Serial Correlation	205
13 Structural Change	251
14 Vector Autoregression	257
15 Dynamic Heteroskedasticity	277
IV Appendices	301
A Probability and Statistics Review	303
B Construction of the Wage Datasets	315
C Some Popular Books Worth Encountering	321

Detailed Table of Contents

About the Author	xix
About the Cover	xxi
Preface	xxix
I Beginnings	1
1 Introduction to Econometrics	3
1.1 Welcome	3
1.1.1 Who Uses Econometrics?	3
1.1.2 What Distinguishes Econometrics?	5
1.2 Types of Recorded Economic Data	5
1.3 Online Information and Data	6
1.4 Software	6
1.5 Tips on How to use this book	8
1.6 Exercises, Problems and Complements	10
1.7 Notes	11
2 Graphics and Graphical Style	13
2.1 Simple Techniques of Graphical Analysis	13
2.1.1 Univariate Graphics	14
2.1.2 Multivariate Graphics	14
2.1.3 Summary and Extension	17
2.2 Elements of Graphical Style	19
2.3 U.S. Hourly Wages	21
2.4 Concluding Remark	22
2.5 Exercises, Problems and Complements	22

2.6 Notes	26
2.7 Graphics Legend: Edward Tufte	28
II Cross Sections	29
3 Regression Under Ideal Conditions	31
3.1 Preliminary Graphics	31
3.2 Regression as Curve Fitting	33
3.2.1 Bivariate, or Simple, Linear Regression	33
3.2.2 Multiple Linear Regression	36
3.2.3 Onward	37
3.3 Regression as a Probability Model	38
3.3.1 A Population Model and a Sample Estimator	38
3.3.2 Notation, Assumptions and Results: The Ideal Conditions	39
A Bit of Matrix Notation	39
Assumptions: The Ideal Conditions (IC)	40
Results Under the IC	41
3.4 A Wage Equation	42
3.4.1 Mean dependent var 2.342	45
3.4.2 S.D. dependent var .561	45
3.4.3 Sum squared resid 319.938	45
3.4.4 Log likelihood -938.236	46
3.4.5 <i>F</i> statistic 199.626	48
3.4.6 Prob(<i>F</i> statistic) 0.000000	49
3.4.7 S.E. of regression .492	49
3.4.8 <i>R</i> -squared .232	50
3.4.9 Adjusted <i>R</i> -squared .231	50
3.4.10 Akaike info criterion 1.423	51
3.4.11 Schwarz criterion 1.435	51
3.4.12 A Bit More on AIC and SIC	52
3.4.13 Hannan-Quinn criter. 1.427	52
3.4.14 Durbin-Watson stat. 1.926	52
3.4.15 The Residual Scatter	53
3.4.16 The Residual Plot	53
3.5 Least Squares and Optimal Point Prediction	55

3.6	Optimal Interval and Density Prediction	57
3.7	Regression Output from a Predictive Perspective	59
3.8	Multicollinearity	60
3.8.1	Perfect and Imperfect Multicollinearity	61
3.9	Beyond Fitting the Conditional Mean: Quantile regression	62
3.10	Exercises, Problems and Complements	64
3.11	Notes	70
3.12	Regression's Inventor: Carl Friedrich Gauss	71
4	Misspecification and Model Selection	73
4.1	Information Criteria (Hard Thresholding)	74
4.2	Cross Validation (Hard Thresholding)	80
4.3	Stepwise Selection (Hard Thresholding)	81
4.3.1	Forward	82
4.3.2	Backward	82
4.4	Bayesian Shrinkage (Soft Thresholding)	83
4.5	Selection <i>and</i> Shrinkage (Mixed Hard and Soft Thresholding)	84
4.5.1	Penalized Estimation	84
4.5.2	The Lasso	84
4.6	Distillation: Principal Components	85
4.6.1	Distilling “ X Variables” into Principal Components . .	85
4.6.2	Principal Components Regression	87
4.7	Exercises, Problems and Complements	87
5	Non-Normality	91
5.0.1	Results	91
5.1	Assessing Normality	92
5.1.1	QQ Plots	92
5.1.2	Residual Sample Skewness and Kurtosis	93
5.1.3	The Jarque-Bera Test	93
5.2	Outliers	94
5.2.1	Outlier Detection	94
	Graphics	94
	Leave-One-Out and Leverage	95
5.3	Robust Estimation	95
5.3.1	Robustness Iteration	95

5.3.2	Least Absolute Deviations	96
5.4	Wage Determination	98
5.4.1	<i>WAGE</i>	98
5.4.2	<i>LWAGE</i>	98
5.5	Exercises, Problems and Complements	110
6	Group Heterogeneity and Indicator Variables	113
6.1	0-1 Dummy Variables	113
6.2	Group Dummies in the Wage Regression	115
6.3	Exercises, Problems and Complements	117
6.4	Notes	119
6.5	Dummy Variables, ANOVA, and Sir Ronald Fischer	120
7	Nonlinearity	121
7.1	Models Linear in Transformed Variables	122
7.1.1	Logarithms	122
7.1.2	Box-Cox and GLM	124
7.2	Intrinsically Non-Linear Models	125
7.2.1	Nonlinear Least Squares	125
7.3	Series Expansions	126
7.4	A Final Word on Nonlinearity and the IC	127
7.5	Selecting a Non-Linear Model	128
7.5.1	<i>t</i> and <i>F</i> Tests, and Information Criteria	128
7.5.2	The RESET Test	128
7.6	Non-Linearity in Wage Determination	129
7.6.1	Non-Linearity in Continuous and Discrete Variables Simultaneously	131
7.7	Exercises, Problems and Complements	132
8	Heteroskedasticity	139
8.1	Consequences of Heteroskedasticity for Estimation, Inference, and Prediction	139
8.2	Detecting Heteroskedasticity	140
8.2.1	Graphical Diagnostics	140
8.2.2	Formal Tests	141
	The Breusch-Pagan-Godfrey Test (BPG)	141
	White's Test	142

8.3	Dealing with Heteroskedasticity	143
8.3.1	Adjusting Standard Errors	143
8.3.2	Adjusting Density Forecasts	144
8.4	Exercises, Problems and Complements	145
9	Limited Dependent Variables	151
9.1	Binary Response	151
9.2	The Logit Model	153
9.2.1	Logit	153
9.2.2	Ordered Logit	154
9.2.3	Complications	155
9.3	Classification and “0-1 Forecasting”	156
9.4	Exercises, Problems and Complements	157
10	Causal Estimation	161
10.1	Predictive Modeling vs. Causal Estimation	162
10.1.1	Predictive Modeling and P-Consistency	163
10.1.2	Causal Estimation and T-Consistency	163
10.1.3	Correlation vs. Causality, and P-Consistency vs. T- Consistency	164
10.2	Reasons for Failure of IC2.1	165
10.2.1	Omitted Variables	165
10.2.2	Measurement Error	166
10.2.3	Simultaneity	167
10.3	Confronting Failures of IC2.1	167
10.3.1	Controloing for Omitted Variables	167
10.3.2	Instrumental Variables	168
10.3.3	Randomized Controlled Trials (RCT's) and Their Ap- proximation	169
	Regression Discontinuity Designs (RDD's)	170
	Propensity-Score Matching	171
	Event Studies (“Synthetic Controls”)	172
	Internal Validity and its Problems	173
	External Validity and its Problems	173
10.4	Exercises, Problems and Complements	174

III Time Series	177
11 Trend and Seasonality	179
11.1 Linear Trend	179
11.2 Non-Linear Trend	181
11.2.1 Quadratic Trend	181
11.2.2 Exponential Trend	182
11.2.3 Non-Linearity in Liquor Sales Trend	184
11.3 Seasonality	186
11.3.1 Seasonal Dummies	188
11.3.2 More General Calendar Effects	190
11.4 Trend and Seasonality in Liquor Sales	191
11.5 Exercises, Problems and Complements	193
11.6 Notes	203
12 Serial Correlation	205
12.1 Characterizing Serial Correlation (in Population, Mostly)	206
12.1.1 Covariance Stationary Time Series	207
12.1.2 Estimation and Inference for the Mean, Autocorrelation and Partial Autocorrelation Functions	212
Sample Mean	213
Sample Autocorrelations	214
Sample Partial Autocorrelations	217
12.2 Modeling Serial Correlation (in Population)	218
12.2.1 White Noise	218
12.2.2 The Lag Operator	224
12.2.3 Autoregression	226
The $AR(1)$ Process	227
The $AR(p)$ Process	234
12.3 Modeling Serial Correlation (in Sample)	236
12.3.1 Detecting Serial Correlation	236
Residual Scatterplots	237
Durbin-Watson	237
The Breusch-Godfrey Test	239
12.3.2 The Residual Correlogram	241
12.3.3 Estimating Serial Correlation	242
12.4 Exercises, Problems and Complements	245

13 Structural Change	251
13.1 Gradual Parameter Evolution	251
13.2 Abrupt Parameter Breaks	252
13.2.1 Exogenously-Specified Breaks	252
13.2.2 The Chow test with Endogenous Break Selection	253
13.3 Dummy Variables and Omitted Variables, Again and Again	254
13.3.1 Dummy Variables	254
13.3.2 Omitted Variables	254
13.4 Recursive Analysis and CUSUM	255
13.5 Structural Change in Liquor Sales Trend	255
13.6 Exercises, Problems and Complements	255
14 Vector Autoregression	257
14.1 Predictive Causality	259
14.2 Application: Housing Starts and Completions	261
14.3 Exercises, Problems and Complements	276
15 Dynamic Heteroskedasticity	277
15.1 The Basic ARCH Process	278
15.2 The GARCH Process	283
15.3 Extensions of ARCH and GARCH Models	290
15.3.1 Asymmetric Response	291
15.3.2 Exogenous Variables in the Volatility Function	292
15.3.3 Regression with GARCH disturbances and GARCH-M	292
15.3.4 Component GARCH	293
15.3.5 Mixing and Matching	294
15.4 Estimating, Forecasting and Diagnosing GARCH Models	294
15.5 Exercises, Problems and Complements	296
15.6 Notes	300
IV Appendices	301
A Probability and Statistics Review	303
A.1 Populations: Random Variables, Distributions and Moments .	303
A.1.1 Univariate	303
A.1.2 Multivariate	306

A.2	Samples: Sample Moments	307
A.2.1	Univariate	307
A.2.2	Multivariate	309
A.3	Finite-Sample and Asymptotic Sampling Distributions of the Sample Mean	310
A.3.1	Exact Finite-Sample Results	310
A.3.2	Approximate Asymptotic Results (Under Weaker Assumptions)	311
A.4	Exercises, Problems and Complements	312
B	Construction of the Wage Datasets	315
C	Some Popular Books Worth Encountering	321

About the Author



Francis X. Diebold is Professor of Economics, Finance and Statistics at the University of Pennsylvania. He has won both undergraduate and graduate economics “teacher of the year” awards, and his academic “family” includes thousands of undergraduate students and nearly 75 Ph.D. students. Diebold has published widely in econometrics, forecasting, finance, and macroeconomics. He is an NBER Faculty Research Associate, as well as an elected Fellow of the Econometric Society, the American Statistical Association, and the International Institute of Forecasters. He has also been the recipient of Sloan, Guggenheim, and Humboldt fellowships, Co-Director of the Wharton Financial Institutions Center, and President of the Society for Financial Econometrics. His academic research is firmly linked to practical matters: During 1986-1989 he served as an economist under both Paul Volcker and Alan Greenspan at the Board of Governors of the Federal Reserve System, during 2007-2008 he served as Executive Director of Morgan Stanley Investment Management, and during 2012-2013 he served as Chairman of the Federal Reserve System’s Model Validation Council.

About the Cover



The colorful painting is *Enigma*, by Glen Josselsohn, from Wikimedia Commons. As noted there:

Glen Josselsohn was born in Johannesburg in 1971. His art has been exhibited in several art galleries around the country, with a number of sell-out exhibitions on the South African art scene ... Glen's fascination with abstract art comes from the likes of Picasso, Pollock, Miro, and local African art.

I used the painting mostly just because I like it. But econometrics is indeed something of an enigma, part economics and part statistics, part science and part art, hunting faint and fleeting signals buried in massive noise. Yet, perhaps somewhat miraculously, it often succeeds.

List of Figures

1.1	Resources for Economists Web Page	7
1.2	R Homepage	8
1.3	Python Homepage	9
2.1	1-Year Goverment Bond Yield, Levels and Changes	15
2.2	Histogram of 1-Year Government Bond Yield	16
2.3	Bivariate Scatterplot, 1-Year and 10-Year Government Bond Yields	17
2.4	Scatterplot Matrix, 1-, 10-, 20- and 30-Year Government Bond Yields	18
2.5	Distributions of Wages and Log Wages	22
2.6	Tufte Teaching, with a First Edition Book by Galileo	28
3.1	Distributions of Log Wage, Education and Experience	32
3.2	(Log Wage, Education) Scatterplot	34
3.3	(Log Wage, Education) Scatterplot with Superimposed Regression Line	35
3.4	Regression Output	42
3.5	Wage Regression Residual Scatter	53
3.6	Wage Regression Residual Plot	55
3.7	Carl Friedrich Gauss	71
4.1	Lasso and Ridge Comparison	86
5.1	OLS Wage Regression	99
5.2	OLS Wage Regression: Residual Plot	100
5.3	OLS Wage Regression: Residual Histogram and Statistics	101
5.4	OLS Wage Regression: Residual Gaussian QQ Plot	102
5.5	OLS Wage Regression: Leave-One-Out Plot	103
5.6	LAD Wage Regression	104

5.7 OLS Log Wage Regression	105
5.8 OLS Log Wage Regression: Residual Plot	106
5.9 OLS Log Wage Regression: Residual Histogram and Statistics	107
5.10 OLS Log Wage Regression: Residual Gaussian QQ Plot	108
5.11 OLS Log Wage Regression: Leave-One-Out Plot	109
5.12 LAD Log Wage Regression	110
6.1 Histograms for Wage Covariates	114
6.2 Wage Regression on Education and Experience	116
6.3 Wage Regression on Education, Experience and Group Dummies	116
6.4 Residual Scatter from Wage Regression on Education, Experience and Group Dummies	117
6.5 Sir Ronald Fischer	120
7.1 Basic Linear Wage Regression	130
7.2 Quadratic Wage Regression	131
7.3 Wage Regression on Education, Experience, Group Dummies, and Interactions	132
7.4 Wage Regression with Continuous Non-Linearities and Interactions, and Discrete Interactions	133
7.5 Regression Output	137
8.1 Final Wage Regression	141
8.2 Squared Residuals vs. Years of Education	142
8.3 BPG Test Regression and Results	143
8.4 White Test Regression and Results	144
8.5 Wage Regression with Heteroskedasticity-Robust Standard Errors	145
8.6 Regression Weighted by Fit From White Test Regression . . .	148
11.1 Various Linear Trends	180
11.2 Various Quadratic Trends	182
11.3 Various Exponential Trends	183
11.4 Log-Quadratic Trend Estimation	185
11.5 Residual Plot, Log-Quadratic Trend Estimation	186
11.6 Liquor Sales Log-Quadratic Trend Estimation with Seasonal Dummies	187

11.7 Residual Plot, Liquor Sales Log-Quadratic Trend Estimation With Seasonal Dummies	188
11.8 Liquor Sales	191
11.9 Log Liquor Sales	192
11.10 Linear Trend Estimation	193
11.11 Residual Plot, Linear Trend Estimation	194
11.12 Estimation Results, Linear Trend with Seasonal Dummies	195
11.13 Residual Plot, Linear Trend with Seasonal Dummies	196
11.14 Seasonal Pattern	197
12.1	213
12.2	214
12.3	220
12.4	221
12.5	222
12.6	228
12.7	229
12.8	230
12.9	231
12.10	232
12.11	236
12.12	237
12.13	238
12.14 BG Test Equation, 4 Lags	240
12.15 BG Test Equation, 8 Lags	241
12.16 Residual Correlogram From Trend + Seasonal Model	243
12.17 Trend + Seasonal Model with Four Autoregressive Lags	244
12.18 Trend + Seasonal Model with Four Lags of y , Residual Plot	245
12.19 Trend + Seasonal Model with Four Autoregressive Lags, Residual Scatterplot	246
12.20 Trend + Seasonal Model with Four Autoregressive Lags, Residual Autocorrelations	247
14.1 Housing Starts and Completions, 1968 - 1996	262
14.2 Housing Starts Correlogram	263
14.3 Housing Starts Autocorrelations and Partial Autocorrelations	264
14.4 Housing Completions Correlogram	265

14.5 Housing Completions Autocorrelations and Partial Autocorrelations	266
14.6 Housing Starts and Completions Sample Cross Correlations	266
14.7 VAR Starts Model	267
14.8 VAR Starts Residual Correlogram	268
14.9 VAR Starts Equation - Sample Autocorrelation and Partial Autocorrelation	269
14.10 VAR Completions Model	270
14.11 VAR Completions Residual Correlogram	271
14.12 VAR Completions Equation - Sample Autocorrelation and Partial Autocorrelation	272
14.13 Housing Starts and Completions - Causality Tests	273
14.14 Housing Starts Forecast	274
14.15 Housing Starts Forecast and Realization	274
14.16 Housing Completions Forecast	275
14.17 Housing Completions Forecast and Realization	275

List of Tables

2.1 Yield Statistics	25
--------------------------------	----

Preface

Econometric Data Science: A Predictive Modeling Approach should be useful to students in a variety of fields – in economics, of course, but also statistics, business, finance, public policy, and even engineering. The “predictive modeling” perspective, emphasized throughout, connects students to the modern perspectives of machine learning, data science, etc., in both causal and non-causal environments.

I have used the material successfully for many years in my undergraduate econometrics course at Penn, as background for various other undergraduate courses, and in master’s-level executive education courses given to professionals in economics, business, finance and government. It is directly accessible at the undergraduate and master’s levels, as the only prerequisite is an introductory probability and statistics course.

Many people have contributed to the development of this book. One way or another, all of the following deserve thanks: Xu Cheng, University of Pennsylvania; Barbara Chizzolini, Bocconi University; Frank Di Traglia, University of Pennsylvania; Carlo Favero, Bocconi University; Bruce Hansen, University of Wisconsin; Frank Schorfheide, University of Pennsylvania; Jim Stock, Harvard University; Mark Watson, Princeton University.

I am especially grateful to the University of Pennsylvania, which for many years has provided an unparalleled intellectual home, the perfect incubator for the ideas that have congealed here. Related, I am grateful to an army of

energetic and enthusiastic Penn graduate and undergraduate students, who read and improved much of the manuscript and code.

Finally, I apologize and accept full responsibility for the many errors and shortcomings that undoubtedly remain – minor and major – despite ongoing efforts to eliminate them.

Francis X. Diebold

Philadelphia

Monday 14th January, 2019

Econometric Data Science

Part I

Beginnings

Chapter 1

Introduction to Econometrics

1.1 Welcome

1.1.1 Who Uses Econometrics?

Econometrics is important — it is used constantly in business, finance, economics, government, consulting and many other fields. Econometric models are used routinely for tasks ranging from data description to policy analysis, and ultimately they guide many important decisions.

To develop a feel for the tremendous diversity of econometric applications, let's explore some of the areas where they feature prominently, and the corresponding diversity of decisions supported.

One key field is economics (of course), broadly defined. Governments, businesses, policy organizations, central banks, financial services firms, and economic consulting firms around the world routinely use econometrics.

Governments, central banks and policy organizations use econometric models to guide monetary policy, fiscal policy, as well as education and training, health, and transfer policies.

Businesses use econometrics for strategic planning tasks. These include management strategy of all types including operations management and control (hiring, production, inventory, investment, ...), marketing (pricing, distributing, advertising, ...), accounting (budgeting revenues and expenditures),

and so on.

Sales modeling is a good example. Firms routinely use econometric models of sales to help guide management decisions in inventory management, sales force management, production planning, and new market entry.

More generally, business firms use econometric models to help 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 also use forecasts of future prices and availability of inputs to guide production decisions.

Econometric models are also crucial in financial services, including asset management, asset pricing, mergers and acquisitions, investment banking, and insurance. Portfolio managers, for example, are keenly interested in the empirical modeling and understanding of asset returns (stocks, bonds, exchange rates, commodity prices, ...).

Econometrics is similarly central to financial risk management. In recent decades, econometric methods for volatility modeling have been developed and widely applied to evaluate and insure risks associated with asset portfolios, and to price assets such as options and other derivatives.

Finally, econometrics is central to the work of a wide variety of consulting firms, many of which support the business functions already mentioned. Litigation support, for example, is also a very active area, in which econometric models are routinely used for damage assessment (e.g., lost earnings), “but for” analyses, and so on.

Indeed these examples are just the tip of the iceberg. Surely you can think

of many more situations in which econometrics is used.

1.1.2 What Distinguishes Econometrics?

Econometrics is much more than just “statistics using economic data,” although it is of course very closely related to statistics.

- Econometrics has special focus on prediction. In many respects the goal of econometrics is to help agents (consumers, firms, investors, policy makers, ...) make better decisions, and good forecasts are key inputs to good decisions.
- Econometrics must confront the special issues and features that arise routinely in economic data, such as trends, seasonality and cycles.
- Econometrics must confront the special problems arising due to its largely non-experimental nature: Model mis-specification, structural change, etc.

With so many applications and issues in econometrics, you might fear that a huge variety of econometric techniques exists, and that you’ll have to master all of them. Fortunately, that’s not the case. Instead, a relatively small number of tools form the common core of much econometric modeling. We will focus on those underlying core principles.

1.2 Types of Recorded Economic Data

Several aspects of economic data will concern us frequently.

One issue is whether the data are continuous or binary. **Continuous data** take values on a continuum, as for example with GDP growth, which in principle can take any value in the real numbers. **Binary data**, in contrast, take just two values, as with a 0-1 indicator for whether or not someone purchased a particular product during the last month.

Another issue is whether the data are recorded over time, over space, or some combination of the two. **Time series data** are recorded over time, as for example with U.S. GDP, which is measured once per quarter. A GDP dataset might contain quarterly data for, say, 1960 to the present. **Cross sectional data**, in contrast, are recorded over space (at a point in time), as with yesterday’s closing stock price for each of the U.S. S&P 500 firms. The data structures can be blended, as for example with a **time series of cross sections**. If, moreover, the cross-sectional units are identical over time, we speak of **panel data**, or **longitudinal data**. An example would be the daily closing stock price for each of the U.S. S&P 500 firms, recorded over each of the last 30 days.

1.3 Online Information and Data

Much useful information 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. (See Figure 1.1.) It contains hundreds of links to data sources, journals, professional organizations, and so on. [FRED \(Federal Reserve Economic Data\)](#) is a tremendously convenient source for economic data. 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. [Quandl](#) provides access to millions of data series on the web.

1.4 Software

Econometric software tools are widely available. Two good and time-honored high-level environments with extensive capabilities are [Stata](#) and [Eviews](#).

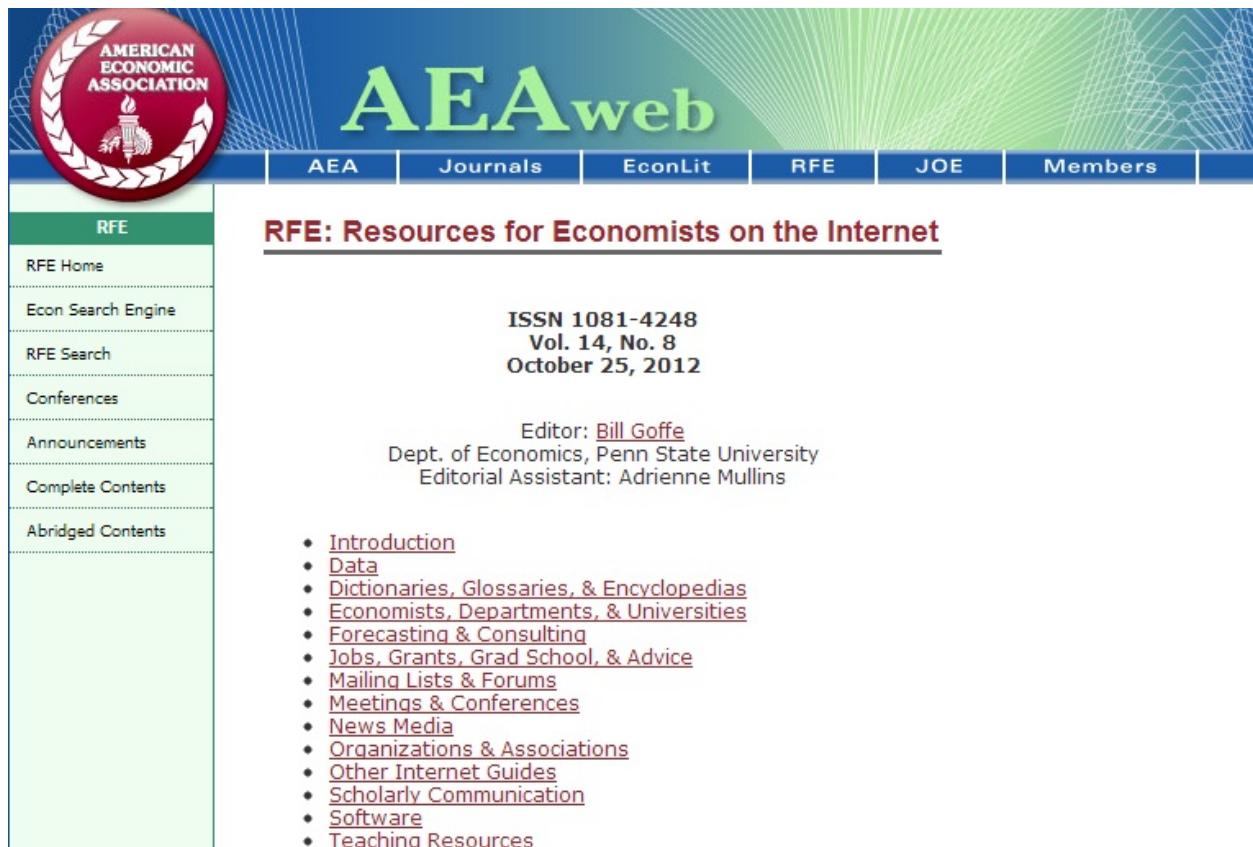


Figure 1.1: Resources for Economists Web Page

Stata has particular strength in **cross sections** and **panels**, and Eviews has particular strength in **time series**. Both reflect a balance of generality and specialization well-suited to the sorts of tasks that will concern us. If you feel more comfortable with another environment, however, that's fine – none of our discussion is wed to Stata or Eviews in any way.

There are also many flexible and more open-ended “mid-level” environments in which you can quickly program, evaluate, and apply new tools and techniques. **R** is one popular such environment, with special strengths in modern statistical methods and graphical data analysis. (See Figure 1.2.) Other notable environments include **Python** (see Figure 1.3) and **Julia**.

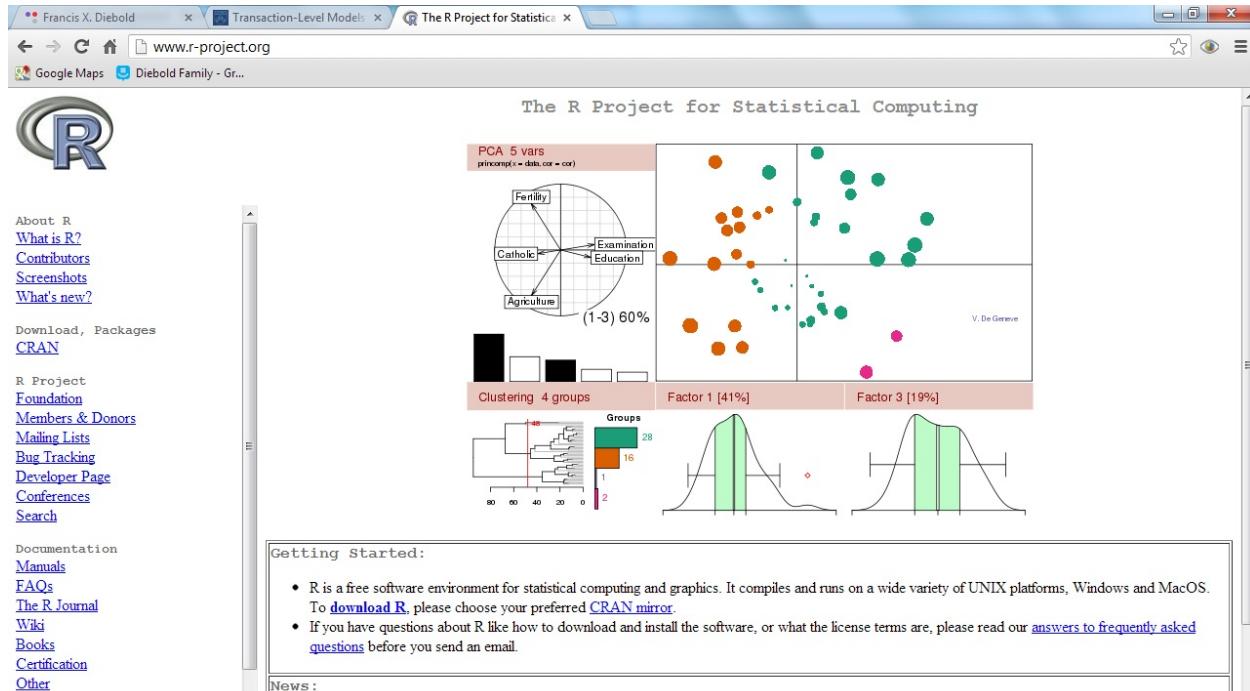


Figure 1.2: R Homepage

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 bibliographic references appear in green.
- Hyperlinks to the web appear in cyan.¹
- Hyperlinks to external files (e.g., video) appear in blue.
- Many images are clickable to reach related material.
- Key concepts appear in bold, and they also appear in the book's (hyperlinked) index.

¹Obviously web links sometimes go dead. I attempt effort to keep them updated. If you're encountering an unusual number of dead links, you're probably using an outdated edition of the book.

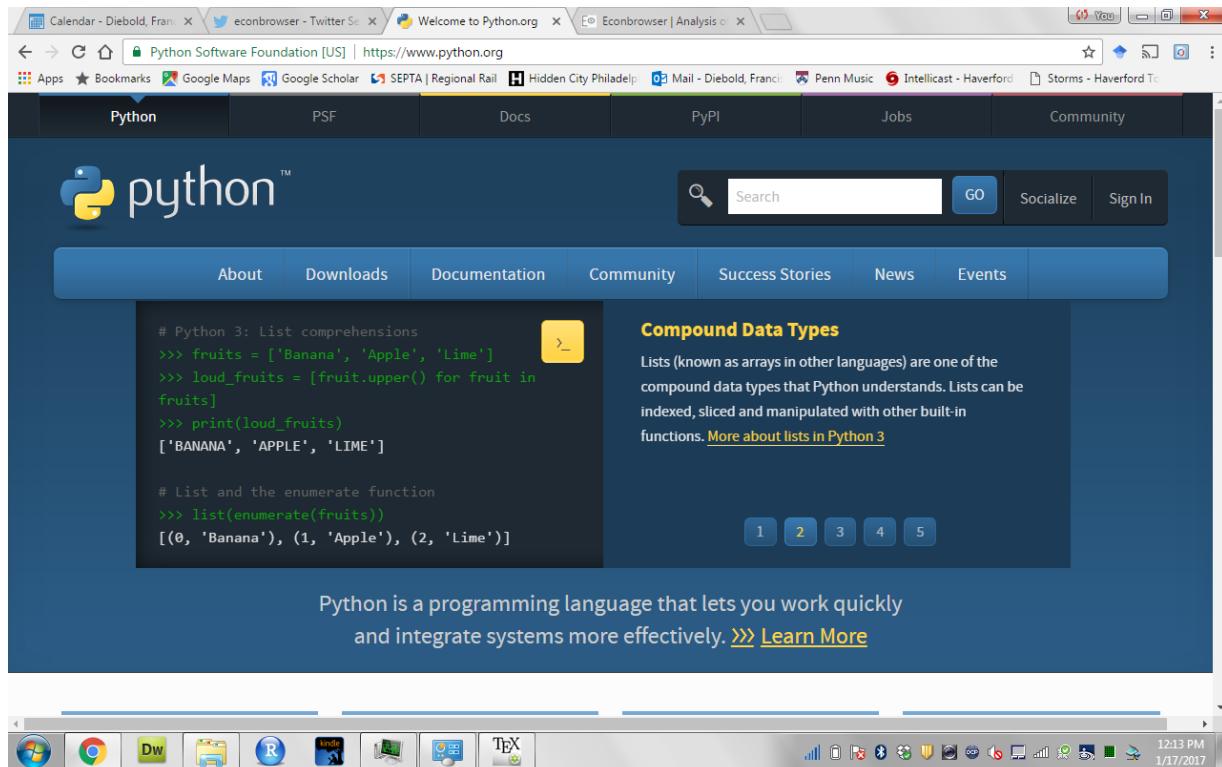


Figure 1.3: Python Homepage

- Additional related materials appear on [the book's web page](#). These may include book updates, presentation slides, datasets, and computer code.
- Facebook group: [Diebold Econometrics](#).
- Additional relevant material sometimes appears on Facebook groups [Diebold Forecasting](#) and [Diebold Time Series Econometrics](#), on Twitter @FrancisDiebold, and on the [No Hesitations](#) blog.
- The data that we use in the book from national income accounts, firms, people, financial and other markets, etc. – are fictitious. Sometimes they data 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 “Exercises, Problems and Complements” sections are of central importance and should be studied carefully. Exercises are generally straightforward checks of your understanding. Problems, in contrast, are generally significantly more involved, whether analytically or computationally. Complements generally introduce important auxiliary material not covered in the main text.

1.6 Exercises, Problems and Complements

1. (No empirical example is definitive)

Recall that, as mentioned in the text, most chapters contain at least one extensive empirical example. At the same time, those examples 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” at the end of each chapter as “Obtain the relevant data for the empirical modeling in this chapter, and produce a superior analysis”.

2. (Nominal, ordinal, interval and ratio data)

We emphasized time series, cross-section and panel data, whether continuous or discrete, but there are other complementary categorizations. In particular, distinctions are often made among **nominal data**, **ordinal data**, **interval data**, and **ratio data**. Which are most common and useful in economics and related fields, and why?

3. (Software differences and bugs: caveat emptor)

Be warned: no software is perfect. In fact, all software is highly imperfect. The results obtained when modeling 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.

1.7 Notes

R is available for free as part of a large and [highly-successful open-source project](#). RStudio provides a fine R working environment, and, like R, it's free. A good R tutorial, first given on Coursera and then moved to YouTube, is [here](#). R-bloggers is a massive compendium with all sorts of information about all things R. Quandl has a nice [R interface](#).

Python and Julia are also free.

Chapter 2

Graphics and Graphical Style

It's almost always a good idea to begin an econometric analysis with graphical data analysis. When compared to the modern array of econometric methods, graphical analysis might seem trivially simple, perhaps even so simple as to be incapable of delivering serious insights. Such is not the case: in many respects the human eye is a far more sophisticated tool for data analysis and modeling than even the most sophisticated statistical techniques. Put differently, graphics *is* a sophisticated technique. That's certainly not to say that graphical analysis alone will get the job done – certainly, graphical analysis has its limitations of its own – but it's usually the best place to start. With that in mind, we introduce in this chapter some simple graphical techniques, and we consider some basic elements of graphical style.

2.1 Simple Techniques of Graphical Analysis

We will segment our discussion into two parts: **univariate** (one variable) and **multivariate** (more than one variable). Because graphical analysis “lets the data speak for themselves,” it is most useful when the dimensionality of the data is low; that is, when dealing with univariate or low-dimensional multivariate data.

2.1.1 Univariate Graphics

First consider time series data. Graphics is used to reveal patterns in time series data. The great workhorse of univariate time series graphics is the simple **time series plot**, in which the series of interest is graphed against time.

In the top panel of Figure 2.1, for example, we present a time series plot of a 1-year Government bond yield over approximately 500 months. A number of important features of the series are apparent. Among other things, its movements appear sluggish and persistent, it appears to trend gently upward until roughly the middle of the sample, and it appears to trend gently downward thereafter.

The bottom panel of Figure 2.1 provides a different perspective; we plot the *change* in the 1-year bond yield, which highlights volatility fluctuations. Interest rate volatility is very high in mid-sample.

Univariate graphical techniques are also routinely used to assess distributional shape, whether in time series or cross sections. A **histogram**, for example, provides a simple estimate of the probability density of a random variable. The observed range of variation of the series is split into a number of segments of equal length, and the height of the bar placed at a segment is the percentage of observations falling in that segment.¹ In Figure 2.2 we show a histogram for the 1-year bond yield.

2.1.2 Multivariate Graphics

When two or more variables are available, the possibility of relations between the variables becomes important, and we use graphics to uncover the existence and nature of such relationships. We use **relational graphics** to

¹In some software packages (e.g., Eviews), the height of the bar placed at a segment is simply the number, not the percentage, of observations falling in that segment. Strictly speaking, such histograms are not density estimators, because the “area under the curve” doesn’t add to one, but they are equally useful for summarizing the shape of the density.

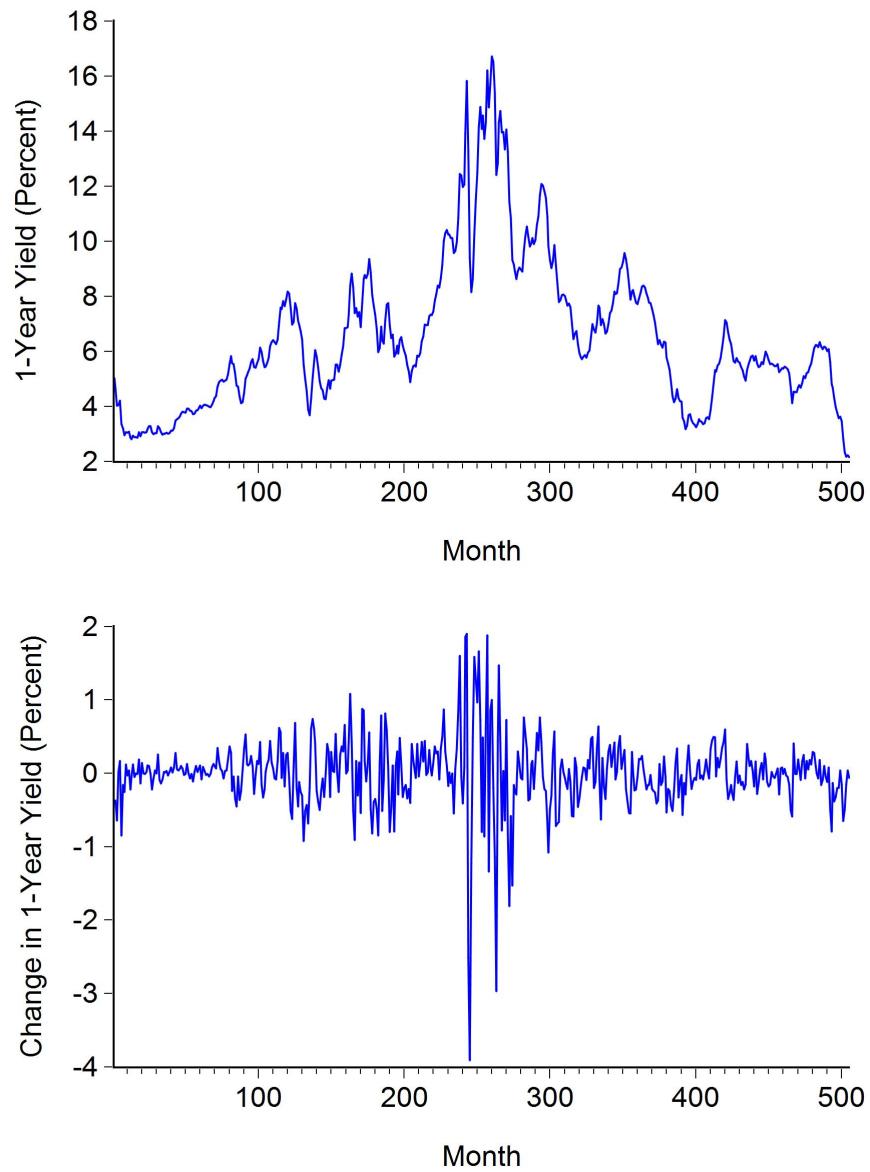


Figure 2.1: 1-Year Goverment Bond Yield, Levels and Changes

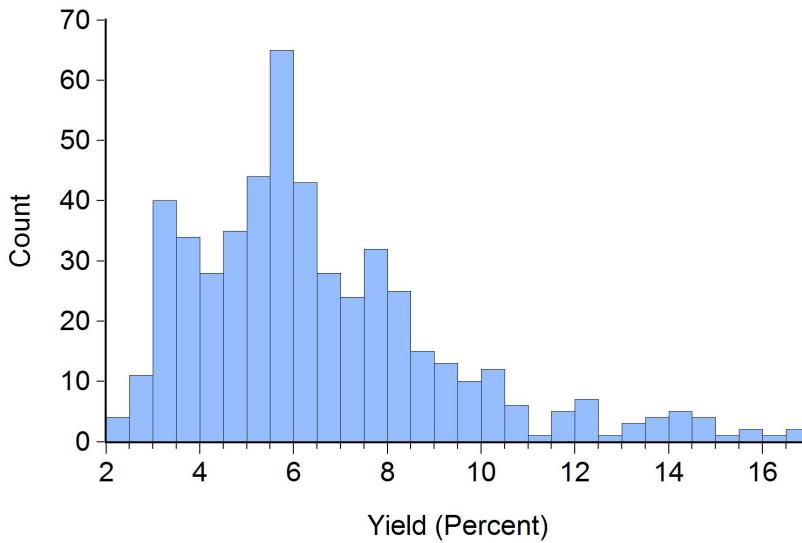


Figure 2.2: Histogram of 1-Year Government Bond Yield

display relationships and flag anomalous observations. You already understand the idea of a bivariate scatterplot.² In Figure 2.3, for example, we show a bivariate scatterplot of the 1-year U.S. Treasury bond rate vs. the 10-year U.S. Treasury bond rate, 1960.01-2005.03. The scatterplot indicates that the two move closely together; in particular, they are *positively correlated*.

Thus far all our discussion of multivariate graphics has been bivariate. That's because graphical techniques are best-suited to low-dimensional data. Much recent research has been devoted to graphical techniques for high-dimensional data, but all such high-dimensional graphical analysis is subject to certain inherent limitations.

One simple and popular scatterplot technique for high-dimensional data – and one that's been around for a long time – is the **scatterplot matrix**, or **multiway scatterplot**. The scatterplot matrix is just the set of all possible bivariate scatterplots, arranged in the upper right or lower left part of a matrix to facilitate comparisons. If we have data on N variables, there are

²Note that “connecting the dots” is generally not useful in scatterplots. This contrasts to time series plots, for which connecting the dots is fine and is typically done.

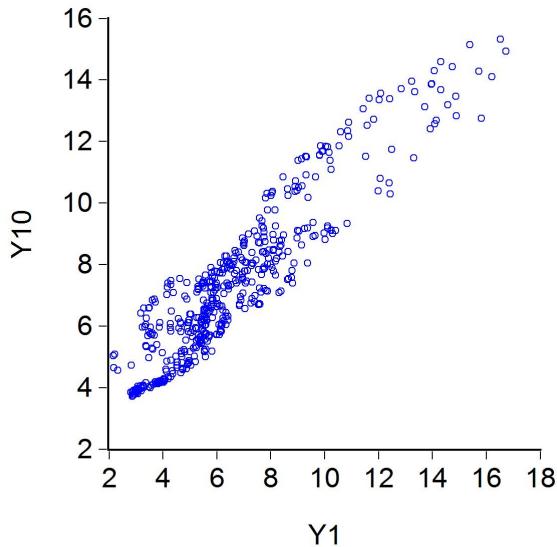


Figure 2.3: Bivariate Scatterplot, 1-Year and 10-Year Government Bond Yields

$\frac{N^2-N}{2}$ such pairwise scatterplots. In Figure 2.4, for example, we show a scatterplot matrix for the 1-year, 10-year, 20-year, and 30-year U.S. Treasury Bond rates, 1960.01-2005.03. There are a total of six pairwise scatterplots, and the multiple comparison makes clear that although the interest rates are closely related in each case, with a regression slope of approximately one, the relationship is more precise in some cases (e.g., 20- and 30-year rates) than in others (e.g., 1- and 30-year rates).

2.1.3 Summary and Extension

Let's summarize and extend what we've learned about the power of graphics:

- Graphics helps us summarize and reveal patterns in univariate time-series data. Time-series plots are helpful for learning about many features of time-series data, including trends, seasonality, cycles, the nature and location of any aberrant observations (“outliers”), structural breaks, etc.
- Graphics helps us summarize and reveal patterns in univariate cross-section data. Histograms are helpful for learning about distributional shape.

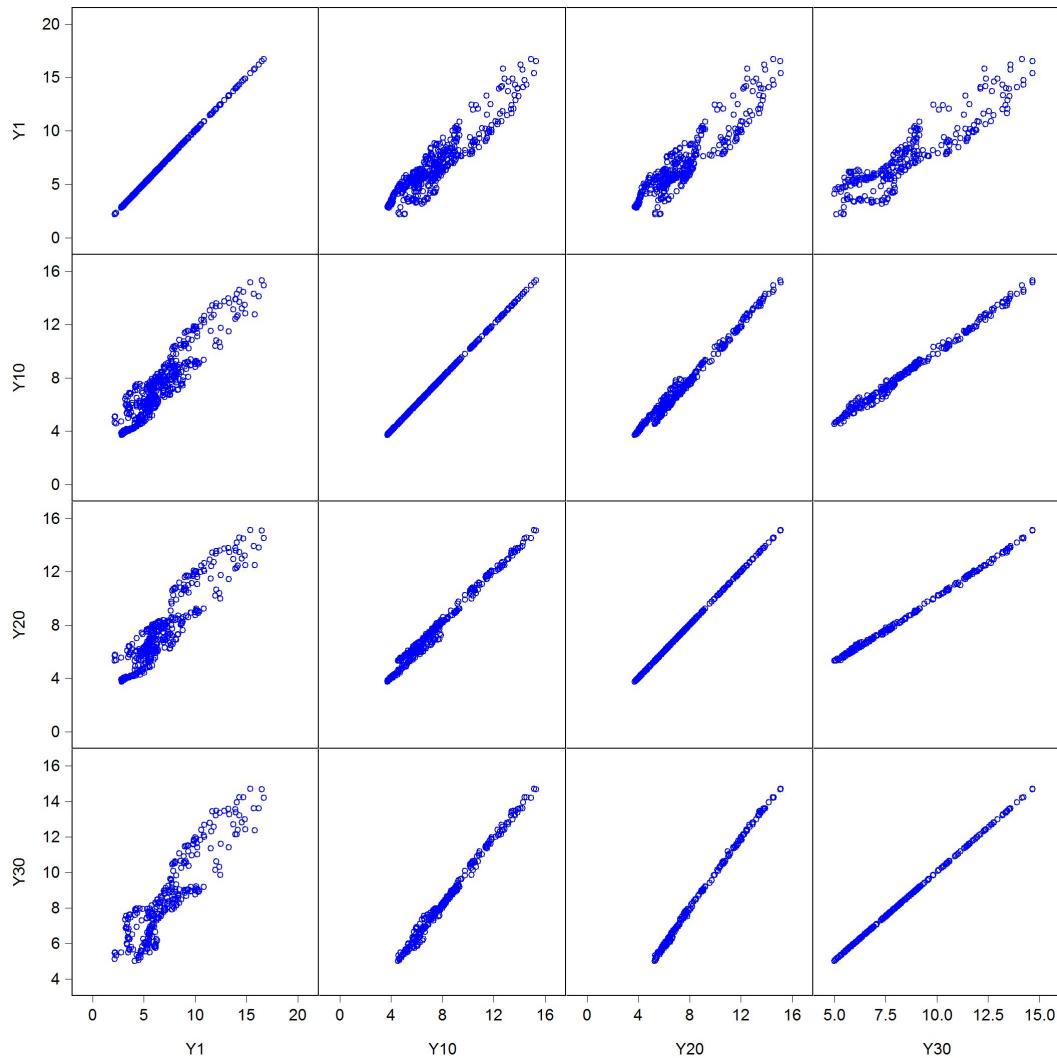


Figure 2.4: Scatterplot Matrix, 1-, 10-, 20- and 30-Year Government Bond Yields

- c. Graphics helps us identify relationships and understand their nature, in both multivariate time-series and multivariate cross-section environments. The key graphical device is the scatterplot, which can help us to begin answering many questions, including: Does a relationship exist? Is it linear or nonlinear? Are there outliers?
- d. Graphics helps us identify relationships and understand their nature in panel data. One can, for example, examine cross-sectional histograms across time periods, or time series plots across cross-sectional units.
- e. Graphics facilitates and encourages comparison of different pieces of data via **multiple comparisons**. The scatterplot matrix is a classic example of a multiple comparison graphic.

We might add to this list another item of tremendous relevance in our age of big data: Graphics enables us to summarize and learn from huge datasets.

2.2 Elements of Graphical Style

In the preceding sections we emphasized the power of graphics and introduced various graphical tools. As with all tools, however, graphical tools can be used effectively or ineffectively, and bad graphics can be far worse than no graphics. In this section you'll learn what makes good graphics good and bad graphics bad. In doing so you'll learn to use graphical tools effectively.

Bad graphics is like obscenity: it's hard to define, but you know it when you see it. Conversely, producing good graphics is like good writing: it's an iterative, trial-and-error procedure, and very much an art rather than a science. But that's not to say that anything goes; as with good writing, good graphics requires discipline. There are at least three keys to good graphics:

- a. Know your audience, and know your goals.

- b. Show the data, and only the data, within the bounds of reason.
- c. Revise and edit, again and again (and again). Graphics produced using software defaults are almost *never* satisfactory.

We can use a number of devices to *show the data*. First, avoid distorting the data or misleading the viewer, in order to reveal true data variation rather than spurious impressions created by design variation. Thus, for example, avoid changing scales in midstream, use **common scales** when performing multiple comparisons, and so on. The sizes of effects in graphics should match their size in the data.

Second, minimize, within reason, **non-data ink** (ink used to depict anything other than data points). Avoid **chartjunk** (elaborate shadings and grids that are hard to decode, superfluous decoration including spurious 3-D perspective, garish colors, etc.)

Third, choose a graph's **aspect ratio** (the ratio of the graph's height, h , to its width, w) to maximize pattern revelation. A good aspect ratio often makes the average absolute slope of line segments connecting the data points approximately equal 45 degrees. This procedure is called **banking to 45 degrees**.

Fourth, maximize graphical data density. Good graphs often display lots of data, indeed so much data that it would be impossible to learn from them in tabular form.³ Good graphics can present a huge amount of data in a concise and digestible form, revealing facts and prompting new questions, at both “micro” and “macro” levels.⁴

Graphs can often be shrunken greatly with no loss, as with **sparklines** (tiny graphics, typically time-series plots, meant to flow with text) and the

³Conversely, for small amounts of data, a good table may be much more appropriate and informative than a graphic.

⁴Note how maximization of graphical data density complements our earlier prescription to maximize the ratio of data ink to non-data ink, which deals with maximizing the *relative* amount of data ink. High data density involves maximizing as well the *absolute* amount of data ink.

sub-plots in multiple comparison graphs, increasing the amount of data ink per unit area.

2.3 U.S. Hourly Wages

We now begin our examination of CPS wage data, which we will use extensively. Here we use the 1995 CPS hourly wage data; for a detailed description see Appendix B. Figure 6.1 has four panels; consider first the left panels. In the upper left we show a histogram of hourly wage for the 1000+ people in the dataset. The distribution is clearly skewed right, with a mean around \$12/hour. In the lower left panel we show a density estimate (basically just a smoothed histogram) together with the best fitting normal distribution (a normal with mean and variance equal to the sample mean and sample variance of the wage data). Clearly the normal fits poorly.

The right panels of Figure 6.1 have the same structure, except that we now work with (natural) logarithms of the wages rather than the original “raw” wage data.⁵ The log is often used as a “symmetrizing” transformation for data with a right-skewed distribution, because the log transformation compresses things, pulling in long right tails. Sometimes taking logs can even produce approximate normality.⁶ Inspection of the log wage histogram in the upper right panel reveals that the log wage does indeed appear more symmetrically distributed, and comparison of the density estimate to the best-fitting normal in the lower-right panel indicates approximate normality of the log wage.

⁵Whenever we say “log” in this book, we mean “natural log”.

⁶Recall the famous lognormal density: A random variable x is defined to be lognormal if $\log(x)$ is normal. Hence if the wage data is approximately lognormally distributed, then, $\log(\text{wage})$ will be approximately normal. Of course lognormality may or may hold – whether data are lognormal is entirely an empirical matter.

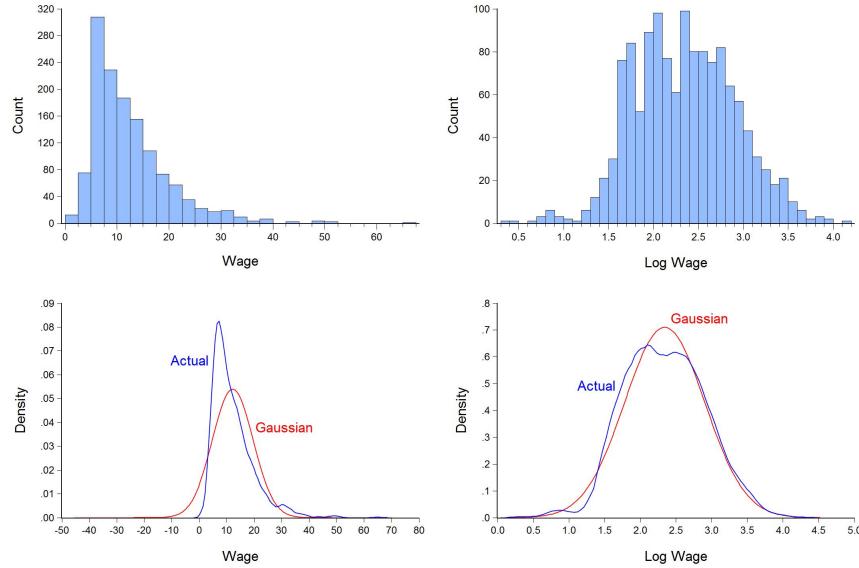


Figure 2.5: Distributions of Wages and Log Wages

2.4 Concluding Remark

Ultimately good graphics proceeds just like good writing, so if good writing is good thinking, then so too is good graphics. So the next time you hear an ignorant pronouncement along the lines of “I don’t like to write; I like to think,” rest assured, his writing, thinking, and graphics are likely all poor.

2.5 Exercises, Problems and Complements

1. (NBER recession bars: A useful graphical device)

In U.S. time-series situations it’s often useful to superimpose “NBER Recession Bars” on time-series plots, to help put things in context. You can find the dates of NBER expansions and contractions at <http://www.nber.org/cycles.html>.

2. (Empirical warm-up)
 - (a) Obtain time series of quarterly real GDP and quarterly real consumption for a country of your choice. Provide details.

- (b) Display time-series plots and a scatterplot (put consumption on the vertical axis).
- (c) Convert your series to growth rates in percent, and again display time series plots.
- (d) From now on use the growth rate series only.
- (e) For each series, provide summary statistics (e.g., mean, standard deviation, range, skewness, kurtosis, ...).
- (f) For each series, perform t-tests of the null hypothesis that the population mean growth rate is 2 percent.
- (g) For each series, calculate 90 and 95 percent confidence intervals for the population mean growth rate. For each series, which interval is wider, and why?
- (h) Regress consumption on GDP. Discuss.

3. (Simple vs. partial correlation)

The set of pairwise scatterplots that comprises a multiway scatterplot provides useful information about the joint distribution of the set of variables, but it's incomplete information and should be interpreted with care. A pairwise scatterplot summarizes information regarding the **simple correlation** between, say, x and y . But x and y may appear highly related in a pairwise scatterplot even if they are in fact unrelated, if each depends on a third variable, say z . The crux of the problem is that there's no way in a pairwise scatterplot to examine the correlation between x and y *controlling* for z , which we call **partial correlation**. When interpreting a scatterplot matrix, keep in mind that the pairwise scatterplots provide information only on simple correlation.

4. (Graphics and Big Data)

Another aspect of the power of statistical graphics comes into play in the analysis of large datasets, so it's increasingly more important in our era of “Big Data”: Graphics enables us to present a huge amount of data in a small space, and hence helps to make huge datasets coherent. We might, for example, have supermarket-scanner data, recorded in five-minute intervals for a year, on the quantities of goods sold in each of four food categories – dairy, meat, grains, and vegetables. Tabular or similar analysis of such data is simply out of the question, but graphics is still straightforward and can reveal important patterns.

5. (Color)

There is a temptation to believe that color graphics is always better than grayscale. That's often far from the truth, and in any event, color is typically best used sparingly.

- a. Color can be (and often is) chartjunk. How and why?
- b. Color has no natural ordering, despite the evident belief in some quarters that it does. What are the implications for “heat map” graphics? Might shades of a single color (e.g., from white or light gray through black) be better?
- c. On occasion, however, color can aid graphics both in showing the data and in appealing to the viewer. One key “show the data” use is in annotation. Can you think of others? What about uses in appealing to the viewer?
- d. Keeping in mind the principles of graphical style, formulate as many guidelines for color graphics as you can.

6. (Principles of tabular style)

The power of tables for displaying data and revealing patterns is very limited compared to that of graphics, especially in this age of Big Data.

Table 2.1: Yield Statistics

Maturity (Months)	\bar{y}	$\hat{\sigma}_y$	$\hat{\rho}_y(1)$	$\hat{\rho}_y(12)$
6	4.9	2.1	0.98	0.64
12	5.1	2.1	0.98	0.65
24	5.3	2.1	0.97	0.65
36	5.6	2.0	0.97	0.65
60	5.9	1.9	0.97	0.66
120	6.5	1.8	0.97	0.68

Notes: We present descriptive statistics for end-of-month yields at various maturities. We show sample mean, sample standard deviation, and first- and twelfth-order sample autocorrelations. Data are from the Board of Governors of the Federal Reserve System. The sample period is January 1985 through December 2008.

Nevertheless, tables are of course sometimes helpful, and there are principles of tabular style, just as there are principles of graphical style. Compare, for example, the nicely-formatted Table 2.1 (no need to worry about what it is or from where it comes...) to what would be produced by a spreadsheet such as Excel.

Try to formulate a set of principles of tabular style. (Hint: One principle is that vertical lines should almost never appear in tables, as in the table above.)

7. (More on graphical style: Appeal to the viewer)

Other graphical guidelines help us *appeal to the viewer*. First, use clear and modest type, avoid mnemonics and abbreviations, and use labels rather than legends when possible. Second, make graphics self-contained; a knowledgeable reader should be able to understand your graphics without reading pages of accompanying text. Third, as with our prescriptions for showing the data, avoid chartjunk.

8. (The “golden” aspect ratio, visual appeal, and showing the data)

A time-honored approach to visual graphical appeal is use of an aspect ratio such that height is to width as width is to the sum of height and width. This turns out to correspond to height approximately sixty percent of width, the so-called “**golden ratio**.” Graphics that conform to the golden ratio, with height a bit less than two thirds of width, are visually appealing. Other things the same, it’s a good idea to keep the golden ratio in mind when producing graphics. Other things are not always the same, however. In particular, the golden aspect ratio may not be the one that maximizes pattern revelation (e.g., by banking to 45 degrees).

9. (Graphics, non-profit and for-profit)

Check out the non-profit “community of creative people” at www.visualizing.org.

Check out Google Charts at <https://developers.google.com/chart/>. Poke around. What’s good? What’s bad? Can you use it to do sparklines?

Check out www.zevross.com.

2.6 Notes

R implements a variety of sophisticated graphical techniques and in many respects represents the cutting edge of statistical graphics software.

ggplot2 is a key R package that provides a broad catalog of graphics capabilities; see www.ggplot2.org. It implements the grammar of graphics developed by Leland Wilkenson, which allows you to produce highly customized graphics in a modular fashion. This grammar leads to a slightly unusual syntax, which must be learned, but once learned you can do almost anything. (The simple plot commands in R allow for some customization and have a

shorter learning curve, but they’re not as powerful.) ggplot2 documentation is at www.cran.r-project.org/web/packages/ggplot2/ggplot2.pdf. A helpful “cheatsheet” is at www.zevross.com/blog/2014/08/04/beautiful-plotting-in-r/#change-the-grid-lines-panel.grid.major.

2.7 Graphics Legend: Edward Tufte

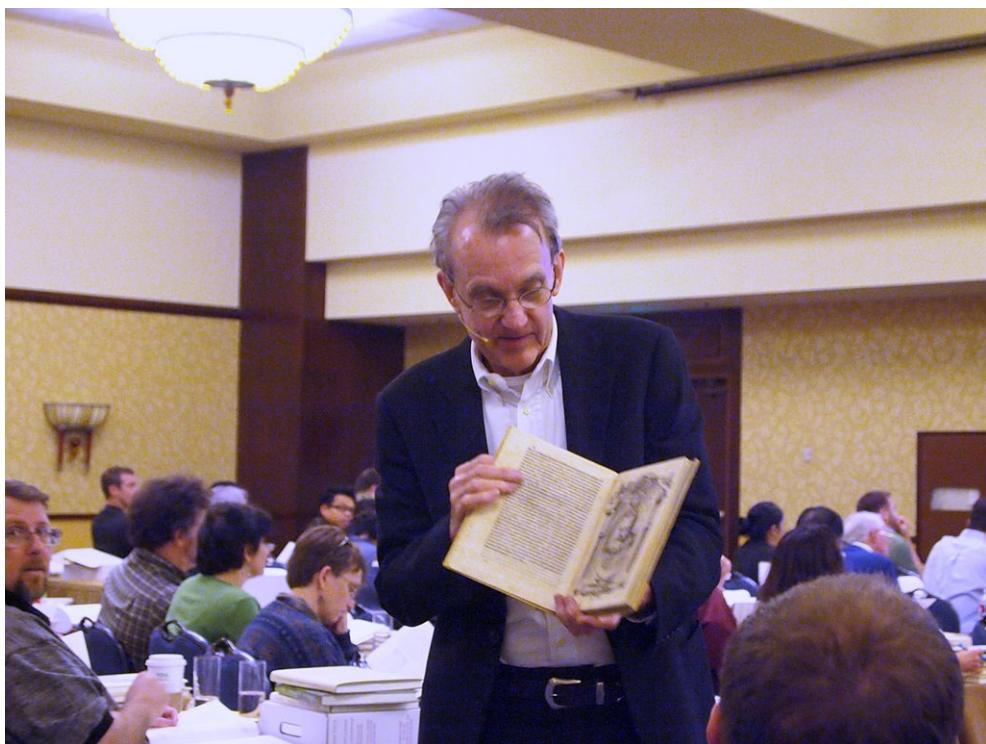


Figure 2.6: Tufte Teaching, with a First Edition Book by Galileo

This chapter has been heavily influenced by [Tufte \(1983\)](#), as are all modern discussions of statistical graphics.⁷ Tufte's book is an insightful and entertaining masterpiece on graphical style, and I recommend enthusiastically. Be sure to check out his [web page](#) and other books, which go far beyond his 1983 work.

⁷Photo details follow.

Date: 7 February 2011.

Source: <http://www.flickr.com/photos/roebot/5429634725/in/set-72157625883623225>.

Author: Aaron Fulkerson.

Originally posted to Flickr by Roebot at <http://flickr.com/photos/40814689@N00/5429634725>. Reviewed on 24 May 2011 by the FlickreviewR robot and confirmed to be licensed under the terms of the cc-by-sa-2.0. Licensed under the Creative Commons Attribution-Share Alike 2.0 Generic license.

Part II

Cross Sections

Chapter 3

Regression Under Ideal Conditions

You have already been introduced to probability and statistics, but chances are that you could use a bit of review before plunging into regression, so begin by studying Appendix A. Be warned, however: it is no substitute for a full-course introduction to probability and statistics, which you should have had already. Instead it is intentionally much more narrow, reviewing some material related to moments of random variables, which we will use repeatedly. It also introduces notation, and foreshadows certain ideas, that we develop subsequently in greater detail.

3.1 Preliminary Graphics

In this chapter we'll be working with cross-sectional data on log wages, education and experience. We already examined the distribution of log wages. For convenience we reproduce it in Figure 3.1, together with the distributions of the new data on education and experience.

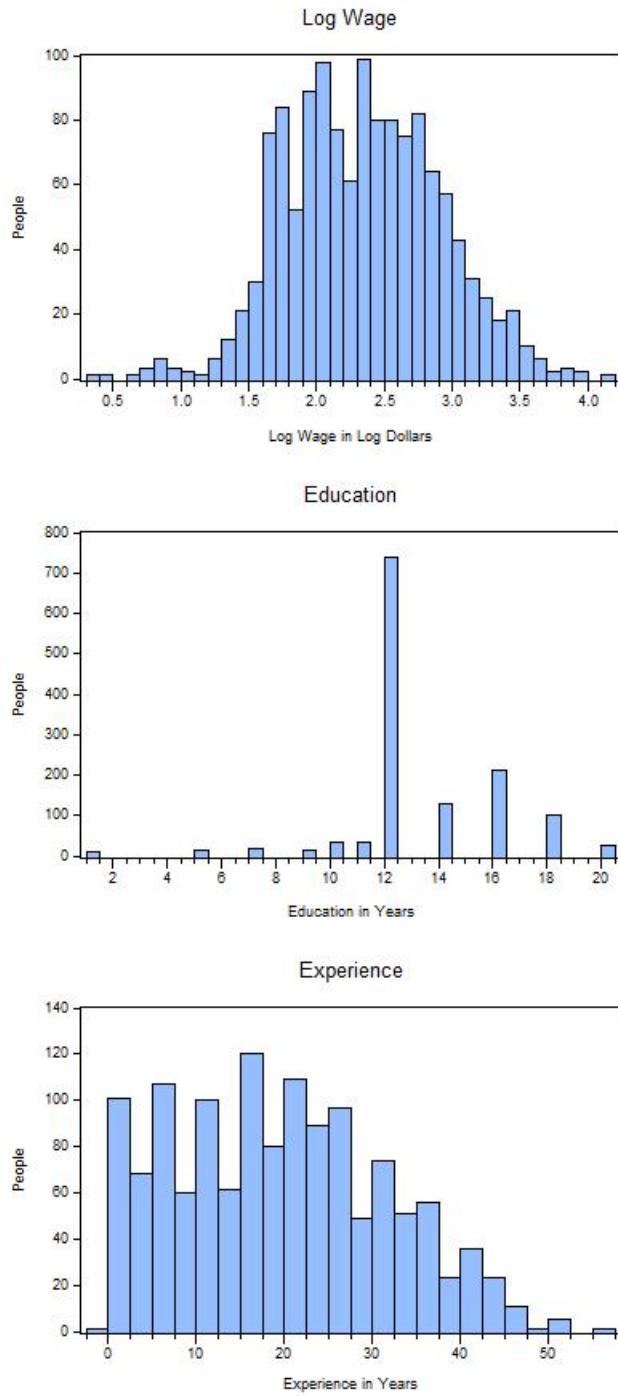


Figure 3.1: Distributions of Log Wage, Education and Experience

3.2 Regression as Curve Fitting

3.2.1 Bivariate, or Simple, Linear Regression

Suppose that we have data on two variables, y and x , as in Figure 3.2, and suppose that we want to find the linear function of x that best fits y , where “best fits” means that the sum of squared (vertical) deviations of the data points from the fitted line is as small as possible. When we “run a regression,” or “fit a regression line,” that’s what we do. The estimation strategy is called least squares, or sometimes “ordinary least squares” to distinguish it from fancier versions that we’ll introduce later.

The specific data that we show in Figure 3.2 are log wages (LWAGE, y) and education (EDUC, x) for a random sample of nearly 1500 people, as described in Appendix B.

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

$$\min_{\beta} \sum_{i=1}^N (y_i - \beta_1 - \beta_2 x_i)^2,$$

where β is shorthand notation for the set of two parameters, β_1 and β_2 . We denote the set of fitted parameters by $\hat{\beta}$, and its elements by $\hat{\beta}_1$ and $\hat{\beta}_2$.

It turns out that the β_1 and β_2 values that solve the least squares problem have well-known mathematical formulas. (More on that later.) We can use a computer to evaluate the formulas, simply, stably and instantaneously.

The fitted values are

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i,$$

$i = 1, \dots, N$. The residuals are the difference between actual and fitted values,

$$e_i = y_i - \hat{y}_i,$$

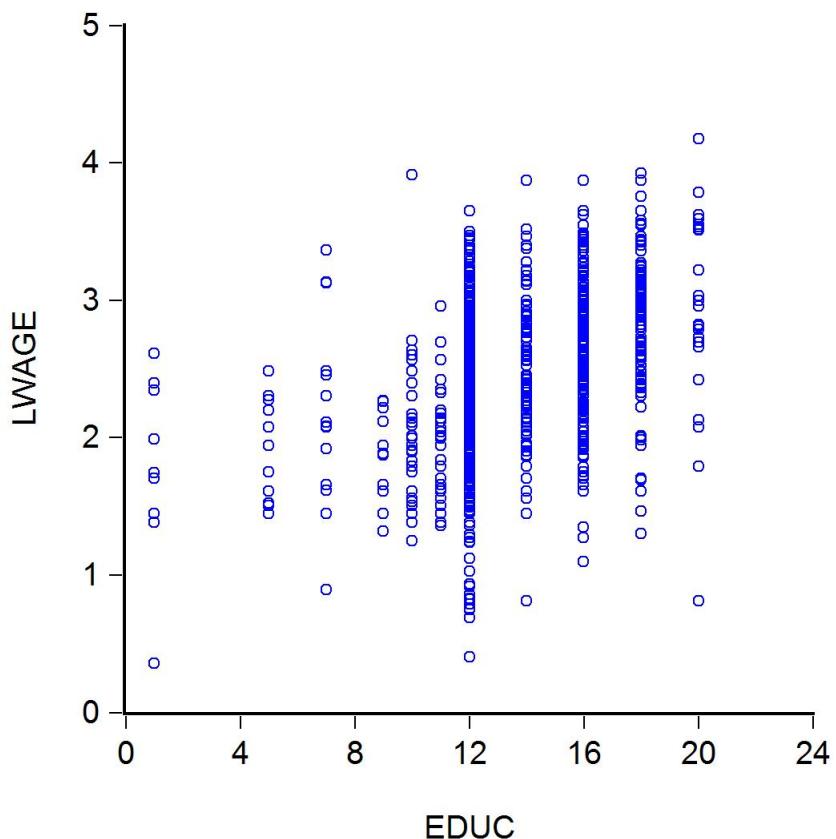


Figure 3.2: (Log Wage, Education) Scatterplot

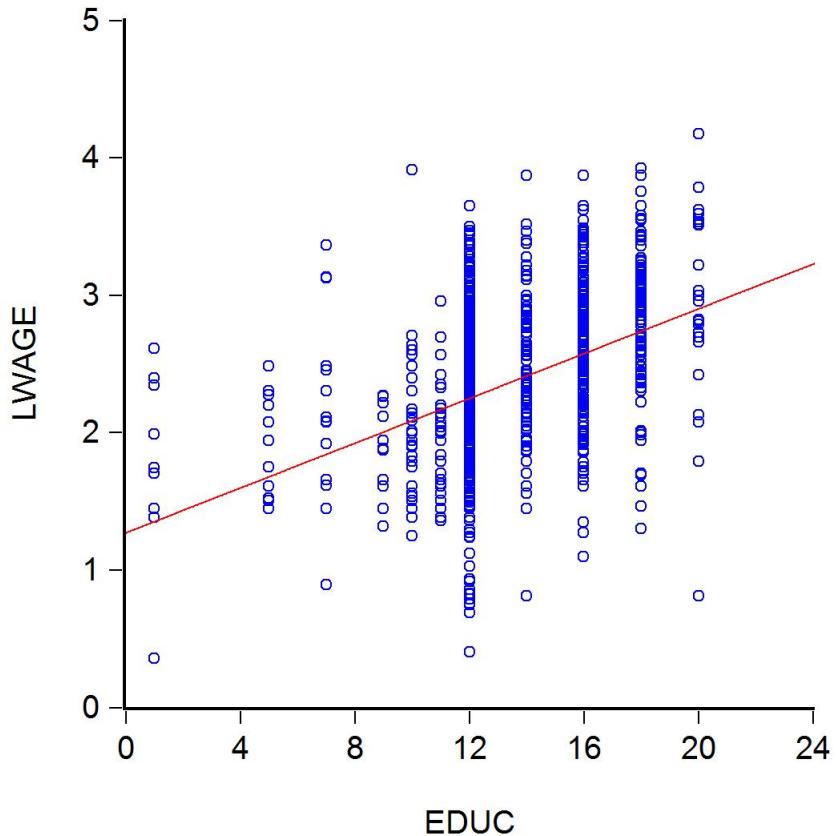


Figure 3.3: (Log Wage, Education) Scatterplot with Superimposed Regression Line

$i = 1, \dots, N$.

In Figure 3.3, we illustrate graphically the results of regressing LWAGE on EDUC. The best-fitting line slopes upward, reflecting the positive correlation between LWAGE and EDUC.¹ Note that the data points don't satisfy the fitted linear relationship exactly; rather, they satisfy it on average. To predict LWAGE for any given value of EDUC, we use the fitted line to find the value of LWAGE that corresponds to the given value of EDUC.

¹Note that use of *log* wage promotes several desiderata. First, it promotes normality, as we discussed in Chapter 2. Second, it enforces positivity of the fitted wage, because $\widehat{WAGE} = \exp(\widehat{LWAGE})$, and $\exp(x) > 0$ for any x .

Numerically, the fitted line is

$$\widehat{LWAGE} = 1.273 + .081 EDUC.$$

We conclude with a brief comment on notation. A standard cross-section notation for indexing the cross-sectional units is $i = 1, \dots, N$. A standard time-series notation for indexing time periods is $t = 1, \dots, T$. Much of our discussion will be valid in *both* cross-section and time-series environments, but we generally attempt to use the more standard notation in each environment.

3.2.2 Multiple Linear Regression

Everything generalizes to allow for more than one RHS variable. This is called multiple linear regression.

Suppose, for example, that we have two RHS variables, x_2 and x_3 . Before we fit a least-squares line to a two-dimensional data cloud; now we fit a least-squares plane to a three-dimensional data cloud. We use the computer to find the values of β_1 , β_2 , and β_3 that solve the problem

$$\min_{\beta} \sum_{i=1}^N (y_i - \beta_1 - \beta_2 x_{2i} - \beta_3 x_{3i})^2,$$

where β denotes the set of three model parameters. We denote the set of estimated parameters by $\hat{\beta}$, with elements $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$. The fitted values are

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i},$$

and the residuals are

$$e_i = y_i - \hat{y}_i,$$

$$i = 1, \dots, N.$$

For our wage data, the fitted model is

$$\widehat{LWAGE} = .867 + .093EDUC + .013EXPER.$$

Extension to the general multiple linear regression model, with an arbitrary number of right-hand-side (RHS) variables (K , including the constant), is immediate. The computer again does all the work. The fitted line is

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \dots + \hat{\beta}_K x_{Ki},$$

which we sometimes write more compactly as

$$\hat{y}_i = \sum_{k=1}^K \hat{\beta}_k x_{ki},$$

where $x_{1i} = 1$ for all i .

3.2.3 Onward

Before proceeding, two aspects of what we've done so far are worth noting. First, we now have two ways to analyze data and reveal its patterns. One is the graphical scatterplot of Figure 3.2, with which we started, which provides a visual view of the data. The other is the fitted regression line of Figure 3.3, which summarizes the data through the lens of a linear fit. Each approach has its merit, and the two are complements, not substitutes, but note that linear regression generalizes more easily to high dimensions.

Second, least squares as introduced thus far has little to do with statistics or econometrics. Rather, it is simply a way of instructing a computer to fit a line to a scatterplot in a way that's rigorous, replicable and arguably reasonable. We now turn to a probabilistic interpretation.

3.3 Regression as a Probability Model

We work with the full multiple regression model (simple regression is of course a special case). Collect the RHS variables into the vector x , where $x'_i = (1, x_{2i}, \dots, x_{Ki})$.

3.3.1 A Population Model and a Sample Estimator

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

$$y_i = \beta_1 + \beta_2 x_{2i} + \dots + \beta_K x_{Ki} + \varepsilon_i$$

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

$i = 1, \dots, N$. The intercept of the line is β_1 , the slope parameters are the other β 's, and the variance of the disturbance is σ^2 .² Collectively, we call the β 's (and σ) the model's parameters.

We assume that the 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 σ^2 . Our job is to *estimate* them using a sample of data from the population. We estimate the β 's precisely as before, using the computer to solve $\min_{\beta} \sum_{i=1}^N \varepsilon_i^2$.

²We speak of the regression intercept and the regression slope.

3.3.2 Notation, Assumptions and Results: The 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 we obtain.

A Bit of Matrix Notation

It will be useful to arrange all RHS variables into a matrix X . X has K columns, one for each regressor. Inclusion of a constant in a regression amounts to including a special RHS variable that is always 1. We put that in the leftmost column of the X matrix, which is just ones. The other columns contain the data on the other RHS variables, over the cross section in the cross-sectional case, or over time in the time-series case. Notationally, X is a $N \times K$ matrix.

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

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

$$y_i = \beta_1 + \beta_2 x_{2i} + \dots + \beta_K x_{Ki} + \varepsilon_i, \quad i = 1, \dots, N.$$

Alternatively, stack $y_i, i = 1, \dots, N$ into the vector y , where $y' = (y_1, y_2, \dots, y_N)$, and stack $\beta_j, j = 1, \dots, K$ into the vector β , where $\beta' = (\beta_1, \beta_2, \dots, \beta_K)$, and stack $\varepsilon_i, i = 1, \dots, N$, into the vector ε , where $\varepsilon' = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N)$. Then we can write the complete model over all observations as

$$y = X\beta + \varepsilon. \tag{3.1}$$

In addition,

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

becomes

$$\varepsilon \sim N(\underline{0}, \sigma^2 I). \quad (3.2)$$

This concise representation is very convenient.

Indeed representation (3.1)-(3.2) is crucially important, not simply because it is concise, but also because key results for estimation and inference may be stated very simply withing it, and because the various assumptions that we need to make to get various statistical results are most naturally and simply stated on X and ε in equation (3.1). We now proceed to discuss such assumptions.

Assumptions: The Ideal Conditions (IC)

1. The data-generating process (DGP) is:

$$y_i = \beta_1 + \beta_2 x_{2i} + \dots + \beta_K x_{Ki} + \varepsilon_i$$

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

and the fitted model matches it exactly.

2. ε_i is independent of (x_{1i}, \dots, x_{Ki}) , for all i

IC1 has many important sub-conditions embedded. For example:

1. The fitted model is correctly specified
2. The disturbances are Gaussian
3. The coefficients (β 's) are fixed (whether over space or time, depending on whether we're working in a time-series or cross-section environment)
4. The relationship is linear

- 5. The ε_i 's have constant variance σ^2
- 6. The ε_i 's are uncorrelated (whether over space or time, depending on whether we're working in a time-series or cross-section environment)

IC2 is more subtle, and it may seem obscure at the moment, but it is very important in the context of causal estimation, which we will discuss in chapter 10.

The IC's are surely heroic in many contexts, and much of econometrics is devoted to detecting and dealing with various IC failures. But before we worry about IC failures, it's invaluable first to understand what happens when they hold.³

Results Under the IC

The least squares estimator is

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

and under the IC it is (among other things) consistent, asymptotically efficient, and asymptotically normally distributed. We write

$$\hat{\beta}_{LS} \stackrel{a}{\sim} N(\beta, V).$$

We consistently estimate the covariance matrix V using $\hat{V} = s^2(X'X)^{-1}$, where $s^2 = \sum_{i=1}^N e_i^2 / (N - K)$.

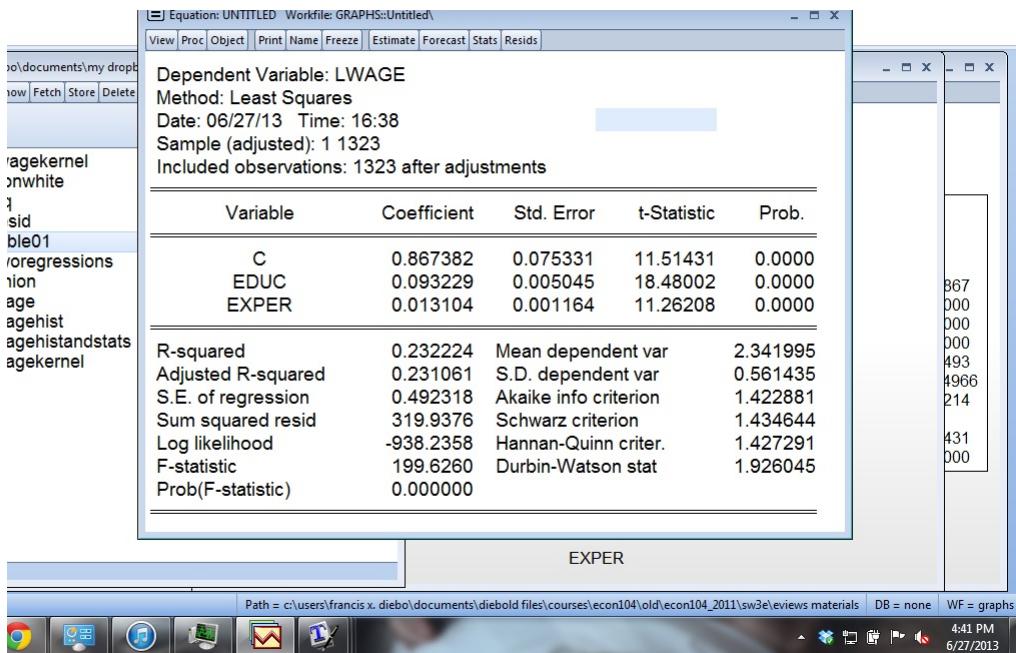


Figure 3.4: Regression Output

3.4 A Wage Equation

Now let's do more than a simple graphical analysis of the regression fit. Instead, let's look in detail at the computer output, which we show in Figure 6.2 for a regression of *LWAGE* on an intercept, *EDUC* and *EXPER*. We run regressions 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.

Before proceeding, note well that the IC may not be satisfied for this dataset, yet we will proceed assuming that they *are* satisfied. As we proceed through this book, we will confront violations of the various assumptions – indeed that's what econometrics is largely about – and we'll return repeatedly to this dataset and others. But we must begin at the beginning.

The software output begins by reminding us that we're running a least-

³Certain variations of the IC as stated above can be entertained, and in addition we have omitted some technical details.

squares (LS) regression, and that the left-hand-side (LHS) variable is the log wage (LWAGE), using a total of 1323 observations.

Next comes a table listing each RHS variable together with four statistics. The RHS variables EDUC and EXPER are education and experience, and the C variable refers to the earlier-mentioned intercept. The C variable always equals one, so the estimated coefficient on C is the estimated intercept of the regression line.⁴

The four statistics associated with each RHS variable are the estimated coefficient (“Coefficient”), its standard error (“Std. Error”), a t statistic, and a corresponding probability value (“Prob.”). The standard errors of the estimated coefficients indicate their likely sampling variability, and hence their reliability. The estimated coefficient plus or minus one standard error is approximately a 68% confidence interval for the true but unknown population parameter, 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, which will be true if the regression disturbance is normally distributed or if the sample size is large. 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 is zero, so that the corresponding variable contributes nothing to the regression and can therefore be dropped. One way to test 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

⁴Sometimes the population coefficient on C is called the constant term, and the regression estimate is called the estimated constant term.

by checking whether the t statistic is greater than two in absolute value.⁵

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 absolute value as the one actually obtained, assuming that the irrelevance hypothesis true. Hence if a t statistic were two, the corresponding probability value would be approximately .05. 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 distribution. 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 .867, so that conditional on zero education and experience, our best forecast of the log wage would be 86.7 cents. Moreover, the intercept is very precisely estimated, as evidenced by the small standard error of .08 relative to the estimated coefficient. An approximate 95% confidence interval for the true but unknown population intercept is $.867 \pm 2(.08)$, or [.71, 1.03]. 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 EDUC is .093, 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. The coefficient is positive, so that LWAGE tends to rise when EDUC rises. In fact, the interpretation of the estimated coefficient of .09 is that, holding everything else constant, a one-

⁵If the sample size is small, or if we want a significance level other than 5%, we must refer to a table of critical values of the t distribution. We also note that use of the t distribution in small samples also requires an assumption of normally distributed disturbances.

year increase in EDUC will produce a .093 increase in LWAGE.

The estimated coefficient on EXPER is .013. Its standard error is also small, and hence its t statistic is large, with a very small probability value. Hence we reject the hypothesis that EXPER contributes nothing to the forecasting regression. A one-year increase in *EXPER* tends to produce a .013 increase in LWAGE.

A variety of diagnostic statistics follow; they help us to evaluate the adequacy of the regression. We provide detailed discussions of many of them elsewhere. Here we introduce them very briefly:

3.4.1 Mean dependent var 2.342

The sample mean of the dependent variable is

$$\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i.$$

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

3.4.2 S.D. dependent var .561

The sample standard deviation of the dependent variable is

$$SD = \sqrt{\frac{\sum_{i=1}^N (y_i - \bar{y})^2}{N - 1}}.$$

It measures the dispersion, or scale, of y .

3.4.3 Sum squared resid 319.938

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

$$SSR = \sum_{i=1}^N e_i^2.$$

3.4.4 Log likelihood -938.236

The likelihood function is the joint density function of the data, viewed as a function of the model parameters. Hence a natural estimation strategy, called maximum likelihood estimation, is to find (and use as estimates) the parameter values that maximize the likelihood function. After all, by construction, those parameter values maximize the likelihood of obtaining the data that were actually obtained. In the leading case of normally-distributed regression disturbances, maximizing the likelihood function (or equivalently, the log likelihood function, because the log is a monotonic transformation) turns out to be equivalent to minimizing the sum of squared residuals, hence the maximum-likelihood parameter estimates are identical to the least-squares parameter estimates. The number reported is the maximized value of the log of the likelihood function.⁶ Like the sum of squared residuals, it's not of direct use, but it's useful for comparing models and testing hypotheses.

Let us now dig a bit more deeply into the likelihood function, maximum-likelihood estimation, and related hypothesis-testing procedures. A natural estimation strategy with wonderful asymptotic properties, called maximum likelihood estimation, is to find (and use as estimates) the parameter values that maximize the likelihood function. After all, by construction, those parameter values maximize the likelihood of obtaining the data that were actually obtained.

In the leading case of normally-distributed regression disturbances, maximizing the likelihood function turns out to be equivalent to minimizing the

⁶Throughout this book, “log” refers to a natural (base e) logarithm.

sum of squared residuals, hence the maximum-likelihood parameter estimates are identical to the least-squares parameter estimates.

To see why maximizing the Gaussian log likelihood gives the same parameter estimate as minimizing the sum of squared residuals, let us derive the likelihood for the Gaussian linear regression model with non-stochastic regressors,

$$\begin{aligned} y_i &= x'_i \beta + \varepsilon \\ \varepsilon_i &\sim iidN(0, \sigma^2). \end{aligned}$$

The model implies that

$$y_i \sim iidN(x'_i \beta, \sigma^2),$$

so that

$$f(y_i) = (2\pi\sigma^2)^{-\frac{1}{2}} e^{\frac{-1}{2\sigma^2}(y_i - x'_i \beta)^2}.$$

Hence $f(y_1, \dots, y_N) = f(y_1)f(y_2) \cdots f(y_N)$ (by independence of the y_i 's). In particular,

$$L = \prod_{i=1}^N (2\pi\sigma^2)^{-\frac{1}{2}} e^{\frac{-1}{2\sigma^2}(y_i - x'_i \beta)^2}$$

so

$$\begin{aligned} \ln L &= \ln \left((2\pi\sigma^2)^{-\frac{N}{2}} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - x'_i \beta)^2 \\ &= \frac{-N}{2} \ln(2\pi) - \frac{N}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - x'_i \beta)^2. \end{aligned}$$

Note in particular that the β vector that maximizes the likelihood (or log likelihood – the optimizers must be identical because the log is a positive monotonic transformation) is the β vector that minimizes the sum of squared residuals.

The log likelihood is also useful for hypothesis testing via likelihood-ratio

tests. Under very general conditions we have asymptotically that:

$$-2(\ln L_0 - \ln L_1) \sim \chi_d^2,$$

where $\ln L_0$ is the maximized log likelihood under the restrictions implied by the null hypothesis, $\ln L_1$ is the unrestricted log likelihood, and d is the number of restrictions imposed under the null hypothesis.

t and F tests are likelihood ratio tests under a normality assumption. That's why they can be written in terms of minimized SSR 's rather than maximized $\ln L$'s.

3.4.5 F statistic 199.626

We use the F statistic to test the hypothesis that the coefficients of all variables in the regression except the intercept are jointly zero.⁷ That is, we test whether, taken jointly as a set, the variables included in the forecasting model have any explanatory value. This contrasts with the t statistics, which we use to examine the explanatory value of the variables one at a time.⁸ If no variable has explanatory value, the F statistic follows an F distribution with $k - 1$ and $T - k$ degrees of freedom. The formula is

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

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 the 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 has explanatory content.

⁷We 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. See Problem 6.

⁸In the degenerate case of only one RHS variable, the t and F statistics contain exactly the same information, and $F = t^2$. When there are two or more RHS variables, however, the hypotheses tested differ, and $F \neq t^2$.

3.4.6 Prob(F statistic) 0.000000

The probability value for the F statistic gives the significance level at which we can just reject the hypothesis that the set of RHS variables has no predictive value. Here, the value is indistinguishable from zero, so we reject the hypothesis overwhelmingly.

3.4.7 S.E. of regression .492

If we knew the elements of β and predicted y_i using $x'_i\beta$, then our prediction errors would be the ε_i 's, with variance σ^2 . We'd like an estimate of σ^2 , because it tells us whether our prediction errors are likely to be large or small. The observed residuals, the e_i 's, are effectively estimates of the unobserved population disturbances, the ε_i 's. Thus the sample variance of the e 's, which we denote s^2 (read “ s -squared”), is a natural estimator of σ^2 :

$$s^2 = \frac{\sum_{i=1}^N e_i^2}{N - K}.$$

s^2 is an estimate of the dispersion of the regression disturbance and hence is used to assess goodness of fit of the model, as well as the magnitude of prediction errors that we're likely to make. The larger is s^2 , the worse the model's fit, and the larger the prediction errors we're likely to make. s^2 involves a degrees-of-freedom correction (division by $N - K$ rather than by $N - 1$, reflecting the fact that K regression coefficients have been estimated), which is an attempt to get a good estimate of the out-of-sample prediction error variance on the basis of the in-sample residuals.

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

$$SER = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^N e_i^2}{N - K}}.$$

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.

3.4.8 R -squared .232

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 percent of the variance of y explained by the variables included in the regression. R^2 measures the in-sample success of the regression equation in predicting y ; hence it is widely used as a quick check of goodness of fit, or predictability, of y based on the variables included in the regression. Here the R^2 is about 23% – well above zero but not great. The formula is

$$R^2 = 1 - \frac{\sum_{i=1}^N e_i^2}{\sum_{i=1}^N (y_i - \bar{y})^2}.$$

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

$$R^2 = 1 - \frac{\frac{1}{N} \sum_{i=1}^N e_i^2}{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2},$$

which makes clear that the numerator in the large fraction is very close to s^2 , and the denominator is very close to the sample variance of y .

3.4.9 Adjusted R -squared .231

The interpretation is the same as that of R^2 , but the formula is a bit different. Adjusted R^2 incorporates adjustments for degrees of freedom used in fitting the model, in an attempt to offset the inflated appearance of good fit if many RHS variables are tried and the “best model” selected. Hence adjusted R^2

is a more trustworthy goodness-of-fit measure than R^2 . As long as there is more than one RHS variable in the model fitted, adjusted R^2 is smaller than R^2 ; here, however, the two are extremely close (23.1% vs. 23.2%). Adjusted R^2 is often denoted \bar{R}^2 ; the formula is

$$\bar{R}^2 = 1 - \frac{\frac{1}{N-K} \sum_{i=1}^N e_i^2}{\frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2},$$

where K is the number of RHS variables, including the constant term. Here the numerator in the large fraction is precisely s^2 , and the denominator is precisely the sample variance of y .

3.4.10 Akaike info criterion 1.423

The Akaike information criterion, or AIC , is effectively an estimate of the out-of-sample forecast error variance, as is s^2 , but it penalizes degrees of freedom more harshly. It is used to select among competing models. The formula is:

$$AIC = e^{(\frac{2K}{N}) \frac{\sum_{i=1}^N e_i^2}{N}},$$

and “smaller is better”. That is, we select the model with smallest AIC . We will discuss AIC in greater depth in Chapter 4.

3.4.11 Schwarz criterion 1.435

The Schwarz information criterion, or SIC , is an alternative to the AIC with the same interpretation, but a still harsher degrees-of-freedom penalty. The formula is:

$$SIC = N^{(\frac{K}{N})} \frac{\sum_{i=1}^N e_i^2}{N},$$

and “smaller is better”. That is, we select the model with smallest SIC . We will discuss SIC in greater depth in Chapter 4.

3.4.12 A Bit More on AIC and SIC

The *AIC* and *SIC* are tremendously important for guiding model selection in a ways that avoid **data mining** and **in-sample overfitting**.

You will want to start using *AIC* and *SIC* immediately, so we provide a bit more information here. Model selection by maximizing R^2 , or equivalently minimizing residual SSR , is ill-advised, because they don't penalize for degrees of freedom and therefore tend to prefer models that are “too big.” Model selection by maximizing \bar{R}^2 , or equivalently minimizing residual s^2 , is still ill-advised, even though \bar{R}^2 and s^2 penalize somewhat for degrees of freedom, because they don't penalize harshly enough and therefore still tend to prefer models that are too big. In contrast, *AIC* and *SIC* get things just right. *SIC* has a wonderful asymptotic optimality property when the set of candidate models is viewed as fixed: Basically *SIC* “gets it right” asymptotically, selecting either the *DGP* (if the *DGP* is among the models considered) or the best predictive approximation to the *DGP* (if the *DGP* is not among the models considered). *AIC* has a different and also-wonderful asymptotic optimality property, known as “efficiency,” when the set of candidate models is viewed as expanding as the sample size grows. In practice, the models selected by *AIC* and *SIC* rarely disagree.

3.4.13 Hannan-Quinn criter. 1.427

Hannan-Quinn is yet another information criterion for use in model selection. We will not use it in this book.

3.4.14 Durbin-Watson stat. 1.926

The Durbin-Watson (DW) statistic is used in time-series contexts, and we will study it later.

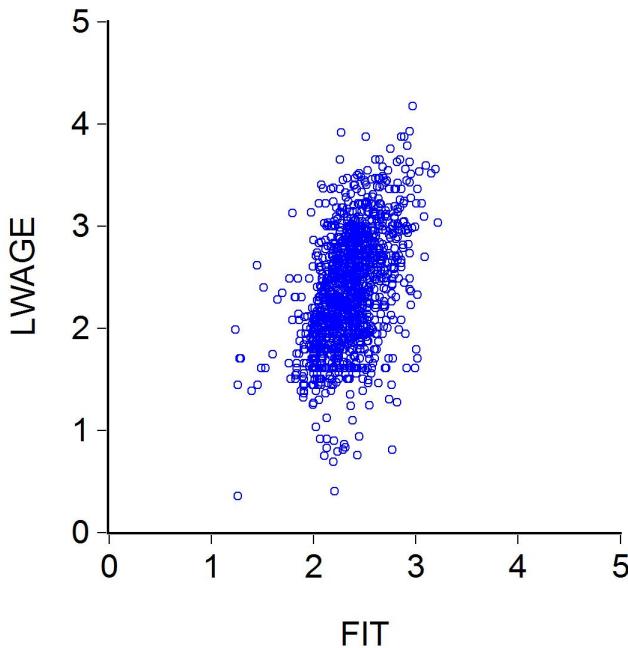


Figure 3.5: Wage Regression Residual Scatter

3.4.15 The Residual Scatter

The residual scatter is often useful in both cross-section and time-series situations. It is a plot of y vs \hat{y} . A perfect fit ($R^2 = 1$) corresponds to all points on the 45 degree line, and no fit ($R^2 = 0$) corresponds to all points on a vertical line corresponding to $y = \bar{y}$.

In Figure 3.5 we show the residual scatter for the wage regression. It is not a vertical line, but certainly also not the 45 degree line, corresponding to the positive but relatively low R^2 of .23.

3.4.16 The Residual Plot

In time-series settings, it's always a good idea to assess visually the adequacy of the model via time series plots of the actual data (y_i 's), the fitted values (\hat{y}_i 's), and the residuals (e_i 's). Often we'll refer to such plots, shown together

in a single graph, as a residual plot.⁹ We'll make use of residual plots throughout this book. Note that even with many RHS variables in the regression model, both the actual and fitted values of y , and hence the residuals, are simple univariate series that can be plotted easily.

The reason we examine the residual plot is that patterns would indicate violation of our *iid* assumption. In time series situations, we are particularly interested in inspecting the residual plot for evidence of serial correlation in the e_i 's, which would indicate failure of the assumption of *iid* regression disturbances. More generally, residual plots can also help assess the overall performance of a model by flagging anomalous residuals, due for example to outliers, neglected variables, or structural breaks.

Our wage regression is cross-sectional, so there is no natural ordering of the observations, and the residual plot is of limited value. But we can still use it, for example, to check for outliers.

In Figure 3.6, we show the residual plot for the regression of LWAGE on EDUC and EXPER. The actual and fitted 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; their scale is on the left. It's important to note that the scales differ; the e_i 's are in fact substantially smaller and less variable than either the y_i 's or the \hat{y}_i 's. We draw the zero line through the residuals for visual comparison. No outliers are apparent.

⁹Sometimes, however, we'll use “residual plot” to refer to a plot of the residuals alone. The intended meaning should be clear from context.

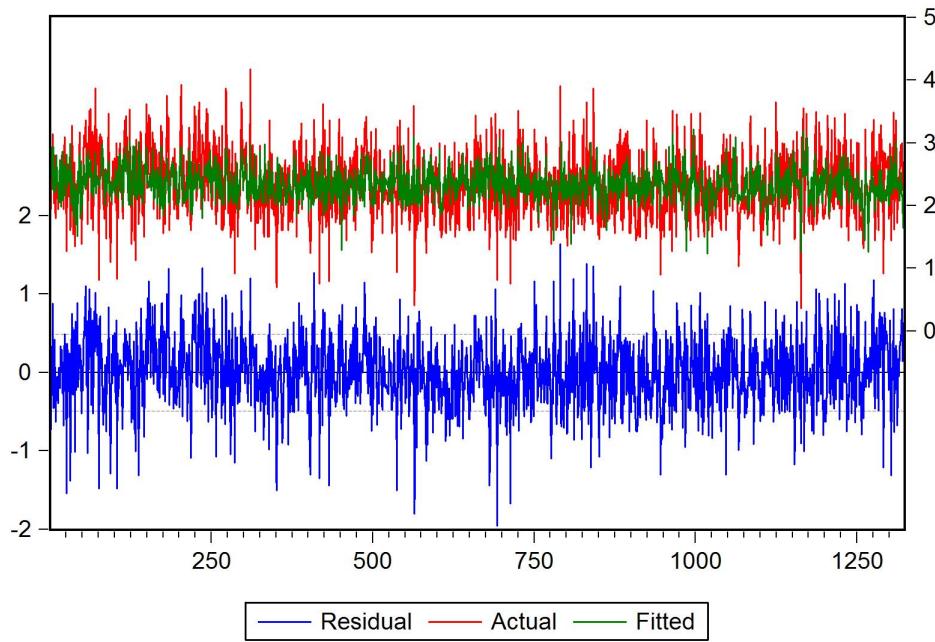


Figure 3.6: Wage Regression Residual Plot

3.5 Least Squares and Optimal Point Prediction

The linear regression DGP under the ideal conditions implies the conditional mean function,

$$E(y_i \mid x_{1i} = 1, x_{2i} = x_{2i}^*, \dots, x_{Ki} = x_{Ki}^*) = \beta_1 + \beta_2 x_{2i}^* + \dots + \beta_K x_{Ki}^*$$

$$\text{(or } E(y_i \mid x_i = x_i^*) = x_i^* \beta \text{)}.$$

And as also already noted much earlier in Chapter 1, a major goal in econometrics is predicting y . The question is “If a new person arrives with characteristics x^* , what is my minimum-MSE prediction of her y ? It turns out, very intuitively, that the answer is $E(y|x=x^*) = x_i^* \beta$. That is, “the conditional mean is the minimum MSE (point) predictor”. (Indeed if it were anything else you’d surely be suspicious.) The non-operational version (i.e., pretending that we know β) is $E(y_i|x_i=x_i^*) = x_i^* \beta$, and the operational

version (using $\hat{\beta}_{LS}$) is $E(\widehat{y_i|x_i = x_i^*}) = x_i^{*\prime} \hat{\beta}_{LS}$.

Let's now introduce a very basic and powerful result. Notice that the β 's in the conditional mean expression give the weights on the various x 's for forming the optimal predictor. Hence under the IC, consistency of OLS ensures that asymptotically the operational point prediction (based on $\hat{\beta}_{LS}$) will use the right weights (based on β). That is, under the IC, LS is consistent for the right predictive weights. Now here's the really amazing thing (although it's obvious when you think about it): *Under great generality, in particular even if the IC fail, LS is still consistent for the right predictive weights, simply by virtue of the MSE-optimization problem that it solves directly.* The bottom line: *Forecasting is of central importance in economics, and LS regression delivers optimal forecasts under great generality.*

If LS provides optimal forecasts even without the IC, you might wonder why we introduced the IC. There are two key sets of reasons. First, even for standard forecasting situations of the form “If a new person arrives with characteristics x^* , what is my minimum-MSE prediction of her y ,” once we drop the IC, so that the fitted model does not necessarily match the true DGP, there is a crucial issue of *what model to use*. Many questions arise. Which x 's should we include, and which should we exclude? Is a linear model really adequate, or should we incorporate some non-linearity? And so on. For any given model, LS will deliver the optimal parameter configuration for forecasting, but again, a crucial issue is what features a “good” or “the best” model should incorporate.

Second, what we've considered so far is called “non-causal” prediction. It exploits correlation between y and x to generate forecasts, but there is no presumption (or need) that x truly *causes* y in a deep scientific sense. (Remember, correlation does not imply causation!) But there is a causal form of prediction that differs from the one sketched thus far. In particular, thus far we've considered “If a new person arrives with characteristics x^* ,

what is my minimum-MSE prediction of her y ?”, but we might alternatively be interested in predicting the effects of an active *treatment*, or *intervention*, or *policy*, along the lines of “If I randomly select someone and *change* her characteristics in some way, what is my minimum-MSE prediction of the corresponding change in her y ?”. It turns out that LS does *not* always perform well for such “causal prediction” questions. So when *does* LS perform well for causal prediction? Under the IC! Effectively LS solves both the non-causal and causal prediction problems under the IC (or, put differently, the two problems are identical under the IC), but when the IC fail LS continues to solve the non-causal prediction problem but fails for the causal prediction problem.

Summarizing, here’s what true:

1. Non-causal prediction is important in economics
2. LS succeeds for non-causal prediction under great generality
3. Causal prediction is important in economics
4. LS fails for causal prediction unless the IC hold, so credible causal prediction is much harder.

Given the combination of 1 and 2 above, it makes obvious sense to start with non-causal prediction and treat it extensively, reserving 4 for separate treatment (which we do in Chapter 10). That has been the successful strategy of econometrics for many decades, and it is very much at the center of modern “data science” and “machine learning”.

3.6 Optimal Interval and Density Prediction

Prediction as introduced thus far is so-called point prediction (a single best – i.e., minimum MSE – guess).

Forecasts stated as confidence intervals (“interval forecasts”) are also of interest. The linear regression DGP under the IC implies the conditional variance function

$$\text{var}(y_i \mid x_i = x_i^*) = \sigma^2,$$

which we can use to form interval forecasts. The non-operational version is

$$y_i \in [x_i^{*\prime} \beta \pm 1.96 \sigma] \quad w.p. \ 0.95,$$

and the operational version is

$$y_i \in [x_i^{*\prime} \hat{\beta}_{LS} \pm 1.96 s] \quad w.p. \ 0.95.$$

Finally full density forecasts are of interest. The linear regression DGP under the IC implies the conditional density function

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

Hence a non-operational density forecast is

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

with operational version

$$y_i \mid x_i = x_i^* \sim N(x_i^{*\prime} \hat{\beta}_{LS}, s^2).$$

Notice that the interval and density forecasts rely for validity on more parts of the IC than do the point forecasts: Gaussian disturbances and constant disturbance variances – which makes clear in even more depth why violations of the IC are generally problematic even in non-causal forecasting situations.

3.7 Regression Output from a Predictive Perspective

In light of our predictive emphasis throughout this book, here we offer some predictive perspective on the regression statistics discussed earlier.

The sample, or historical, mean of the dependent variable, \bar{y} , an estimate of the *unconditional* mean of y , is a benchmark forecast. It is 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 \bar{y} .

The OLS fitted values, $\hat{y}_i = x_i' \hat{\beta}$, are effectively in-sample regression predictions.

The OLS residuals, $e_i = y_i - \hat{y}_i$, are effectively in-sample prediction errors corresponding to use of those in-sample regression predictions.

OLS coefficient signs and sizes relate to the weights put on the various x variables in forming the best in-sample prediction of y .

The standard errors, t statistics, and p -values let us do statistical inference as to which regressors are most relevant for predicting y .

SSR measures “total” in-sample accuracy of the regression predictions. It is closely related to in-sample MSE :

$$MSE = \frac{1}{N} SSR = \frac{1}{N} \sum_{i=1}^N e_i^2$$

(“average” in-sample accuracy of the regression predictions)

The F statistic effectively compares the accuracy of the regression-based forecast to that of the unconditional-mean forecast. It helps us assess whether the x variables, taken as a set, have predictive value for y . That contrasts with the t statistics, which assess predictive value of the x variables one at a time.

s^2 is just SSR scaled by $N - K$, so again, it’s a measure of the in-sample

accuracy of the regression-based forecast. It's like MSE, but corrected for degrees of freedom.

R^2 and \bar{R}^2 effectively compare the in-sample accuracy of conditional-mean ($x'_i \hat{\beta}$) and unconditional-mean (\bar{y}) forecasts. R^2 is not corrected for d.f. and has MSE on top:

$$R^2 = 1 - \frac{\frac{1}{N} \sum_{i=1}^N (y_i - x'_i \hat{\beta})^2}{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2}.$$

In contrast, \bar{R}^2 is corrected for d.f. and has s^2 on top:

$$\bar{R}^2 = 1 - \frac{\frac{1}{N-K} \sum_{i=1}^N (y_i - x'_i \hat{\beta})^2}{\frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2}.$$

Residual plots are useful for visually flagging neglected things that impact forecasting. Residual correlation (in time-series contexts) indicates that point forecasts could possibly be improved. Non-constant residual volatility indicates that interval and density forecasts could be possibly improved.

3.8 Multicollinearity

Collinearlty and **multicollinearity** don't really involve failure of the ideal conditions, but they nevertheless are sometimes issues and should be mentioned.

Collinearity refers to two x variables that are highly correlated. But even if all pairwise correlations are small an x variable could nevertheless be highly correlated with a *linear combination* of other x variables. That raises the idea of multicollinearity, where an x variable is highly correlated with a *linear combination* of other x variables. Collinearity is of course a special case of multicollinearity, so henceforth we will simply speak of multicollinearity.

3.8.1 Perfect and Imperfect Multicollinearity

There are two types of multicollinearity, perfect and imperfect.

Perfect multicollinearity refers to perfect correlation among some regressors, or linear combinations of regressors. Perfect multicollinearity is indeed a problem; the $X'X$ matrix is singular, so $(X'X)^{-1}$ does not exist, and the OLS estimator cannot even be computed!¹⁰ *Perfect* multicollinearity is disastrous, but it's unlikely to occur unless you do something really silly, like entering the same regressor twice.¹¹ In any event the solution is trivial: simply drop one of the redundant variables.

Imperfect multicollinearity, in contrast, occurs routinely but is not necessarily problematic, although in extreme cases it may require some attention. Imperfect collinearity/multicollinearity refers to (imperfect) correlation among some regressors, or linear combinations of regressors. Imperfect multicollinearity is not a “problem” in the sense that something was done incorrectly, and it is not a violation of the IC. Rather, it just reflects the nature of economic and financial data. But we still need to be aware of it and understand its effects. Telltale symptoms are large F and R^2 , yet small t 's (large s.e.'s), and/or coefficients that are sensitive to small changes in sample period. That is, OLS has trouble parsing individual influences, yet it's clear that there is an overall relationship. OLS is in some sense just what the doctor ordered – orthogonal projection.

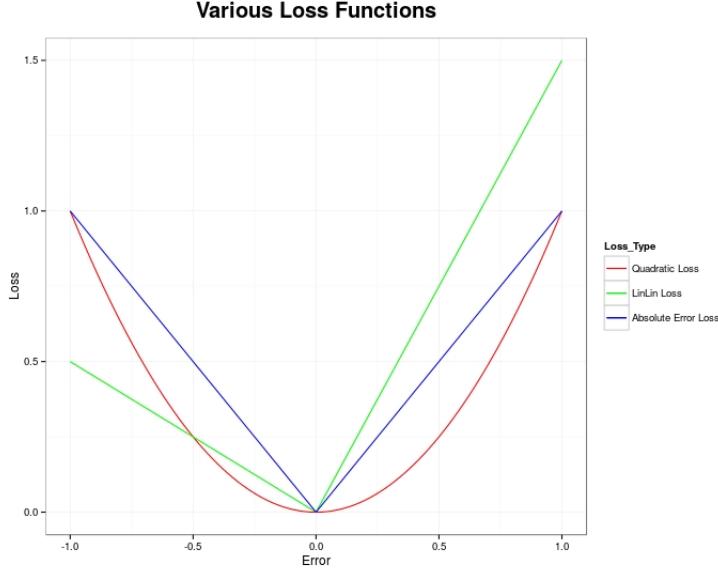
It can be shown, and it is very intuitive, that

$$\text{var}(\hat{\beta}_k) = f \left(\underbrace{\sigma^2}_{+}, \underbrace{\sigma_{x_k}^2}_{-}, \underbrace{R_k^2}_{+} \right)$$

where R_k^2 is the R^2 from a regression of x_k on all other regressors. In the

¹⁰For this reason people sometimes view non-singular $X'X$ as part of the IC.

¹¹A classic and more sophisticated example involves the “dummy variable trap,” in which we include as regressors a full set of dummy variables *and* an intercept. We will define dummy variables and note the dummy variable trap in Chapters 6 and 11.



limit, as $R_k^2 \rightarrow 1$, $\text{var}(\hat{\beta}_k) \rightarrow \infty$, because x_k is then perfectly “explained” by the other variables and is therefore completely redundant. R_k^2 is effectively a measure of the “strength” of the multicollinearity affecting β_k . We often measure the strength of multicollinearity by the “**variance inflation factor**”,

$$VIF(\hat{\beta}_k) = \frac{1}{1 - R_k^2},$$

which is just a transformation of R_k^2 .

3.9 Beyond Fitting the Conditional Mean: Quantile regression

Recall that the OLS estimator, $\hat{\beta}_{OLS}$, solves:

$$\min_{\beta} \sum_{i=1}^N (y_i - \beta_1 - \beta_2 x_{2t} - \dots - \beta_K x_{Kt})^2 = \min_{\beta} \sum_{i=1}^N \varepsilon_i^2$$

As you know, the solution has a simple analytic closed-form expression, $(X'X)^{-1}X'y$, with wonderful properties under the IC (unbiased, consistent,

Gaussian, MVUE). But other objectives are possible and sometimes useful. So-called quantile regression (QR) involves an objective function linear on each side of 0 but with (generally) unequal slopes. QR estimator $\hat{\beta}_{QR}$ minimizes “linlin loss,” or “check function loss”:

$$\min_{\beta} \sum_{i=1}^N \text{linlin}(\varepsilon_i),$$

where:

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

$I(x) = 1$ if x is true, and $I(x) = 0$ otherwise.

“ $I(\cdot)$ ” stands for “indicator” variable.

“linlin” refers to linearity on each side of the origin.

QR is not as simple as OLS, but it is still simple solves a linear programming problem).

A key issue is what, precisely, quantile regression fits. QR fits the $d \cdot 100\%$ quantile:

$$\text{quantile}_d(y|X) = x\beta$$

where

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

This is an important generalization of regression (e.g., How do the wages of people in the far left tail of the wage distribution vary with education and experience, and how does that compare to those in the center of the wage distribution?)

3.10 Exercises, Problems and Complements

1. (Regression with and without a constant term)

Consider Figure 3.3, in which we showed a scatterplot of y vs. x with a fitted regression line superimposed.

- a. In fitting that regression line, we included a constant term. How can you tell?
 - b. Suppose that we had not included a constant term. How would the figure look?
 - c. We almost always include a constant term when estimating regressions. Why?
 - d. When, if ever, might you explicitly want to exclude the constant term?
2. (Interpreting coefficients and variables)

Let $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$, where y_i is the number of hot dogs sold at an amusement park on a given day, x_i is the number of admission tickets sold that day, z_i is the daily maximum temperature, and ε_i is a random error. Assume the IC.

- a. State whether each of y_i , x_i , z_i , β_1 , β_2 and β_3 is a coefficient or a variable.
- b. Determine the units of β_1 , β_2 and β_3 , and describe the physical meaning of each.
- c. What do the signs of the coefficients tell you about how the various variables affects the number of hot dogs sold? What are your expectations for the signs of the various coefficients (negative, zero, positive or unsure)?
- d. Is it sensible to entertain the possibility of a non-zero intercept (i.e., $\beta_1 \neq 0$)? $\beta_2 > 0$? $\beta_3 < 0$?

3. (Scatter plots and regression lines)

Draw qualitative scatter plots and regression lines for each of the following two-variable datasets, and state the R^2 in each case:

- a. Data set 1: y and x have correlation 1
- b. Data set 2: y and x have correlation -1
- c. Data set 3: y and x have correlation 0.

4. (Desired values of regression diagnostic statistics)

For each of the diagnostic statistics listed below, indicate whether, other things the same, “bigger is better,” “smaller is better,” or neither. Explain your reasoning. (Hint: Be careful, think before you answer, and be sure to qualify your answers as appropriate.)

- a. Coefficient
- b. Standard error
- c. t statistic
- d. Probability value of the t statistic
- e. R -squared
- f. Adjusted R -squared
- g. Standard error of the regression
- h. Sum of squared residuals
- i. Log likelihood
- j. Mean of the dependent variable
- k. Standard deviation of the dependent variable
- l. Akaike information criterion
- m. Schwarz information criterion

- n. F statistic
 - o. Probability-value of the F statistic
5. (Regression semantics)
- Regression analysis is so important, and used so often by so many people, that a variety of associated terms have evolved over the years, all of which are the same for our purposes. You may encounter them in your reading, so it's important to be aware of them. Some examples:
- a. Ordinary least squares, least squares, OLS, LS.
 - b. y , LHS variable, regressand, dependent variable, endogenous variable
 - c. x 's, RHS variables, regressors, independent variables, exogenous variables, predictors, covariates
 - d. probability value, prob-value, p -value, marginal significance level
 - e. Schwarz criterion, Schwarz information criterion, SIC , Bayes information criterion, BIC
6. (Regression when X Contains Only an Intercept)
- Consider the regression model (3.1)-(3.2), but where X contains only an intercept.
- a. What is the OLS estimator of the intercept?
 - b. What is the distribution of the OLS estimator under the ideal conditions?
 - c. Does the variance-covariance matrix of the OLS estimator under the ideal conditions depend on any unknown parameters, and if so, how would you estimate them?
7. (Dimensionality)

We have emphasized, particularly in Chapter 2, that graphics is a powerful tool with a variety of uses in the construction and evaluation of econometric models. We hasten to add, however, that graphics has its limitations. In particular, graphics loses much of its power as the dimension of the data grows. If we have data in ten dimensions, and we try to squash it into two or three dimensions to make graphs, there's bound to be some information loss.

Thus, in contrast to the analysis of data in two or three dimensions, in which case learning about data by fitting models involves a loss of information whereas graphical analysis does not, graphical methods lose their comparative advantage in higher dimensions. In higher dimensions, graphical analysis can become comparatively laborious and less insightful.

8. (Wage regressions)

The relationship among wages and their determinants is one of the most important in all of economics. In the text we have examined, and will continue to examine, the relationship for 1995 using a CPS subsample. Here you will thoroughly analyze the relationship for 2004 and 2012, compare your results to those for 1995, and think hard about the meaning and legitimacy of your results.

- (a) Obtain the relevant 1995, 2004 and 2012 CPS subsamples.
- (b) Discuss any differences in the datasets. Are the same people in each dataset?
- (c) For now, assume the validity of the ideal conditions. Using each dataset, run the OLS regression $WAGE \rightarrow c, EDUC, EXPER$. (Note that the LHS variable is $WAGE$, not $LWAGE$.) Discuss and compare the results in detail.

- (d) Now think of as many reasons as possible to be *skeptical* of your results. (This largely means think of as many reasons as possible why the IC might fail.) Which of the IC might fail? One? A few? All? Why? Insofar as possible, discuss the IC, one-by-one, how/why failure could happen here, the implications of failure, how you might detect failure, what you might do if failure is detected, etc.
 - (e) Repeat all of the above using *LWAGE* as the LHS variable.
9. (Parallels between the sampling distribution of the sample mean under simple random sampling, and the sampling distribution of the OLS estimator under the IC)

Consider first the sample mean under Gaussian simple random sampling.

- (a) What *is* a Gaussian simple random sample?
- (b) What *is* the sample mean, and what finite-sample properties does it have under Gaussian simple random sampling?
- (c) Display and discuss the exact distribution of the sample mean.
- (d) How would you estimate and plot the exact distribution of the sample mean?

Now consider the OLS regression estimator under the IC.

- (a) What *are* the IC?
- (b) What is the OLS estimator, and what finite-sample properties does it enjoy?
- (c) Display and discuss the exact distribution of the OLS estimator.

Under what conditions, if any, do your “sample mean answers” and “OLS answers” precisely coincide?

10. (OLS regression residuals sum to zero)

Assertion: As long as an intercept is included in a linear regression, the OLS residuals must sum to precisely zero. The intuition is simply that non-zero residual mean (residual “constant term”) would automatically be pulled into the residual constant term, hence guaranteeing a zero residual mean.

- (a) Prove the assertion precisely.
- (b) Evaluate the claim that the assertion implies the regression fits perfectly “on average,” despite the fact that it fits imperfectly point-by-point.

11. (Simulation algorithm for density prediction)

- (a) Take R draws from $N(0, \hat{\sigma}^2)$.
- (b) Add $x^*'\hat{\beta}$ to each disturbance draw.
- (c) Form a density forecast by fitting a density to the output from step 11b.
- (d) Form an interval forecast (95%, say) by sorting the output from step 11b to get the empirical cdf, and taking the left and right interval endpoints as the the .025% and .975% values, respectively.

As $R \rightarrow \infty$, the algorithmic and analytic results coincide.

Note: This simulation algorithm may seem roundabout, but later we will drop normality.

12. (Quantile regression empirics)

For the 1995 CPS subsample (see EPC 8) re-do the regression $LWAGE \rightarrow c, EDUC, EXPER$ using 20%, 50% and 80% quantile regression instead of OLS regression.

3.11 Notes

Dozens of software packages implement linear regression analysis. Most automatically include an intercept in linear regressions unless explicitly instructed otherwise. That is, they automatically create and include a C variable.

The R command for ordinary least squares regression is “lm”. It’s already pre-loaded into R as the default package for estimating linear models. It uses standard R format for such models, where you specify formula, data, and various estimation options. It returns a model estimated by OLS including coefficients, residuals, and fitted values. You can also easily calculate summary statistics using the summary function.

The standard R quantile regression package is `quantreg`, written by [Roger Koenker](#), the inventor of quantile regression. The command “rq” functions similarly to “lm”. It takes as input a formula, data, the quantile to be estimated, and various estimation options.

3.12 Regression's Inventor: Carl Friedrich Gauss



Figure 3.7: Carl Friedrich Gauss

This is a photographic reproduction of a public-domain artwork, an oil painting of German mathematician and philosopher Carl Friedrich Gauss by G. Biermann (1824-1908). Date: 1887 (painting). Source Gau-Gesellschaft Göttingen e.V. (Foto: A. Wittmann).

Chapter 4

Misspecification and Model Selection

The IC's of Chapter 3 are surely heroic in economic contexts, so let us begin to relax them. One aspect of IC 1 is that the fitted model matches the true DGP. In reality we can never know the DGP, and surely any model that we might fit fails to match it, so there is an issue of how to select and fit a “good” model.

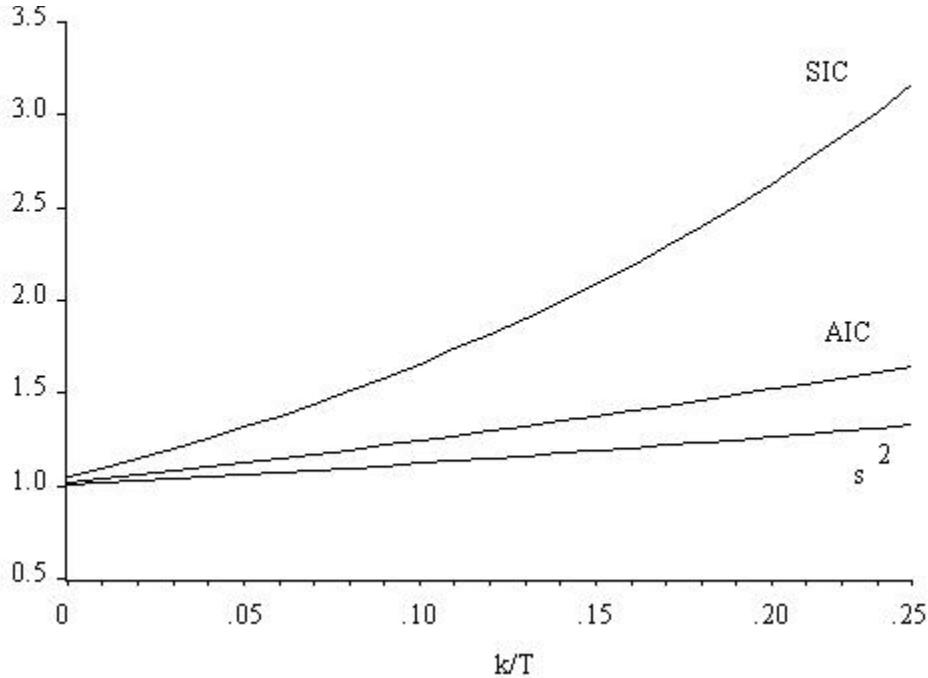
Recall that the **Akaike information criterion**, or *AIC*, is effectively an estimate of the out-of-sample forecast error variance, as is s^2 , but it penalizes degrees of freedom more harshly. It is used to select among competing forecasting models. The formula is:

$$AIC = e^{(\frac{2K}{N})} \frac{\sum_{i=1}^N e_i^2}{N}.$$

Also recall that the **Schwarz information criterion**, or *SIC*, is an alternative to the *AIC* with the same interpretation, but a still harsher degrees-of-freedom penalty. The formula is:

$$SIC = N^{(\frac{K}{N})} \frac{\sum_{i=1}^N e_i^2}{N}.$$

Here we elaborate. 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.

4.1 Information Criteria (Hard Thresholding)

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**,

$$MSE = \frac{\sum_{i=1}^N e_i^2}{N},$$

where N is the sample size and $e_i = y_i - \hat{y}_i$. 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 $1/N$ 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 ,

$$R^2 = 1 - \frac{\sum_{i=1}^N e_i^2}{\sum_{i=1}^N (y_i - \bar{y})^2} = 1 - \frac{MSE}{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2}.$$

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,

$$s^2 = \frac{\sum_{i=1}^N e_i^2}{N - K},$$

where K is the number of degrees of freedom used in model fitting.¹ 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 the standard error of the regression. s^2 is also intimately connected to the

¹The degrees of freedom used in model fitting is simply the number of parameters estimated.

R^2 adjusted for degrees of freedom (the “**adjusted R^2** ,” or \bar{R}^2). Recall that

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^N e_i^2 / (N - K)}{\sum_{i=1}^N (y_i - \bar{y})^2 / (N - 1)} = 1 - \frac{s^2}{\sum_{i=1}^N (y_i - \bar{y})^2 / (N - 1)}.$$

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 ,

$$s^2 = \left(\frac{N}{N - K} \right) \frac{\sum_{i=1}^N e_i^2}{N}.$$

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:

$$AIC = e^{\left(\frac{2K}{N}\right)} \frac{\sum_{i=1}^N e_i^2}{N}$$

and

$$SIC = N^{(\frac{K}{N})} \frac{\sum_{i=1}^N e_i^2}{N}.$$

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/N , the number of parameters estimated per sample observation, and we can compare the penalty factors graphically as K/N varies. In Figure *** we show the penalties as K/N moves from 0 to .25, for a sample size of $N = 100$. The s^2 penalty is small and rises slowly with K/N ; the AIC penalty is a bit larger and still rises only slowly with K/N . The SIC penalty, on the other hand, is substantially larger and rises much more quickly with K/N .

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 ?? 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.² More generally, *AIC* and *SIC* are written not in terms of minimized *SSR*'s, but rather in terms of maximized *lnL*'s. We have:

$$AIC = -2\ln L + 2K$$

and

$$SIC = -2\ln L + K\ln N.$$

These are useful for any model estimated by maximum likelihood, Gaussian or non-Gaussian.

4.2 Cross Validation (Hard Thresholding)

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

²Recall that in the Gaussian case *SSR* minimization and *lnL* maximization are equivalent.

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 N 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 “ $N - fold$ ” CV, because we split the data into N parts (the N individual observations) and predict each of them. More generally we can split the data into M parts ($M < N$) 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.

4.3 Stepwise Selection (Hard Thresholding)

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.

4.3.1 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.

4.3.2 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.

4.4 Bayesian Shrinkage (Soft Thresholding)

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**,³

$$\hat{\beta}_{ridge} = (X'X + \lambda I)^{-1} X'y.$$

$\lambda \rightarrow 0$ produces OLS, whereas $\lambda \rightarrow \infty$ shrinks completely to 0. λ can be chosen by CV. (Notice that λ can *not* be chosen by information criteria, as K regressors are included regardless of λ . Hence CV is a more general selection procedure, useful for selecting various “tuning parameters” (like λ) as opposed to just numbers of variables in hard-threshold procedures.

³The ridge regression estimator can be shown to be the posterior mean for a certain prior and likelihood.

4.5 Selection and Shrinkage (Mixed Hard and Soft Thresholding)

4.5.1 Penalized Estimation

Consider the penalized estimator,

$$\hat{\beta}_{PEN} = \operatorname{argmin}_{\beta} \left(\sum_{i=1}^N \left(y_i - \sum_i \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^K |\beta_i|^q \right),$$

or equivalently

$$\hat{\beta}_{PEN} = \operatorname{argmin}_{\beta} \sum_{i=1}^N \left(y_i - \sum_i \beta_i x_{it} \right)^2$$

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.

4.5.2 The Lasso

The lasso solves the L1-penalized regression problem of finding

$$\hat{\beta}_{LASSO} = \operatorname{argmin}_{\beta} \left(\sum_{i=1}^N \left(y_i - \sum_i \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^K |\beta_i| \right)$$

or equivalently

$$\hat{\beta}_{LASSO} = \operatorname{argmin}_{\beta} \sum_{i=1}^N \left(y_i - \sum_i \beta_i x_{it} \right)^2$$

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 β 's). This means that we can use info criteria to select among “lasso models” for various λ . That is, the lasso is another device for producing an “increasing” sequence of candidate models (as λ increases). The “best” λ can then be chosen by information criteria (or cross-validation, of course).

4.6 Distillation: Principal Components

4.6.1 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 $X'X$

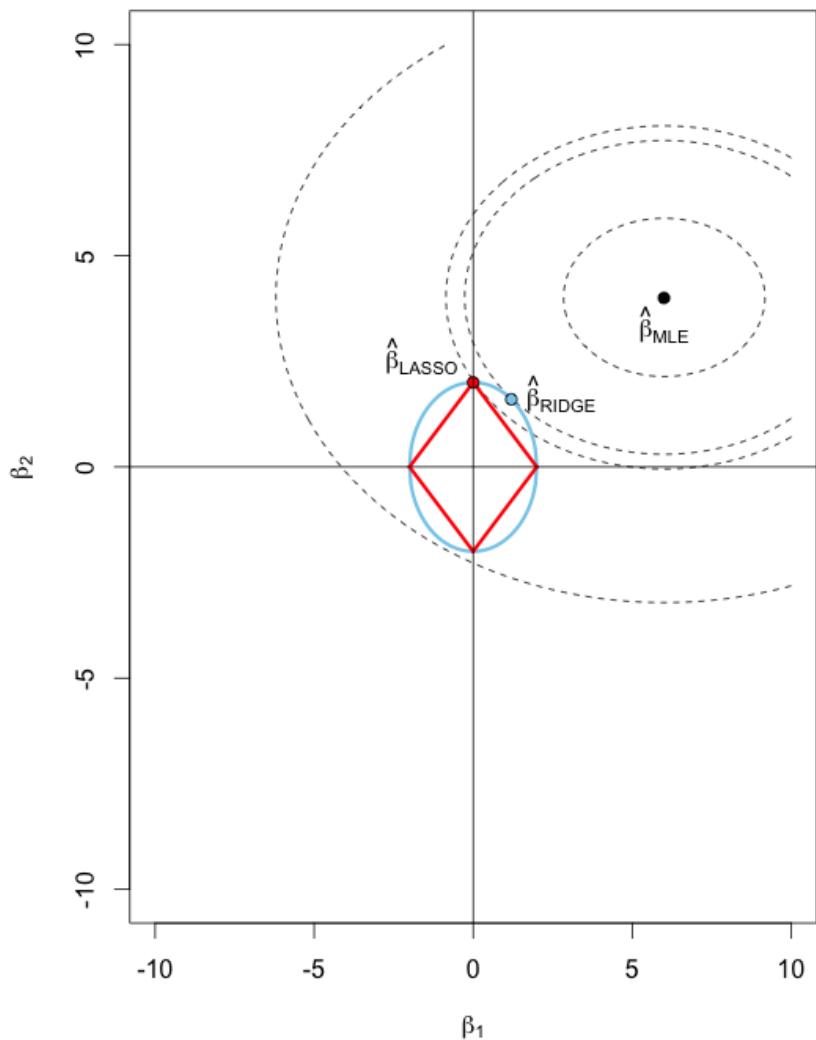


Figure 4.1: Lasso and Ridge Comparison

First principal component (PC):

$$z_1 = Xv_1$$

$$\text{var}(z_1) = d_1^2/N$$

(maximal sample variance among all possible l.c.'s of columns of X)

In general:

$$z_j = Xv_j \perp z_{j'}, j' \neq j$$

$$\text{var}(z_j) \leq d_j^2/N$$

4.6.2 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).

4.7 Exercises, Problems and Complements

1. (The sum of squared residuals, SSR)
 - (a) What is SSR and why is it reported?
 - (b) Do you agree with “bigger is better,” “smaller is better,” or neither?
Be careful.
 - (c) Describe in detail and discuss the use of regression statistics R^2 , \bar{R}^2 , F , SER , and SIC . What role does SSR play in each of the test statistics?
 - (d) Under the IC, is the maximized log likelihood related to the SSR ? If so, how? Would your answer change if we dropped normality?

2. (The variety of “information criteria” reported across software packages)

Some authors, and software packages, examine and report the logarithms of the AIC and SIC,

$$\ln(AIC) = \ln\left(\frac{\sum_{i=1}^N e_i^2}{N}\right) + \left(\frac{2K}{N}\right)$$

$$\ln(SIC) = \ln\left(\frac{\sum_{i=1}^N e_i^2}{N}\right) + \frac{K \ln(N)}{N}.$$

The practice is so common that $\log(AIC)$ and $\log(SIC)$ are often simply called the “AIC” and “SIC.” AIC and SIC must be greater than zero, so $\log(AIC)$ and $\log(SIC)$ are always well-defined and can take on any real value. The important insight, however, is that although these variations will of course change the numerical values of AIC and SIC produced by your computer, they will not change the *rankings* of models under the various criteria. Consider, for example, selecting among three models. If $AIC_1 < AIC_2 < AIC_3$, then it must be true as well that $\ln(AIC_1) < \ln(AIC_2) < \ln(AIC_3)$, so we would select model 1 regardless of the “definition” of the information criterion used.

3. (“All-subset”, “partial-subset”, and “one-shot” model selection)

Note that model selection by information criteria or cross validation are all-subset strategies, insofar as we examine all possible models and pick the one that looks best according to the criterion. Stepwise procedures are partial-subset strategies, insofar as we examine many models, but not all possible models, and pick the one that looks best according to the criterion. Ridge and LASSO, in contrast, are one-shot strategies, insofar as we need to perform only a single estimation. All-subset strategies become unappealing as the number of regressors, K , grows, because there are 2^K subsets of K regressors, requiring running and comparing

2^K regressions. One-shot strategies, in contrast, remain appealing in situations with many regressors, because just one estimation is required.

Chapter 5

Non-Normality

Here we consider another violation of the IC, non-normal disturbances.

Non-normality and **outliers**, which we introduce in this chapter, are closely related, because deviations from Gaussian behavior are often characterized by fatter tails than the Gaussian, which produce outliers. It is important to note that outliers are not necessarily “bad,” or requiring “treatment.” *Every* data set must have *some* most extreme observation, by definition! Statistical estimation efficiency, moreover, *increases* with data variability. The most extreme observations can be the most informative about the phenomena of interest. “Bad” outliers, in contrast, are those associated with things like data recording errors (e.g., you enter .753 when you mean to enter 75.3) or one-off events (e.g., a strike or natural disaster).

5.0.1 Results

To understand the properties of OLS without normality, it is helpful first to consider the properties of the sample mean without normality.

As reviewed in Appendix A, for a non-Gaussian simple random sample,

$$y_i \sim iid(\mu, \sigma^2), i = 1, \dots, N,$$

we have that the sample mean is consistent, asymptotically normal, and

asymptotically efficient, with

$$\bar{y} \xrightarrow{a} N\left(\mu, \frac{\sigma^2}{N}\right).$$

This result forms the basis for asymptotic inference. It is a Gaussian central limit theorem. We consistently estimate σ^2 using s^2 .

Now consider the linear regression under the IC except that we allow non-Gaussian disturbances. OLS remains consistent, asymptotically normal, and asymptotically efficient, with

$$\hat{\beta}_{OLS} \xrightarrow{a} N(\beta, V).$$

We consistently estimate the covariance matrix V using $s^2(X'X)^{-1}$.

5.1 Assessing Normality

There are many methods, ranging from graphics to formal tests.

5.1.1 QQ Plots

We introduced histograms earlier in Chapter 2 as a graphical device for learning about distributional shape. If, however, interest centers on the *tails* of distributions, **QQ plots** often provide sharper insight as to the agreement or divergence between the actual and reference distributions.

The QQ plot is simply a plot of the quantiles of the standardized data against the quantiles of a standardized reference distribution (e.g., normal). If the distributions match, the QQ plot is the 45 degree line. To the extent that the QQ plot does not match the 45 degree line, the nature of the divergence

can be very informative, as for example in indicating fat tails.

5.1.2 Residual Sample Skewness and Kurtosis

Recall skewness and kurtosis, which we reproduce here for convenience:

$$S = \frac{E(y - \mu)^3}{\sigma^3}$$

$$K = \frac{E(y - \mu)^4}{\sigma^4}.$$

Obviously, each tells about a different aspect of non-normality. Kurtosis, in particular, tells about fatness of distributional tails relative to the normal.

A simple strategy is to check various implications of residual normality, such as $S = 0$ and $K = 3$, via informal examination of \hat{S} and \hat{K} .

5.1.3 The Jarque-Bera Test

The **Jarque-Bera test** (JB) effectively aggregates the information in the data about both skewness and kurtosis to produce an overall test of the joint hypothesis that $S = 0$ and $K = 3$, based upon \hat{S} and \hat{K} . The test statistic is

$$JB = \frac{N}{6} \left(\hat{S}^2 + \frac{1}{4}(\hat{K} - 3)^2 \right).$$

Under the null hypothesis of independent normally-distributed observations ($S = 0$, $K = 3$), JB is distributed in large samples as a χ^2 random variable with two degrees of freedom.¹

¹We have discussed the case of an observed time series. If the series being tested for normality is the residual from a model, then N can be replaced with $N - K$, where K is the number of parameters estimated, although the distinction is inconsequential asymptotically.

5.2 Outliers

Outliers refer to big disturbances (in population) or residuals (in sample). Outliers may emerge for a variety of reasons, and they may require special attention because they can have substantial influence on the fitted regression line.

On the one hand, OLS retains its magic in such outlier situations – it is BLUE regardless of the disturbance distribution. On the other hand, the fully-optimal (MVUE) estimator may be highly non-linear, so the fact that OLS remains BLUE is less than fully comforting. Indeed OLS parameter estimates are particularly susceptible to distortions from outliers, because the quadratic least-squares objective *really* hates big errors (due to the squaring) and so goes out of its way to tilt the fitted surface in a way that minimizes them.

How to identify and treat outliers is a time-honored problem in data analysis, and there's no easy answer. If an outlier is simply a data-recording mistake, then it may well be best to discard it if you can't obtain the correct data. On the other hand, every dataset, even a perfectly “clean” dataset, has a “most extreme observation,” but it doesn't follow that it should be discarded. Indeed the most extreme observations are often the most informative – precise estimation requires data variation.

5.2.1 Outlier Detection

Graphics

One obvious way to identify outliers in bivariate regression situations is via graphics: one xy scatterplot can be worth a thousand words. In higher dimensions, the residual $\hat{y}y$ scatterplot remains invaluable, as does the residual plot of $y - \hat{y}$.

Leave-One-Out and Leverage

Another way to identify outliers is a “leave-one-out” coefficient plot, where we use the computer to sweep through the sample, leaving out successive observations, and examining differences in parameter estimates with observation various observations “in” vs. “out”. That is, in an obvious notation, we examine and plot $\hat{\beta}_{OLS}^{(-i)} - \hat{\beta}_{OLS}$, $i = 1, \dots, N$.

It can be shown, however, that the change in $\hat{\beta}_{OLS}$ is

$$\hat{\beta}_{OLS}^{(-i)} - \hat{\beta}_{OLS} = -\frac{1}{1-h_i}(X'X)^{-1}x'_i e_i,$$

where h_i is the i -th diagonal element of the “hat matrix,” $X(X'X)^{-1}X'$. Hence the estimated coefficient change $\hat{\beta}_{OLS}^{(-i)} - \hat{\beta}_{OLS}$ is driven by $\frac{1}{1-h_i}$. h_i is called the **observation- i leverage**. h_i can be shown to be in $[0, 1]$, so that the larger is h_i , the larger is $\hat{\beta}_{OLS}^{(-i)} - \hat{\beta}_{OLS}$. Hence one really just needs to examine the leverage sequence, and scrutinize carefully observations with high leverage.

5.3 Robust Estimation

Robust estimation provides a useful middle ground between completely discarding allegedly-outlying observations (“dumming them out”) and doing nothing. Here we introduce outlier-robust approaches to regression. The first involves OLS regression, but on weighted data, an the second involves switching from OLS to a different estimator.

5.3.1 Robustness Iteration

Fit at robustness iteration 0:

$$\hat{y}^{(0)} = X\hat{\beta}^{(0)}$$

where

$$\hat{\beta}^{(0)} = \operatorname{argmin} \left[\sum_{i=1}^N (y_i - x_i' \beta)^2 \right].$$

Robustness weight at iteration 1:

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

where

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

and $S(z)$ is a function such that $S(z) = 1$ for $z \in [-1, 1]$ but downweights outside that interval.

Fit at robustness iteration 1:

$$\hat{y}^{(1)} = X \hat{\beta}^{(1)}$$

where

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

Continue as desired.

5.3.2 Least Absolute Deviations

Recall that the OLS estimator solves

$$\min_{\beta} \sum_{i=1}^N (y_i - x_i' \beta)^2.$$

Now we simply change the objective to

$$\min_{\beta} \sum_{i=1}^N |y_i - x_i' \beta|.$$

or

$$\min_{\beta} \sum_{i=1}^N |\varepsilon_i|$$

That is, we change from squared-error loss to absolute-error loss. We call the new estimator “**least absolute deviations**” (LAD) and we write $\hat{\beta}_{LAD}$.² By construction, $\hat{\beta}_{LAD}$ is not influenced by outliers as much as $\hat{\beta}_{OLS}$. Put differently, LAD is more robust to outliers than is OLS.

Of course nothing is free, and the price of LAD is a bit of extra computational complexity relative to OLS. In particular, the LAD estimator does not have a tidy closed-form analytical expression like OLS, so we can’t just plug into a simple formula to obtain it. Instead we need to use the computer to find the optimal β directly. If that sounds complicated, rest assured that it’s largely trivial using modern numerical methods, as embedded in modern software.³

It is important to note that whereas OLS fits the conditional mean function:

$$\text{mean}(y|X) = X\beta,$$

LAD fits the conditional *median* function (50% quantile):

$$\text{median}(y|X) = X\beta$$

The conditional mean and median are equal under symmetry and hence under normality, but not under asymmetry, in which case the median is a better measure of central tendency. Hence LAD delivers two kinds of robustness to non-normality: it is robust to outliers and robust to asymmetry.

²Note that LAD regression is just quantile regression for $d = .50$.

³Indeed computation of the *LAD* estimator turns out to be a linear programming problem, which is well-studied and simple.

5.4 Wage Determination

Here we show some empirical results that make use of the ideas sketched above. There are many tables and figures appearing at the end of the chapter. We do not refer to them explicitly, but all will be clear upon examination.

5.4.1 *WAGE*

We run $WAGE \rightarrow c, EDUC, EXPER$. We show the regression results, the residual plot, the residual histogram and statistics, the residual Gaussian QQ plot, the leave-one-out plot, and the results of *LAD* estimation. The residual plot shows lots of positive outliers, and the residual histogram and Gaussian QQ plot indicate right-skewed residuals.

5.4.2 *LWAGE*

Now we run $LWAGE \rightarrow c, EDUC, EXPER$. Again we show the regression results, the residual plot, the residual histogram and statistics, the residual Gaussian QQ plot, the leave-one-out plot, and the results of *LAD* estimation. Among other things, and in sharp contrast to the results for *WAGE* and opposed to *LWAGE*, the residual histogram and Gaussian QQ plot indicate approximate residual normality.

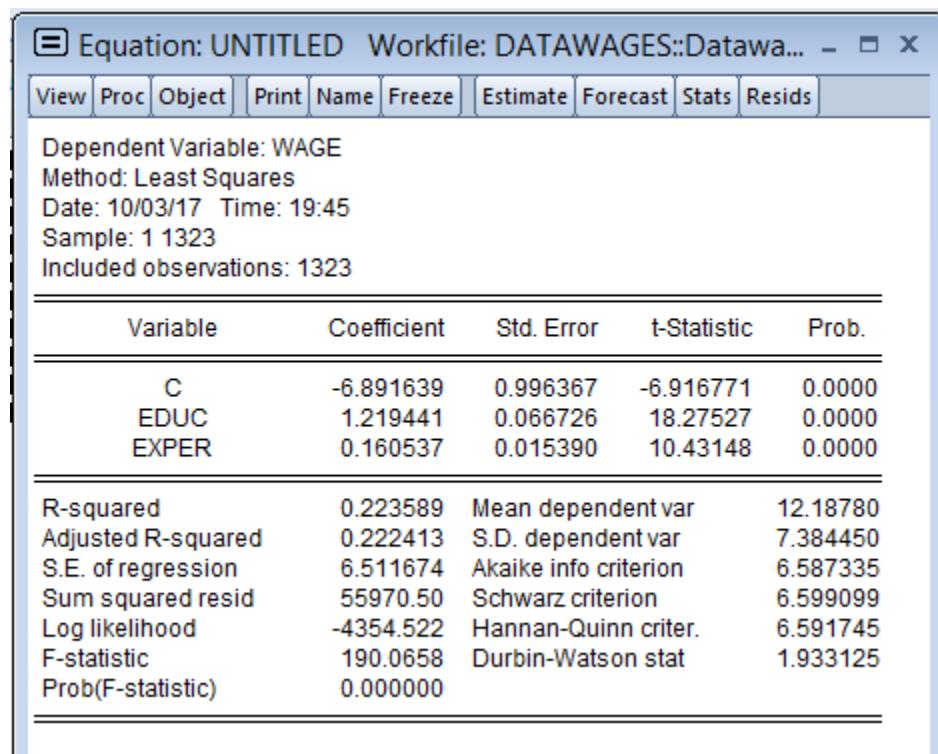


Figure 5.1: OLS Wage Regression

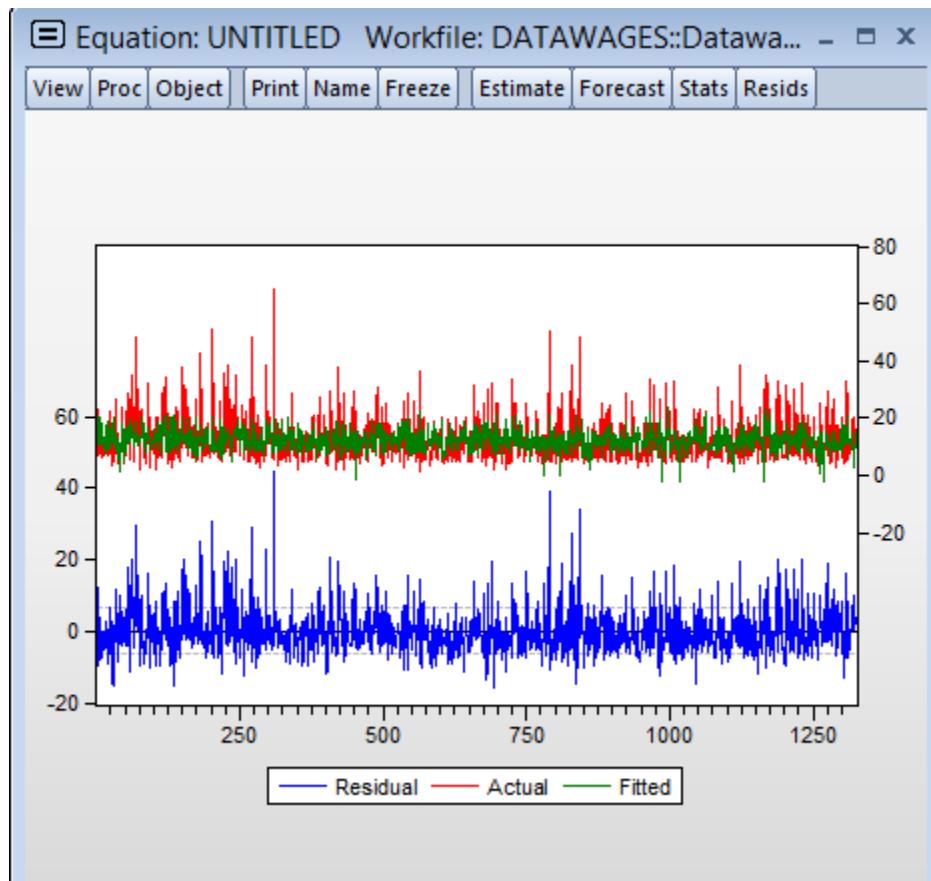


Figure 5.2: OLS Wage Regression: Residual Plot

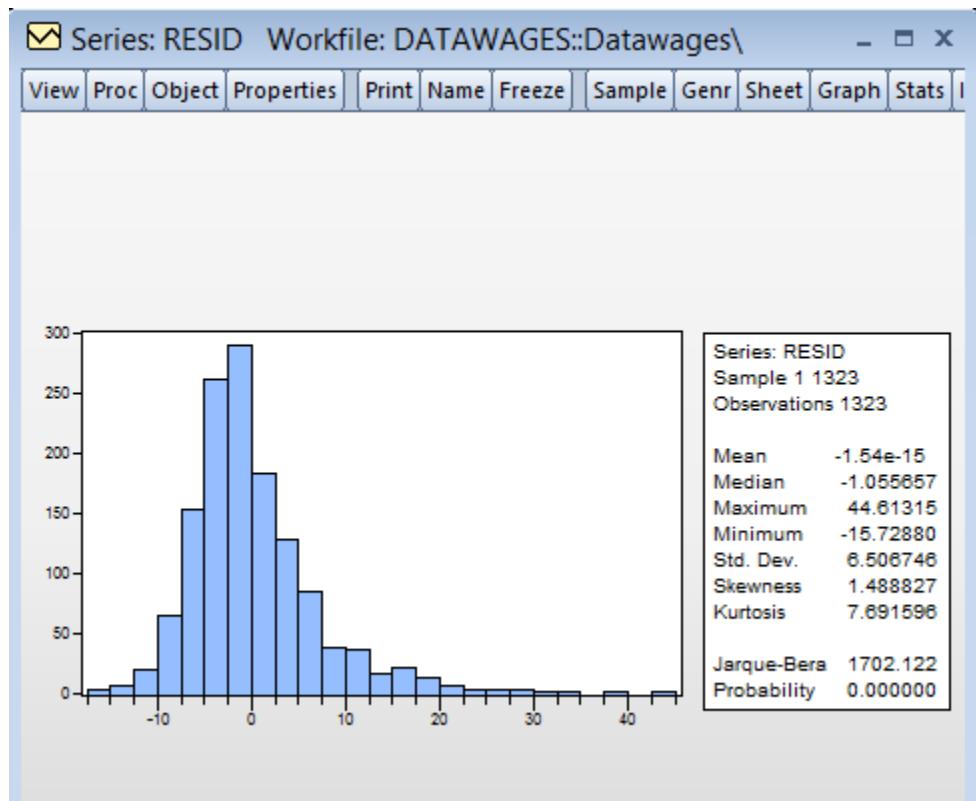


Figure 5.3: OLS Wage Regression: Residual Histogram and Statistics

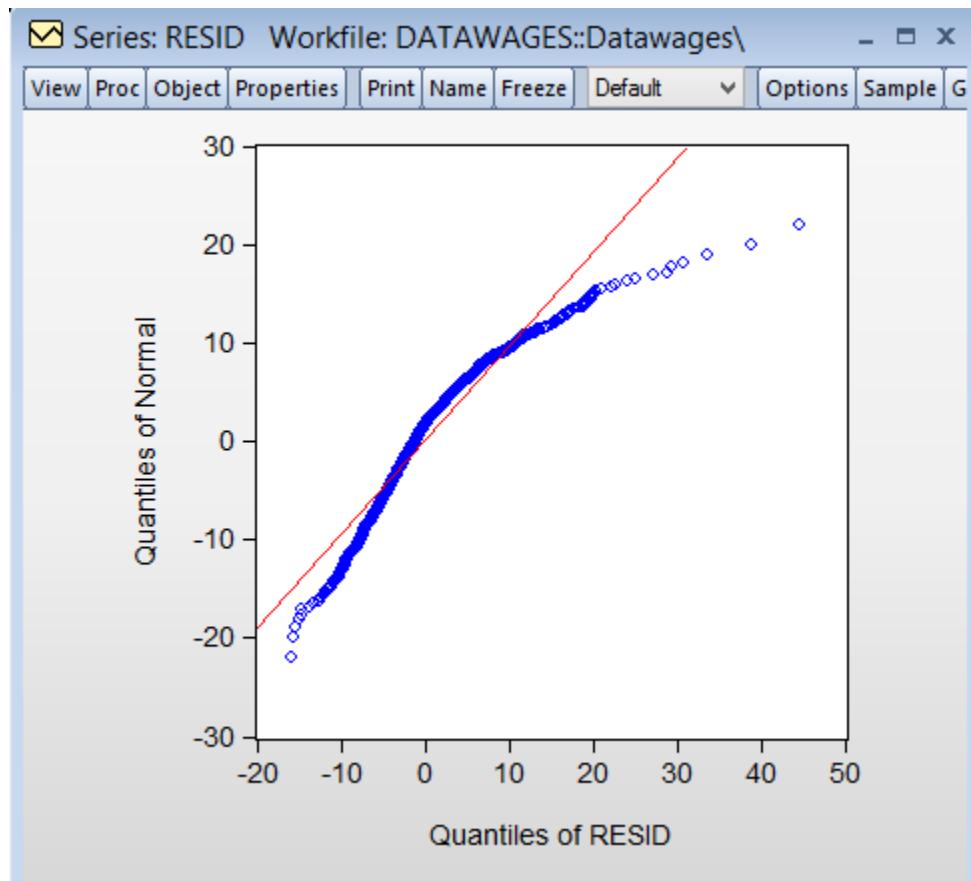


Figure 5.4: OLS Wage Regression: Residual Gaussian QQ Plot

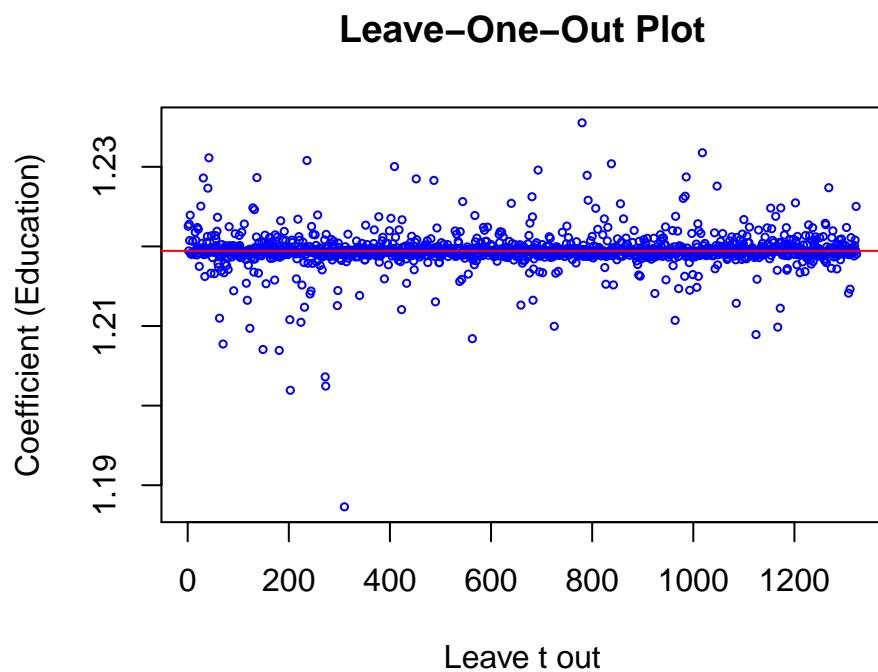


Figure 5.5: OLS Wage Regression: Leave-One-Out Plot

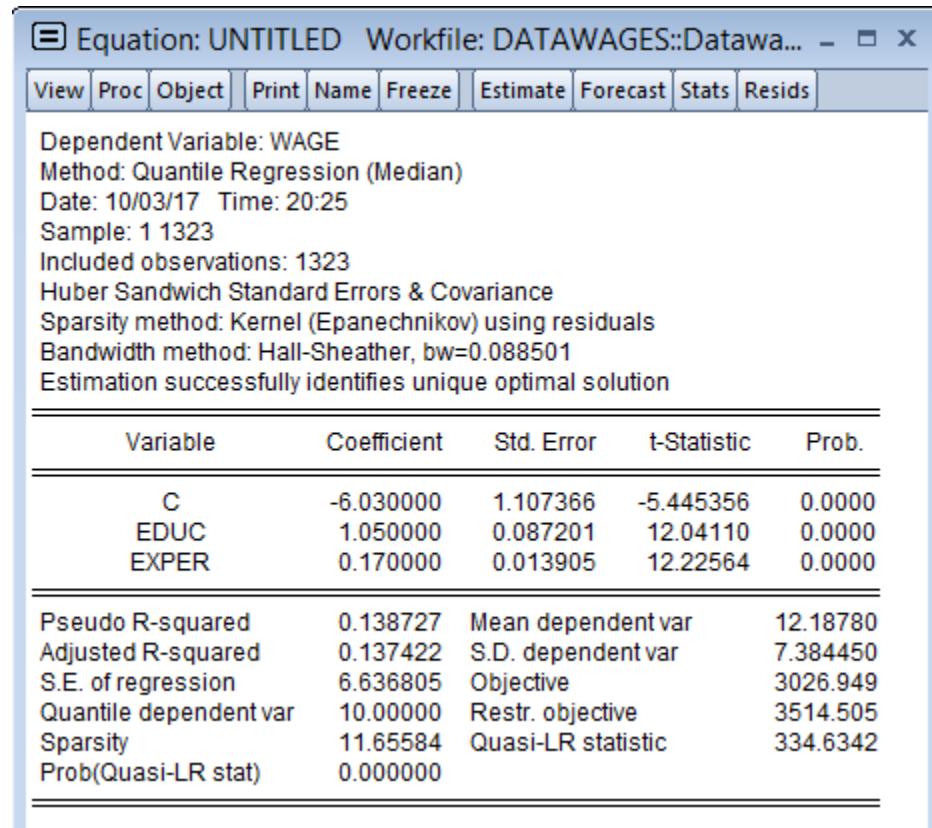


Figure 5.6: LAD Wage Regression

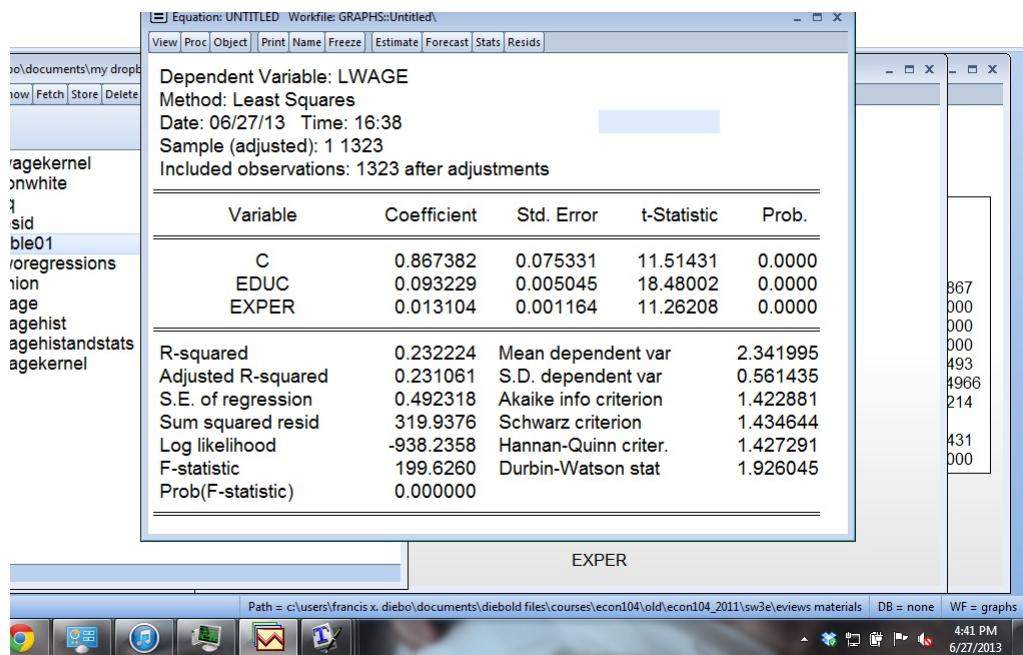


Figure 5.7: OLS Log Wage Regression

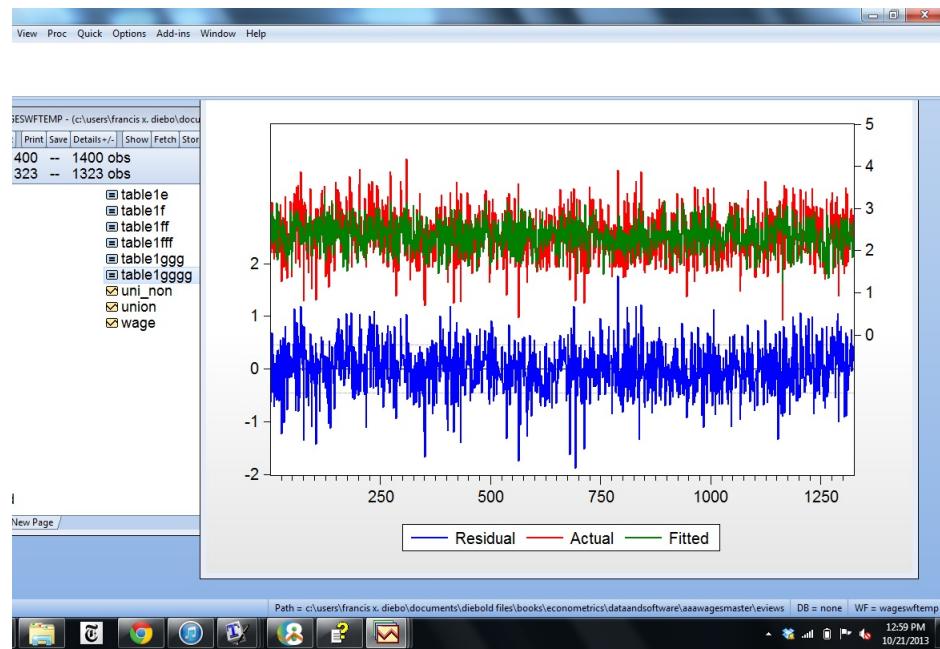


Figure 5.8: OLS Log Wage Regression: Residual Plot

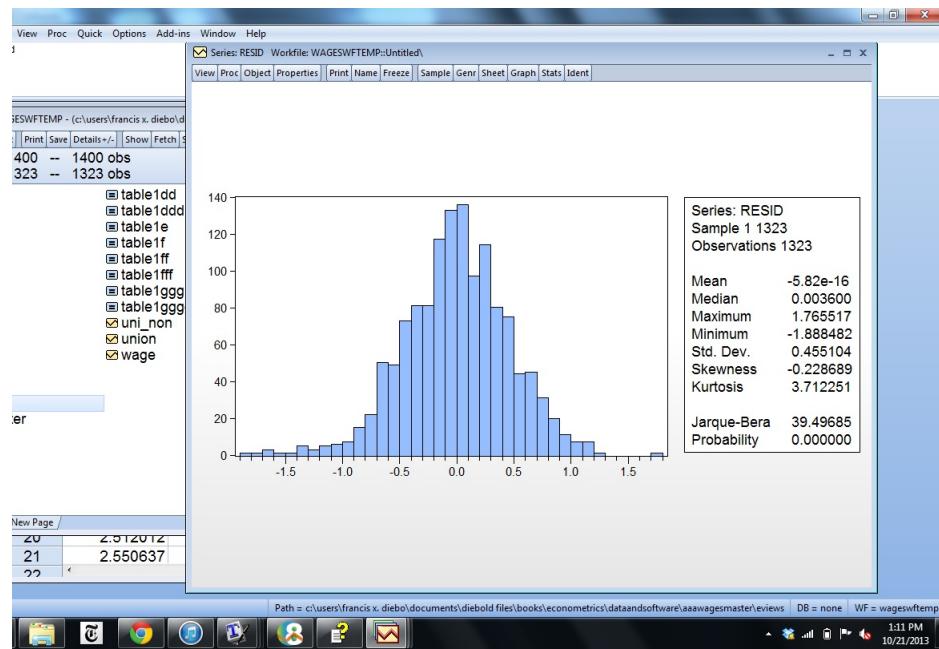


Figure 5.9: OLS Log Wage Regression: Residual Histogram and Statistics

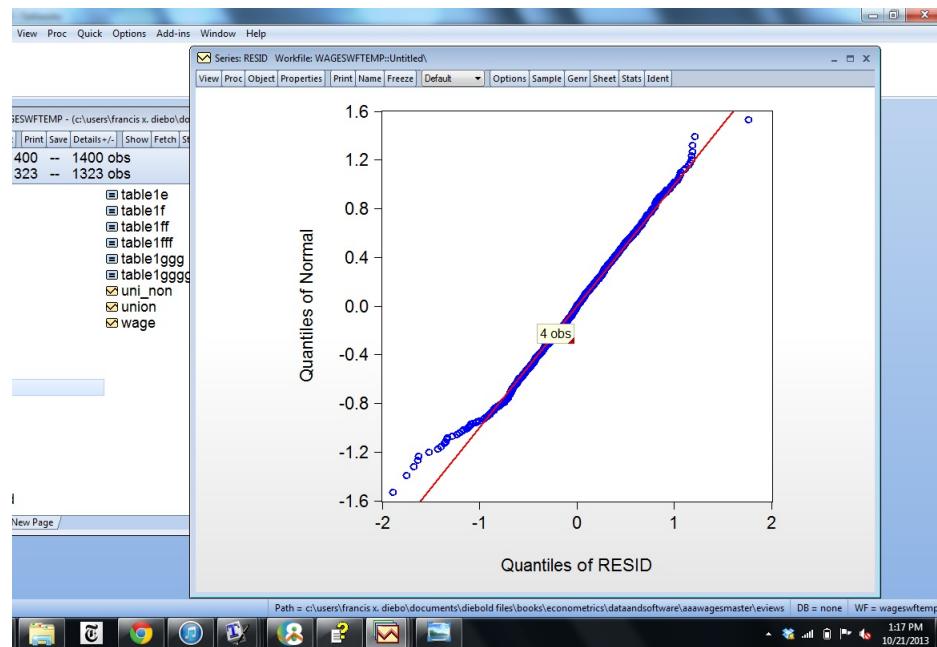


Figure 5.10: OLS Log Wage Regression: Residual Gaussian QQ Plot

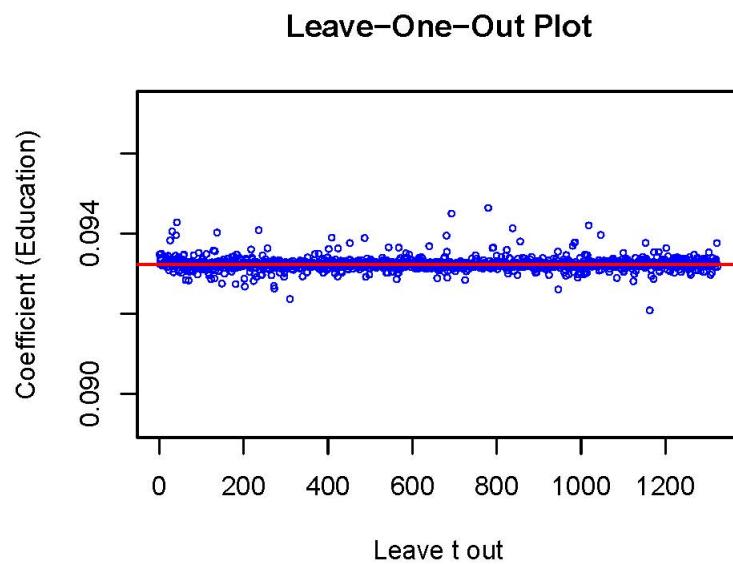


Figure 5.11: OLS Log Wage Regression: Leave-One-Out Plot

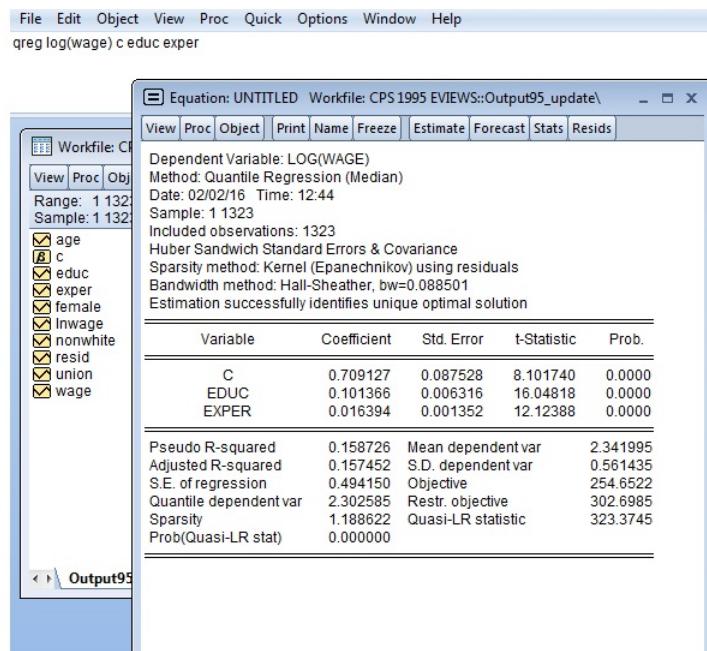


Figure 5.12: LAD Log Wage Regression

5.5 Exercises, Problems and Complements

1. (Taleb's *The Black Swan*)

Nassim Taleb is a financial markets trader turned pop author. His book, *The Black Swan* (Taleb (2007)), deals with many of the issues raised in this chapter. “Black swans” are seemingly impossible or very low-probability events – after all, swans are supposed to be *white* – that occur with annoying regularity in reality. Read his book. Where does your reaction fall on the spectrum from A to B below?

- A. Taleb offers crucial lessons for econometricians, heightening awareness in ways otherwise difficult to achieve. After reading Taleb, it’s hard to stop worrying about non-normality, model uncertainty, etc.
- B. Taleb belabors the obvious for hundreds of pages, arrogantly “informing” us that non-normality is prevalent, that all models are misspecified, and so on. Moreover, it takes a model to beat a model, and Taleb offers

little.

2. (Additional ways of quantifying “outliers”)

- (a) Consider the outlier probability,

$$P|y - \mu| > 5\sigma$$

(there is of course nothing magical about our choice of 5). In practice we use a sample version of the population object.

- (b) Consider the “tail index” γ , such that

$$P(y > y^*) = ky^{*- \gamma}.$$

In practice we use a sample version of the population object.

3. (“Leave-one-out” coefficient plots)

Leave-one-out coefficient plots are more appropriate for cross-section data than for time-series data. Why? How might you adapt them to handle time-series data?

Chapter 6

Group Heterogeneity and Indicator Variables

From one perspective we continue working under the FIC. From another we now begin relaxing the FIC, effectively by recognizing RHS variables that were omitted from, but should not have been omitted from, our original wage regression.

6.1 0-1 Dummy Variables

A **dummy variable**, or **indicator variable**, is just a 0-1 variable that indicates something, such as whether a person is female, non-white, or a union member. We use dummy variables to account for such “group effects,” if any. We might define the dummy UNION, for example, to be 1 if a person is a union member, and 0 otherwise. That is,

$$UNION_i = \begin{cases} 1, & \text{if observation } i \text{ corresponds to a union member} \\ 0, & \text{otherwise.} \end{cases}$$

In Figure 6.1 we show histograms and statistics for all potential determinants of wages. Education (EDUC) and experience (EXPER) are standard continuous variables, although we measure them only discretely (in years);

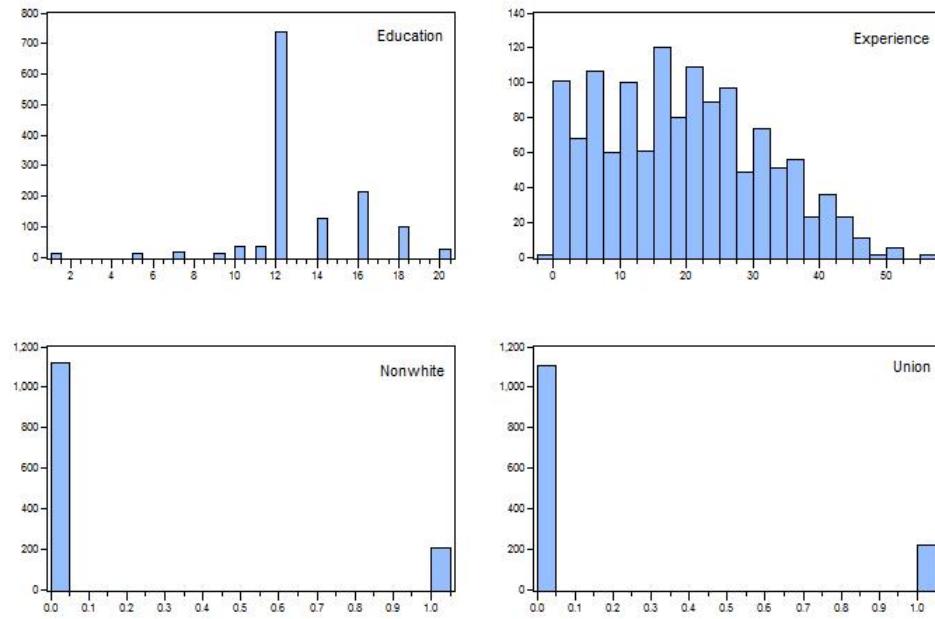


Figure 6.1: Histograms for Wage Covariates

we have examined them before and there is nothing new to say. The new variables are 0-1 dummies, UNION (already defined) and NONWHITE, where

$$NONWHITE_i = \begin{cases} 1, & \text{if observation } i \text{ corresponds to a non - white person} \\ 0, & \text{otherwise.} \end{cases}$$

Note that the sample mean of a dummy variable is the fraction of the sample with the indicated attribute. The histograms indicate that roughly one-fifth of people in our sample are union members, and roughly one-fifth are non-white.

We also have a third dummy, FEMALE, where

$$FEMALE_i = \begin{cases} 1, & \text{if observation } i \text{ corresponds to a female} \\ 0, & \text{otherwise.} \end{cases}$$

We don't show its histogram because it's obvious that FEMALE should be approximately 0 w.p. 1/2 and 1 w.p. 1/2, which it is.

Sometimes dummies like UNION, NONWHITE and FEMALE are called **intercept dummies**, because they effectively allow for a different intercept for each group (union vs. non-union, non-white vs. white, female vs. male). The regression intercept corresponds to the “base case” (zero values for all dummies) and the dummy coefficients give the extra effects when the respective dummies equal one. For example, in a wage regression with an intercept and a single dummy (UNION, say), the intercept corresponds to non-union members, and the estimated coefficient on UNION is the extra effect (up or down) on LWAGE accruing to union members.

Alternatively, we could define and use a full set of dummies for each category (e.g., include both a union dummy and a non-union dummy) and drop the intercept, reading off the union and non-union effects directly.

In any event, never include a full set of dummies *and* an intercept. Doing so would be redundant because the sum of a full set of dummies is just a unit vector, but that’s what the intercept is. If an intercept is included, one of the dummy categories must be dropped.

6.2 Group Dummies in the Wage Regression

Recall our basic wage regression,

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

shown in Figure 6.2. Both explanatory variables are highly significant, with expected signs.

Now consider the same regression, but with our three group dummies added, as shown in Figure 6.3. All dummies are significant with the expected signs, and R^2 is higher. Both SIC and AIC favor including the group dummies. We show the residual scatter in Figure 6.4. Of course it’s hardly the forty-five degree line (the regression R^2 is higher but still only .31), but it’s

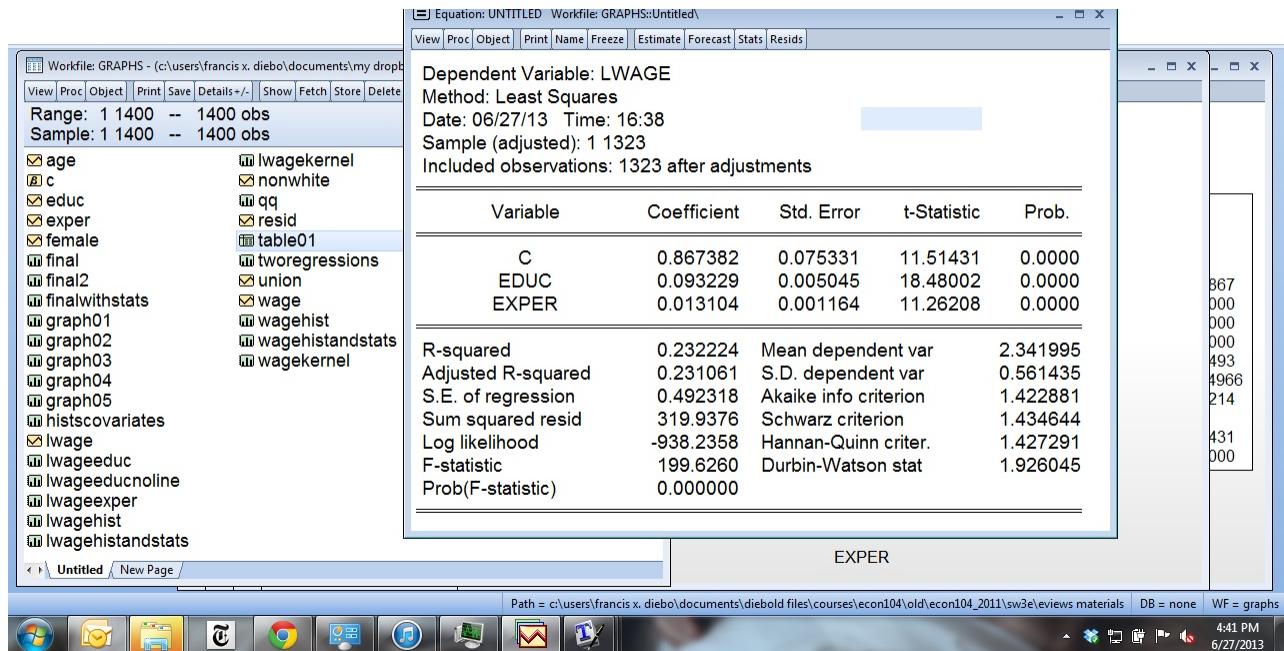


Figure 6.2: Wage Regression on Education and Experience

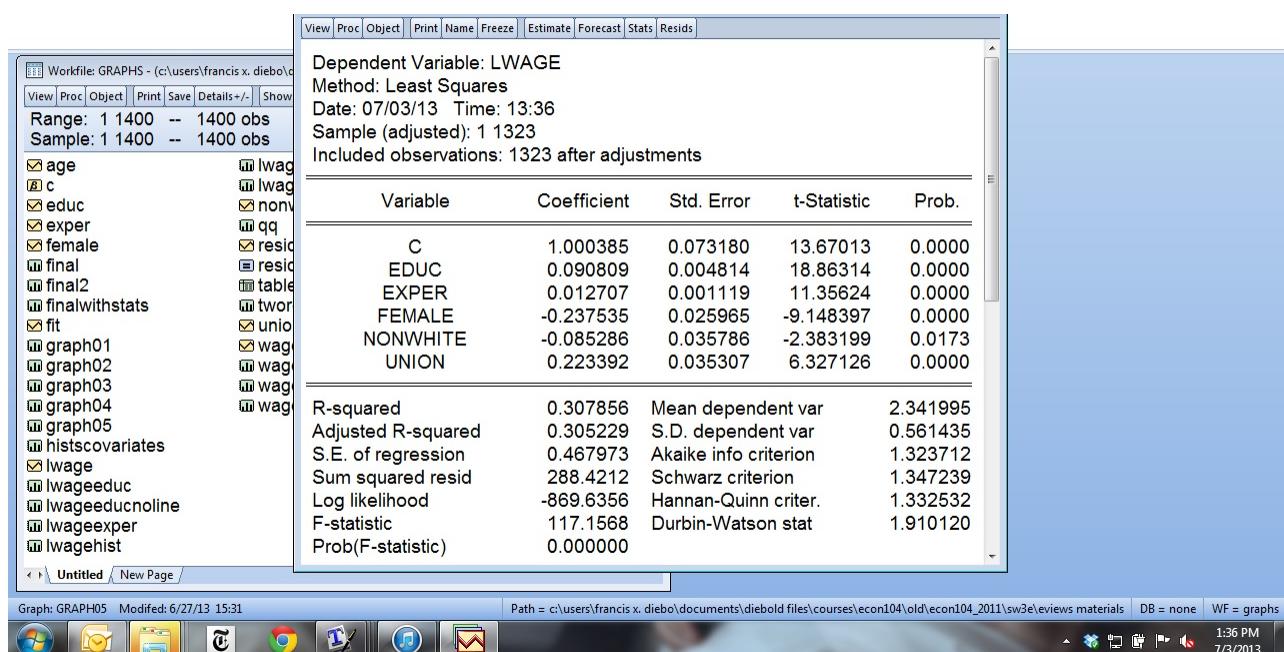


Figure 6.3: Wage Regression on Education, Experience and Group Dummies

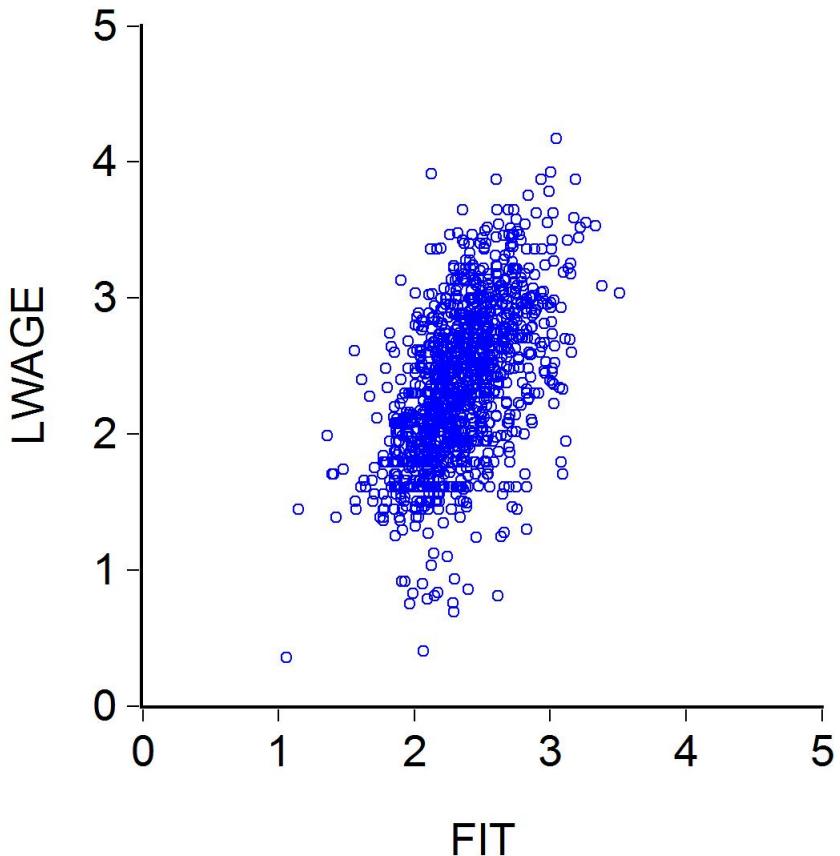


Figure 6.4: Residual Scatter from Wage Regression on Education, Experience and Group Dummies

getting closer.

6.3 Exercises, Problems and Complements

1. (Slope dummies)

Consider the regression

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i.$$

The dummy variable model as introduced in the text generalizes the intercept term such that it can change across groups. Instead of writing

the intercept as β_1 , we write it as $\beta_1 + \delta D_i$.

We can also allow slope coefficients to vary with groups. Instead of writing the slope as β_2 , we write it as $\beta_2 + \gamma D_i$. Hence to capture slope variation across groups we regress y not only on an intercept and x , but also on $D * x$.

Allowing for *both* intercept and slope variation across groups corresponds to regressing on an intercept, D , x , and $D * x$.

2. (Dummies vs. separate regression)

Consider the simple regression, $y \rightarrow c, x$.

- (a) How is inclusion of a group G intercept dummy related to the idea of running separate regressions, one for G and one for non- G ? Are the two strategies equivalent? Why or why not?
- (b) How is inclusion of group G intercept and slope dummies related to the idea of running separate regressions, one for G and one for non- G ? Are the two strategies equivalent? Why or why not?

3. (Analysis of variance (ANOVA) and dummy variable regression)

[You should have learned about **analysis of variance** (ANOVA) in your earlier statistics course. In any event there's good news: If you understand regression on dummy variables, you understand analysis of variance (ANOVA), as any ANOVA analysis can be done via regression on dummies. So here we go.]

You treat each of 1000 randomly-selected farms that presently use no fertilizer. You either do nothing, or you apply one of four experimental fertilizers, A, B, C or D. Using a dummy variable regression setup:

- (a) How would you test the hypothesis that none of the four new fertilizers is effective?

- (b) Assuming that you reject the null, how would you estimate the improvement (or worsening) due to using fertilizer A, B, C or D?

6.4 Notes

ANOVA traces to Sir Ronald Fischer's 1918 article, "The Correlation Between Relatives on the Supposition of Mendelian Inheritance," and it was featured prominently in his classic 1925 book, *Statistical Methods for Research Workers*. Fischer is in many ways the "father" of much of modern statistics.

6.5 Dummy Variables, ANOVA, and Sir Ronald Fischer



Figure 6.5: Sir Ronald Fischer

Photo credit: From Wikimedia commons. Source: http://www.swlearning.com/quant/kohler/stat/biographical_sketches/Fisher_3.jpeg Rationale: Photographer died >70yrs ago => PD. Date: 2008-05-30 (original upload date). Source: Transferred from en.wikipedia. Author: Original uploader was Bletchley at en.wikipedia. Permission (Reusing this file): Released under the GNU Free Documentation License; PD-OLD-70. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled GNU Free Documentation License.

Chapter 7

Nonlinearity

In general there is no reason why the conditional mean function should be linear. That is, the appropriate **functional form** may not be linear. Whether linearity provides an adequate approximation is an empirical matter.

Non-linearity is related to non-normality, which we studied in chapter 5. In particular, in the multivariate normal case, the conditional mean function is linear in the conditioning variables. But once we leave the *terra firma* of multivariate normality, anything goes. The **conditional mean function** and disturbances may be linear and Gaussian, non-linear and Gaussian, linear and non-Gaussian, or non-linear and non-Gaussian.

In the Gaussian case, because the conditional mean is a linear function of the conditioning variable(s), it coincides with the **linear projection**. In non-Gaussian cases, however, linear projections are best viewed as approximations to generally non-linear conditional mean functions. That is, we can view the linear regression model as a linear approximation to a generally non-linear conditional mean function. Sometimes the linear approximation may be adequate, and sometimes not.

7.1 Models Linear in Transformed Variables

Models can be non-linear but nevertheless linear in non-linearly-transformed variables. A leading example involves logarithms, to which we now turn. This can be very convenient. Moreover, coefficient interpretations are special, and similarly convenient.

7.1.1 Logarithms

Logs turn multiplicative models additive, and they neutralize exponentials. Logarithmic models, although non-linear, are nevertheless “linear in logs.”

In addition to turning certain non-linear models linear, they can be used to enforce non-negativity of a left-hand-side variable and to stabilize a disturbance variance. (More on that later.)

Log-Log Regression

First, consider **log-log regression**. We write it out for the simple regression case, but of course we could have more than one regressor. We have

$$\ln y_i = \beta_1 + \beta_2 \ln x_i + \varepsilon_i.$$

y_i is a non-linear function of the x_i , but the function is linear in logarithms, so that ordinary least squares may be applied.

To take a simple example, consider a Cobb-Douglas production function with output a function of labor and capital,

$$y_i = AL_i^\alpha K_i^\beta \exp(\varepsilon_i).$$

Direct estimation of the parameters A, α, β would require special techniques. Taking logs, however, yields

$$\ln y_i = \ln A + \alpha \ln L_i + \beta \ln K_i + \varepsilon_i.$$

This transformed model can be immediately estimated by ordinary least squares. We simply regress $\ln y_i$ on an intercept, $\ln L_i$ and $\ln K_i$. Such log-log regressions often capture relevant non-linearities, while nevertheless maintaining the convenience of ordinary least squares.

Note that the estimated intercept is an estimate of $\ln A$ (not A , so if you want an estimate of A you must exponentiate the estimated intercept), and the other estimated parameters are estimates of α and β , as desired.

Recall that for close y_i and x_i , $(\ln y_i - \ln x_i)$ is approximately the percent difference between y_i and x_i . Hence the coefficients in log-log regressions give the expected percent change in $E(y_i|x_i)$ for a one-percent change in x_i , the so-called *elasticity of y_i with respect to x_i* .

Log-Lin Regression

Second, consider **log-lin regression**, in which $\ln y_i = \beta x_i + \varepsilon_i$. We have a log on the left but not on the right. The classic example involves the workhorse model of exponential growth:

$$y_t = Ae^{rt}\varepsilon_t$$

It's non-linear due to the exponential, but taking logs yields

$$\ln y_t = \ln A + rt + \varepsilon_t,$$

which is linear. The growth rate r gives the approximate percent change in $E(y_t|t)$ for a one-unit change in time (because logs appear only on the left).

Lin-Log Regression

Finally, consider **lin-log Regression**:

$$y_i = \beta \ln x_i + \varepsilon_i$$

It's a bit exotic but it sometimes arises. β gives the effect on $E(y_i|x_i)$ of a one-percent change in x_i , because logs appear only on the right.

7.1.2 Box-Cox and GLM

Box-Cox

The **Box-Cox transformation** generalizes log-lin regression. We have

$$B(y_i) = \beta_1 + \beta_2 x_i + \varepsilon_i,$$

where

$$B(y_i) = \frac{y_i^\lambda - 1}{\lambda}.$$

Hence

$$E(y_i|x_i) = B^{-1}(\beta_1 + \beta_2 x_i).$$

Because

$$\lim_{\lambda \rightarrow 0} \left(\frac{y_i^\lambda - 1}{\lambda} \right) = \ln(y_i),$$

the Box-Cox model corresponds to the log-lin model in the special case of $\lambda = 0$.

GLM

The so-called “generalized linear model” (GLM) provides an even more flexible framework. Almost all models with left-hand-side variable transformations are special cases of those allowed in the **generalized linear model (GLM)**. In the GLM, we have

$$G(y_i) = \beta_1 + \beta_2 x_i + \varepsilon_i,$$

so that

$$E(y_i|x_i) = G^{-1}(\beta_1 + \beta_2 x_i).$$

Wide classes of “link functions” G can be entertained. Log-lin regression, for example, emerges when $G(y_i) = \ln(y_i)$, and Box-Cox regression emerges when $G(y_i) = \frac{y_i^\lambda - 1}{\lambda}$.

7.2 Intrinsically Non-Linear Models

Sometimes we encounter **intrinsically non-linear models**. That is, there is no way to transform them to linearity, so that they can then be estimated simply by least squares, as we have always done so far.

As an example, consider the **logistic model**,

$$y_i = \frac{1}{a + br^{x_i}} + \varepsilon_i,$$

with $0 < r < 1$. The precise shape of the logistic curve of course depends on the precise values of a , b and r , but its “S-shape” is often useful. The key point for our present purposes is that there is no simple transformation of y that produces a model linear in the transformed variables.

7.2.1 Nonlinear Least Squares

The least squares estimator is often called “ordinary” least squares, or OLS. As we saw earlier, the OLS estimator has a simple closed-form analytic expression, which makes it trivial to implement on modern computers. Its computation is fast and reliable.

The adjective “ordinary” distinguishes ordinary least squares from more laborious strategies for finding the parameter configuration that minimizes the sum of squared residuals, such as the **non-linear least squares (NLS)** estimator. When we estimate by non-linear least squares, 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 β values until we find those that appear to minimize the sum of squared residuals. This is not only more laborious (and hence slow), but also less reliable, as, for example, one may arrive at a minimum that is local but not global.

Why then would anyone ever use non-linear least squares as opposed to

OLS? Indeed, when OLS is feasible, we generally *do* prefer it. For example, in all regression models discussed thus far OLS is applicable, so we prefer it. Intrinsically non-linear models can't be estimated using OLS, but they can be estimated using non-linear least squares. We resort to non-linear least squares in such cases.

Intrinsically non-linear models obviously violate the linearity assumption of the IC. But the violation is not a big deal. Under the remaining IC (that is, dropping only linearity), $\hat{\beta}_{NLS}$ has a sampling distribution similar to that under the IC.

7.3 Series Expansions

There is really no such thing as an intrinsically non-linear model. In the bivariate case we can think of the relationship as

$$y_i = g(x_i, \varepsilon_i)$$

or slightly less generally as

$$y_i = f(x_i) + \varepsilon_i.$$

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

$$f(x_i) \approx \beta_1 + \beta_2 x_i$$

and the quadratic (second-order) approximation is

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

In the multiple regression case, Taylor approximations also involves interaction terms. Consider, for example, a function of two regressors, $f(x_i, z_i)$.

The second-order Taylor approximation is:

$$f(x_i, z_i) \approx \beta_1 + \beta_2 x_i + \beta_3 z_i + \beta_4 x_i^2 + \beta_5 z_i^2 + \beta_6 x_i z_i.$$

The final term picks up **interaction effects**. Interaction effects are also relevant in situations involving dummy variables. There we capture interactions by including products of dummies.¹

The ultimate point is that even 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 (e.g., quadratic) so that only a few series terms are required. From this viewpoint non-linearity is in some sense really an omitted-variables problem.

One can also use Fourier series approximations:

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

and 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.

7.4 A Final Word on Nonlinearity and the IC

It is of interest to step back and ask what parts of the IC are violated in our various non-linear models.

Models linear in transformed variables (e.g., log-log regression) actually *don't* violate the IC, after transformation. Neither do series expansion models, if the adopted expansion order is deemed correct, because they too are linear in transformed variables.

The series approach to handling non-linearity is actually very general and

¹Notice that a product of dummies is one if and only if both individual dummies are one.

handles intrinsically non-linear models as well, and low-ordered expansions are often adequate in practice, even if an infinite expansion is required in theory. If series terms are needed, a purely linear model would suffer from misspecification of the X matrix (a violation of the IC) due to the omitted higher-order expansion terms. Hence the failure of the IC discussed in this chapter can be viewed either as:

1. The linearity assumption ($E(y|X) = X'\beta$) is incorrect, or
2. The linearity assumption ($E(y|X) = X'\beta$) is correct, but the assumption that X is correctly specified (i.e., no omitted variables) is incorrect, due to the omitted higher-order expansion terms.

7.5 Selecting a Non-Linear Model

7.5.1 t and F Tests, and Information Criteria

One can use the usual t and F tests for testing linear models against non-linear alternatives in nested cases, and information criteria (AIC and SIC) for testing against non-linear alternatives in non-nested cases. To test linearity against a quadratic alternative in a simple regression case, for example, we can simply run $y \rightarrow c, x, x^2$ and perform a t -test for the relevance of x^2 .

And of course, use AIC and SIC as always.

7.5.2 The RESET Test

Direct inclusion of powers and cross products of the various x variables in the regression can be wasteful of degrees of freedom, however, particularly if there are more than just one or two right-hand-side variables in the regression and/or if the non-linearity is severe, so that fairly high powers and interactions would be necessary to capture it.

In light of this, a useful strategy is first to fit a linear regression $y_i \rightarrow c, x_i$ and obtain the fitted values \hat{y}_i . Then, to test for non-linearity, we run the regression again with various powers of \hat{y}_i included,

$$y_i \rightarrow c, x_i, \hat{y}_i^2, \dots, \hat{y}_i^m.$$

Note that the powers of \hat{y}_i are linear combinations of powers and cross products of the x variables – just what the doctor ordered. There is no need to include the first power of \hat{y}_i , because that would be redundant with the included x variables. Instead we include powers $\hat{y}_i^2, \hat{y}_i^3, \dots$. Typically a small m is adequate. Significance of the included set of powers of \hat{y}_i can be checked using an F test. This procedure is called RESET (Regression Specification Error Test).

7.6 Non-Linearity in Wage Determination

For convenience we reproduce in Figure 7.1 the results of our current linear wage regression,

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

$$FEMALE, UNION, NONWHITE.$$

The RESET test from that regression suggests neglected non-linearity; the p -value is .03 when using \hat{y}_t^2 and \hat{y}_t^3 in the RESET test regression.

Non-Linearity in EDUC and EXPER: Powers and Interactions

Given the results of the RESET test, we proceed to allow for non-linearity.

In Figure 7.2 we show the results of the quadratic regression

$$LWAGE \rightarrow EDUC, EXPER$$

$$EDUC^2, EXPER^2, EDUC * EXPER,$$

$$FEMALE, UNION, NONWHITE$$

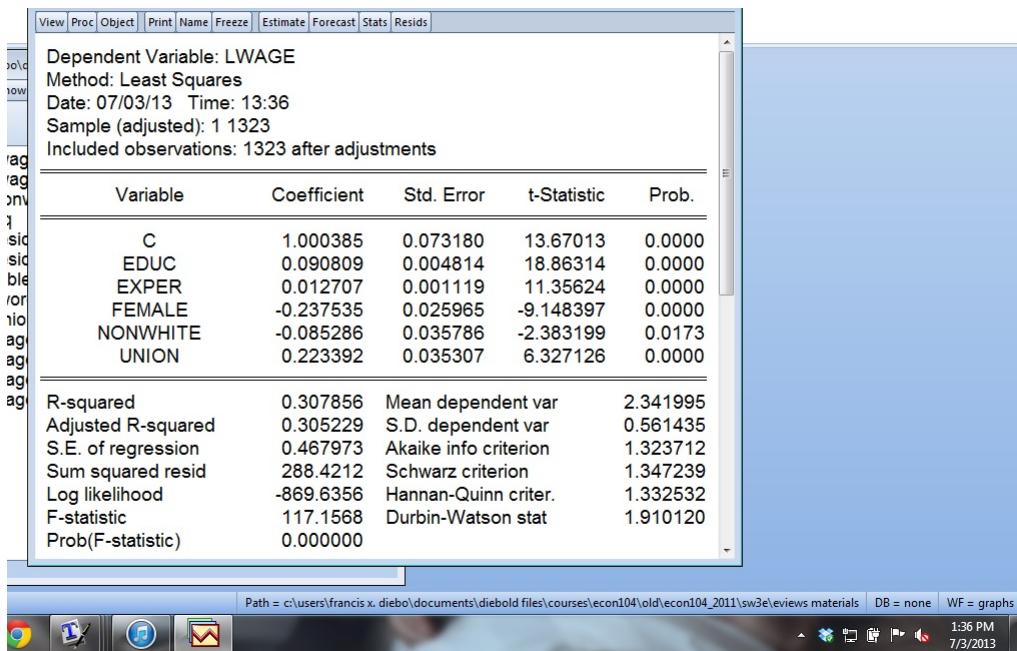


Figure 7.1: Basic Linear Wage Regression

Two of the non-linear effects are significant. The impact of experience is decreasing, and experience seems to trade off with education, insofar as the interaction is negative.

Non-Linearity in FEMALE, UNION and NONWHITE: Interactions

Just as continuous variables like *EDUC* and *EXPER* may interact (and we found that they do), so too may discrete dummy variables. For example, the wage effect of being female *and* non-white might not simply be the sum of the individual effects. We would estimate it as the sum of coefficients on the individual dummies *FEMALE* and *NONWHITE* *plus* the coefficient on the interaction dummy *FEMALE*NONWHITE*.

In Figure 7.4 we show results for

$$LWAGE \rightarrow EDUC, EXPER,$$

$$FEMALE, UNION, NONWHITE,$$

$$FEMALE*UNION, FEMALE*NONWHITE, UNION*NONWHITE.$$

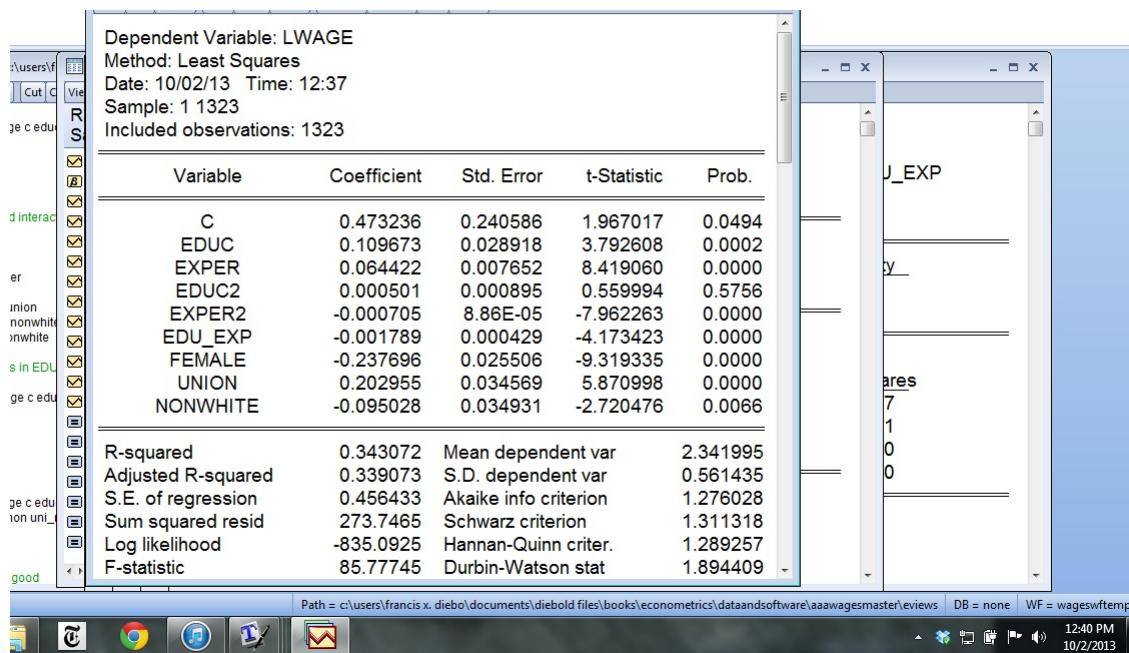


Figure 7.2: Quadratic Wage Regression

The dummy interactions are insignificant.

7.6.1 Non-Linearity in Continuous and Discrete Variables Simultaneously

Now let's incorporate powers and interactions in *EDUC* and *EXPER*, and interactions in *FEMALE*, *UNION* and *NONWHITE*.

In Figure 7.4 we show results for

$$LWAGE \rightarrow EDUC, EXPER,$$

$$EDUC^2, EXPER^2, EDUC * EXPER,$$

$$FEMALE, UNION, NONWHITE,$$

$$FEMALE*UNION, FEMALE*NONWHITE, UNION*NONWHITE.$$

The dummy interactions remain insignificant.

Note that we could explore additional interactions among *EDUC*, *EXPER*

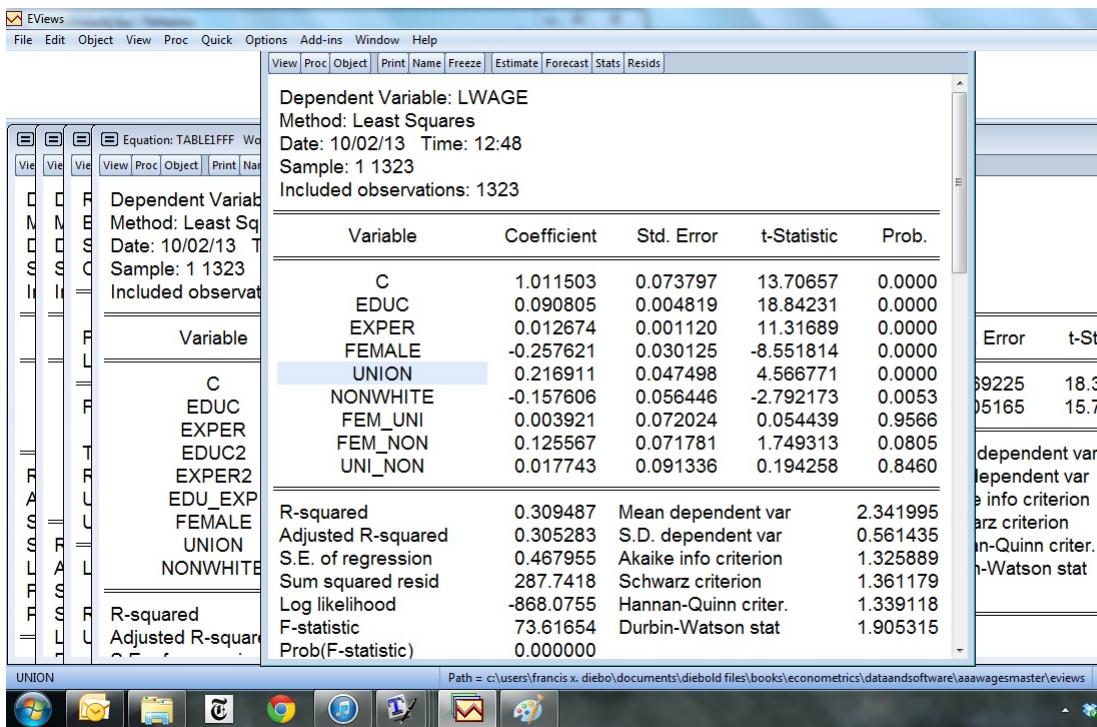


Figure 7.3: Wage Regression on Education, Experience, Group Dummies, and Interactions

and the various dummies. We leave that to the reader.

Assembling all the results, our tentative “best” model thus far is the one of section 7.6,

$$LWAGE \rightarrow EDUC, EXPER,$$

$$EDUC^2, EXPER^2, EDUC * EXPER,$$

$$FEMALE, UNION, NONWHITE.$$

The RESET statistic has a p -value of .19, so we would not reject adequacy of functional form at conventional levels.

7.7 Exercises, Problems and Complements

1. (Tax revenue and the tax rate)

The U.S. Congressional Budget Office (CBO) is helping the president

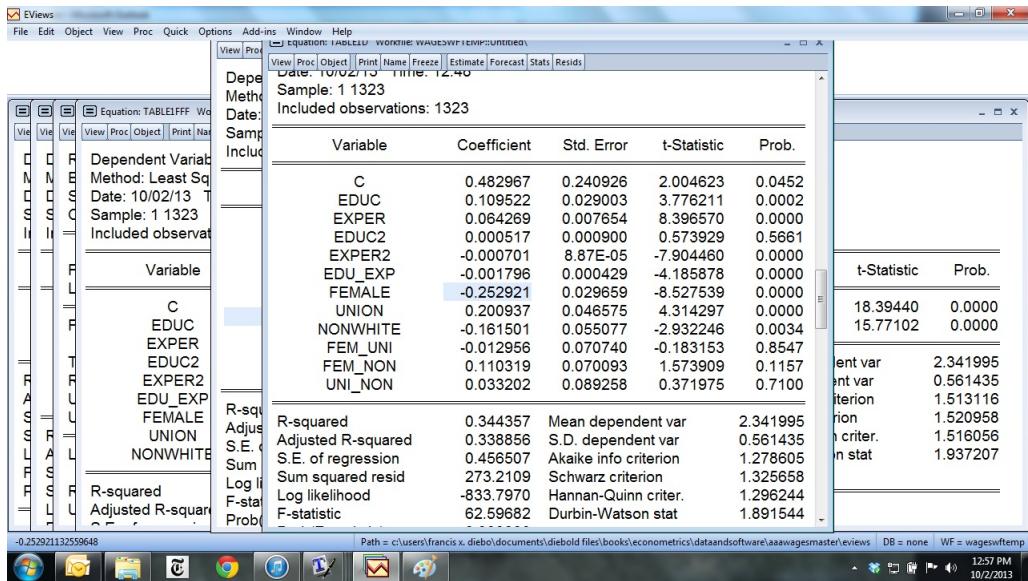


Figure 7.4: Wage Regression with Continuous Non-Linearities and Interactions, and Discrete Interactions

to set tax policy. In particular, the president has asked for advice on where to set the average tax rate to maximize the tax revenue collected per taxpayer. For each of 65 countries the CBO has obtained data on the tax revenue collected per taxpayer and the average tax rate.

- Is tax revenue likely related to the tax rate? (That is, do you think that the mean of tax revenue conditional on the tax rate actually *is* a function of the tax rate?)
 - Is the relationship likely linear? (Hint: how much revenue would be collected at tax rates of zero or one hundred percent?)
 - If not, is a linear regression nevertheless likely to produce a good approximation to the true relationship?
2. (Graphical regression diagnostic: scatterplot of e_i vs. x_i)

This plot helps us assess whether the relationship between y and x is truly linear, as assumed in linear regression analysis. If not, the linear regression residuals will depend on x . In the case where there is only

one right-hand side variable, as above, we can simply make a scatterplot of e_i vs. x_i . When there is more than one right-hand side variable, we can make separate plots for each, although the procedure loses some of its simplicity and transparency.

3. (Difficulties with non-linear optimization)

Non-linear optimization can be a tricky business, fraught with problems. Some problems are generic. It's relatively easy to find a local optimum, for example, but much harder to be confident that the local optimum is global. Simple checks such as trying a variety of startup values and checking the optimum to which convergence occurs are used routinely, but the problem nevertheless remains. Other problems may be software specific. For example, some software may use highly accurate analytic derivatives whereas other software uses approximate numerical derivatives. Even the same software package may change algorithms or details of implementation across versions, leading to different results.

4. (Conditional mean functions)

Consider the regression model,

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \beta_4 z_i + \varepsilon_i$$

under the full ideal conditions. Find the mean of y_i conditional upon $x_i = x_i^*$ and $z_i = z_i^*$. Is the conditional mean linear in $(x_i^*? z_i^*)$?

5. (OLS vs. NLS)

Consider the following three regression models:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

$$y_i = \beta_1 e^{\beta_2 x_i} \varepsilon_i$$

$$y_i = \beta_1 + e^{\beta_2 x_i} + \varepsilon_i.$$

- a. For each model, determine whether OLS may be used for estimation (perhaps after transforming the data), or whether NLS is required.
 - b. For those models for which OLS is feasible, do you expect NLS and OLS estimation results to agree precisely? Why or why not?
 - c. For those models for which NLS is “required,” show how to avoid it using series expansions.
6. (What is linear regression really estimating?)

It is important to note the distinction between a conditional mean and a **linear projection**. The conditional mean is not necessarily a linear function of the conditioning variable(s). The linear projection *is* of course a linear function of the conditioning variable(s), by construction. Linear projections are best viewed as approximations to generally non-linear conditional mean functions. That is, we can view an empirical linear regression as estimating the population linear projection, which in turn is an approximation to the population conditional expectation. Sometimes the linear projection may be an adequate approximation, and sometimes not.

7. Putting lots of things together.

Consider the cross-sectional (log) wage equation that we studied extensively, which appears again in Figure 7.5 for your reference.

- (a) The model was estimated using ordinary least squares (OLS). What loss function is optimized in calculating the OLS estimate? (Give a formula and a graph.) What is the formula (if any) for the OLS estimator?
- (b) Consider instead estimating the same model numerically (i.e., by NLS) rather than analytically (i.e., by OLS). What loss function is

- optimized in calculating the NLS estimate? (Give a formula and a graph.) What is the formula (if any) for the NLS estimator?
- (c) Does the estimated equation indicate a statistically significant effect of union status on log wage? An economically important effect? What is the precise interpretation of the estimated coefficient on UNION? How would the interpretation change if the wage were not logged?
 - (d) Precisely what hypothesis does the F-statistic test? What are the restricted and unrestricted sums of squared residuals to which it is related, and what are the two OLS regressions to which they correspond?
 - (e) Consider an additional regressor, AGE, where $AGE = 6 + EDUC + EXPER$. (The idea is that 6 years of early childhood, followed by EDUC years of education, followed by EXPER years of work experience should, under certain assumptions, sum to a person's age.) Discuss the likely effects, if any, of adding AGE to the regression.
 - (f) The log wage may of course *not* be linear in EDUC and EXPER. How would you assess the possibility of quadratic nonlinear effects using t-tests? An F-test? The Schwarz criterion (SIC)? R^2 ?
 - (g) Suppose you find that the log wage relationship is indeed non-linear but still very simple, with only $EXPER^2$ entering in addition to the variables in Figure 7.5. What is $\frac{\partial E(LWAGE|X)}{\partial EXPER}$ in the expanded model? How does it compare to $\frac{\partial E(LWAGE|X)}{\partial EXPER}$ in the original model of Figure 7.5? What are the economic interpretations of the two derivatives? (X refers to the full set of included right-hand-side variables in a regression.)
 - (h) Return to the original model of Figure 7.5. How would you assess the overall adequacy of the fitted model using the standard error of

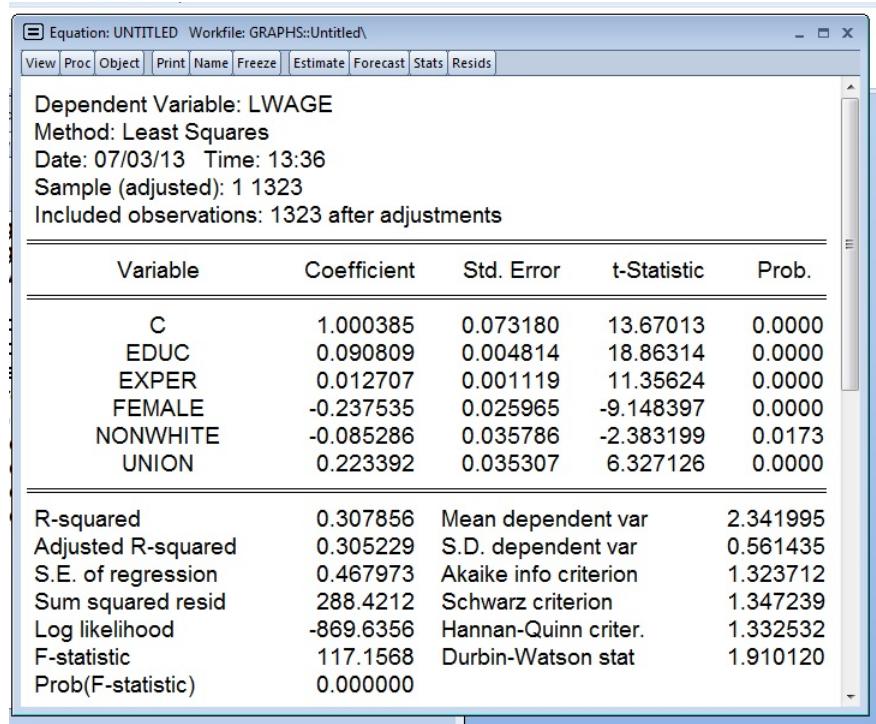


Figure 7.5: Regression Output

the regression? The model residuals? Which is likely to be more useful/informative?

- (i) Consider estimating the model not by OLS or NLS, but rather by quantile regression (QR). What loss function is optimized in calculating the QR estimate? (Give a formula and a graph.) What is the formula (if any) for the QR estimator? How is the least absolute deviations (LAD) estimator related to the QR estimator? Under the IC, are the OLS and LAD estimates likely very close? Why or why not?
- (j) Discuss whether and how you would incorporate trend and seasonality by using a linear time trend variable and a set of seasonal dummy variables.

Chapter 8

Heteroskedasticity

We continue exploring issues associated with possible failure of the ideal conditions. This chapter's issue is “Do we really believe that disturbance variances are constant?” As always, consider: $\varepsilon \sim N(\underline{0}, \Omega)$. Heteroskedasticity corresponds to Ω diagonal but $\Omega \neq \sigma^2 I$

$$\Omega = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_N^2 \end{pmatrix}$$

Heteroskedasticity can arise for many reasons. A leading cause is that σ_i^2 may depend on one or more of the x_i 's. A classic example is an “Engel curve”, a regression relating food expenditure to income. Wealthy people have much more discretion in deciding how much of their income to spend on food, so their disturbances should be more variable, as routinely found.

8.1 Consequences of Heteroskedasticity for Estimation, Inference, and Prediction

As regards point estimation, OLS remains largely OK, insofar as parameter estimates remain consistent and asymptotically normal. They are, however,

rendered inefficient. But consistency is key. Inefficiency is typically inconsequential in large samples, as long as we have consistency.

As regards inference, however, heteroskedasticity wreaks significant havoc. Standard errors are biased and inconsistent. Hence t statistics do not have the t distribution in finite samples and do not even have the $N(0, 1)$ distribution asymptotically.

Finally, as regards prediction, results vary depending on whether we're talking about point or density prediction. Our earlier feasible point forecasts constructed under homoskedasticity remain useful under heteroskedasticity. Because parameter estimates remain consistent, we still have

$$\widehat{E(y_i \mid x_i=x_i^*)} \rightarrow_p E(y_i \mid x_i=x_i^*)$$

In contrast, our earlier feasible density forecasts do not remain useful, because under heteroskedasticity it is no longer appropriate to base them on “identical σ 's for different people”. Now we need to base them on “different σ 's for different people”.

8.2 Detecting Heteroskedasticity

We will consider both graphical heteroskedasticity diagnostics and formal heteroskedasticity tests. The two approaches are complements, not substitutes.

8.2.1 Graphical Diagnostics

The first thing we can do is graph e_i^2 against x_i , for various regressors, looking for relationships. This makes sense because e_i^2 is effectively a proxy for σ_i^2 . Recall, for example, our “Final” wage regression, shown in Figure 8.1.

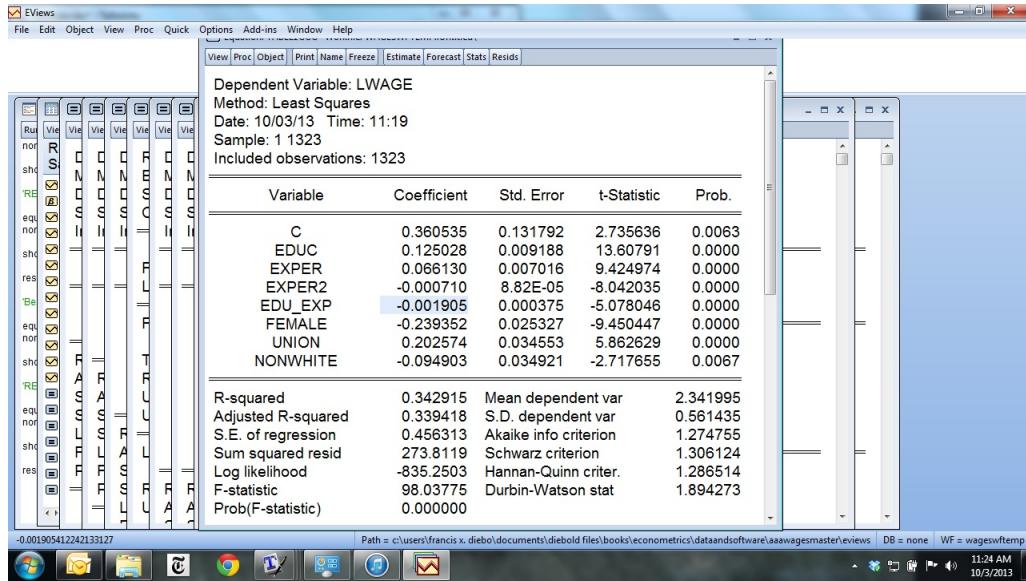


Figure 8.1: Final Wage Regression

In Figure 8.2 we graph the squared residuals against EDUC. There is apparently a positive relationship, although it is noisy. This makes sense, because very low education almost always leads to very low wage, whereas high education can produce a larger variety of wages (e.g., both neurosurgeons and college professors are highly educated, but neurosurgeons typically earn much more).

8.2.2 Formal Tests

The Breusch-Pagan-Godfrey Test (BPG)

An important limitation of the graphical method for heteroskedasticity detection is that it is purely pairwise (we can only examine one x at a time), whereas the disturbance variance might actually depend on more than one x . Formal tests let us blend the information from multiple x 's, and they also let us assess statistical significance.

The BPG test proceeds as follows:

1. Estimate the OLS regression, and obtain the squared residuals

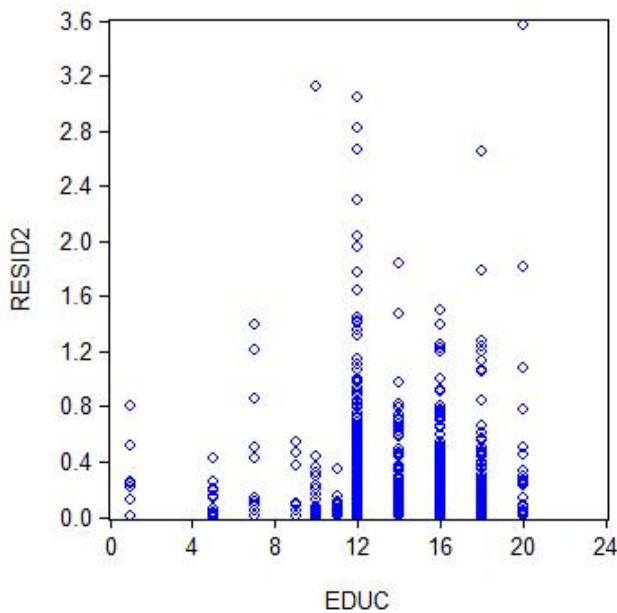


Figure 8.2: Squared Residuals vs. Years of Education

2. Regress the squared residuals on all regressors
3. To test the null hypothesis of no relationship, examine $N \cdot R^2$ from this regression. It can be shown that in large samples $N \cdot R^2 \sim \chi_{K-1}^2$ under the null of homoskedasticity, where K is the number of regressors in the test regression.

We show the BPG test results in Figure 8.3.

White's Test

White's test is a simple extension of BPG, replacing the linear BPG test regression with a more flexible (quadratic) regression:

1. Estimate the OLS regression, and obtain the squared residuals
2. Regress the squared residuals on all regressors, squared regressors, and pairwise regressor cross products

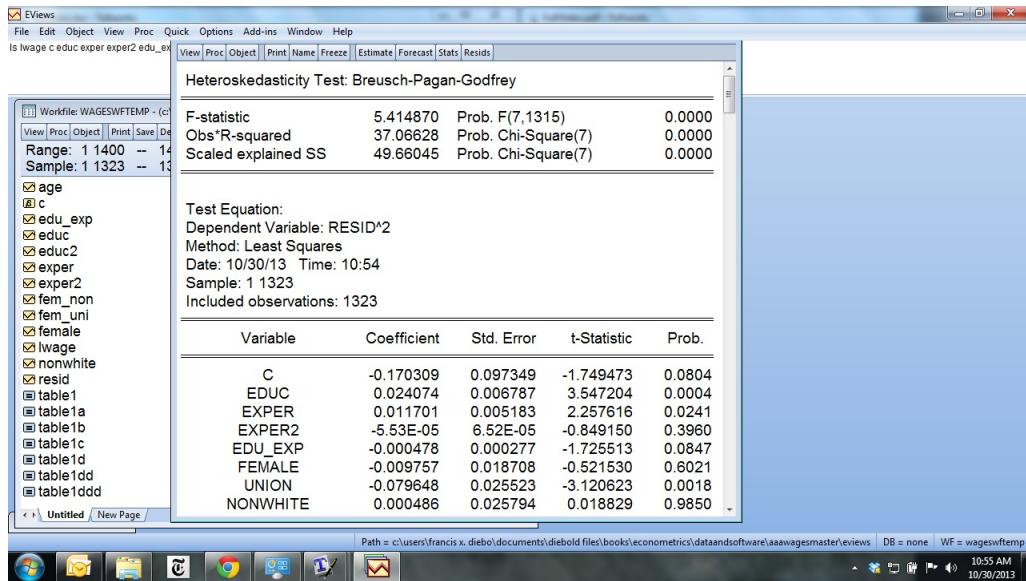


Figure 8.3: BPG Test Regression and Results

3. To test the null hypothesis of no relationship, examine $N \cdot R^2$ from this regression. It can be shown that in large samples $N \cdot R^2 \sim \chi^2_{K-1}$ under the null of homoskedasticity, where K is the number of regressors in the test regression.

We show the White test results in Figure 8.4.

8.3 Dealing with Heteroskedasticity

We will consider both adjusting standard errors and adjusting density forecasts.

8.3.1 Adjusting Standard Errors

Using advanced methods, one *can* obtain consistent standard errors, even when heteroskedasticity is present. Mechanically, it's just a simple regression option. e.g., in EViews, instead of "ls y,c,x", use "ls(cov=white) y,c,x"

Even if you're only interested in prediction, you still might want to use robust standard errors, in order to do credible inference regarding the con-

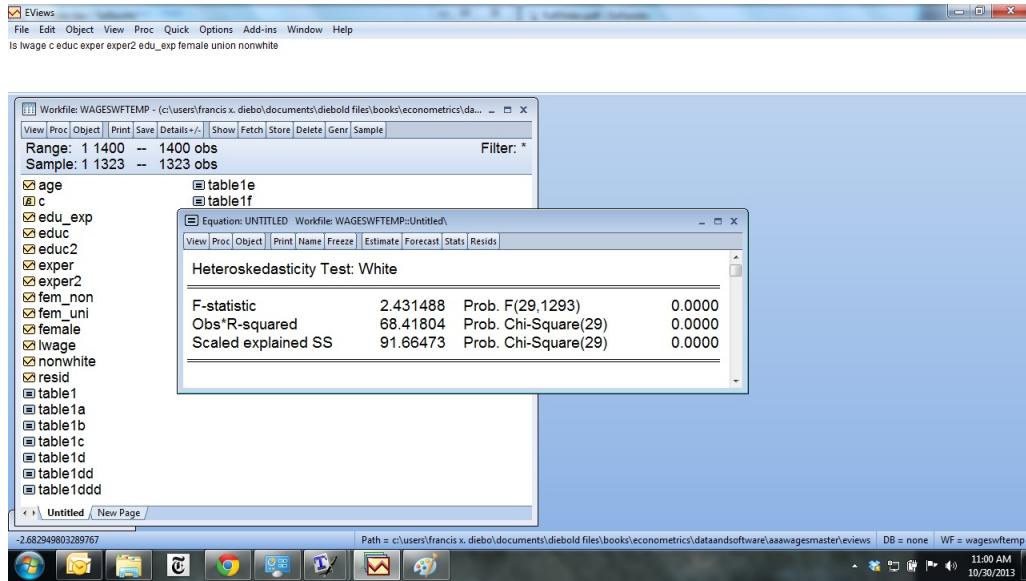


Figure 8.4: White Test Regression and Results

tributions of the various x variables to the point prediction.

In Figure 8.5 we show the final wage regression with robust standard errors. Although the exact values of the standard errors change, it happens in this case that significance of all coefficients is preserved.

8.3.2 Adjusting Density Forecasts

Recall operational density forecast under the ideal conditions (which include, among other things, Gaussian homoskedastic disturbances):

$$y_i \mid x_i=x^* \sim N(x^{*'} \hat{\beta}_{LS}, s^2).$$

Now, under heteroskedasticity (but maintaining normality), we have the natural extension,

$$y_i \mid x_i=x^* \sim N(x^{*'} \hat{\beta}_{LS}, \hat{\sigma}_*^2),$$

where $\hat{\sigma}_*^2$ is the fitted value from the BPG or White test regression evaluated at x^* .

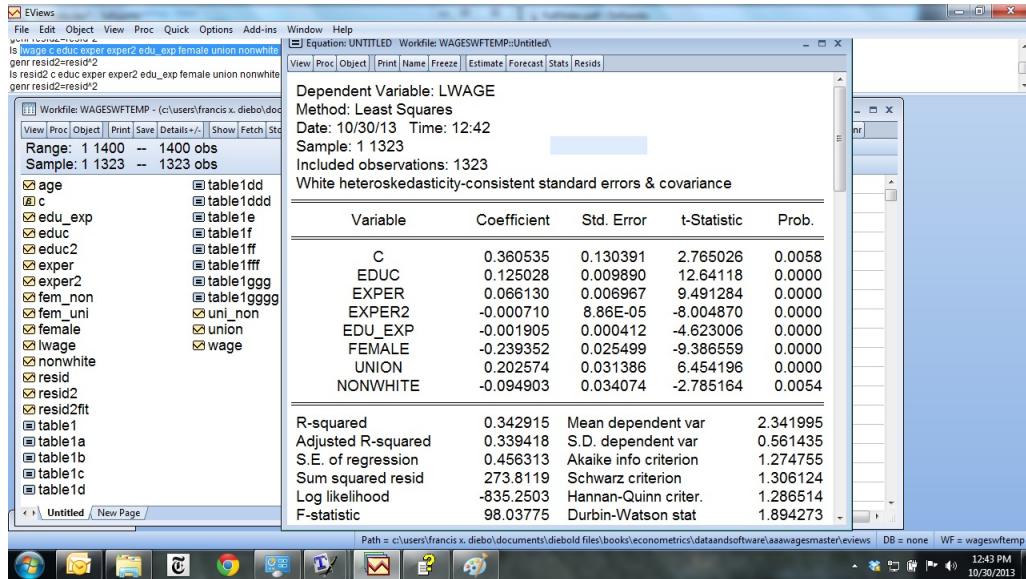


Figure 8.5: Wage Regression with Heteroskedasticity-Robust Standard Errors

8.4 Exercises, Problems and Complements

1. (Vocabulary)

All these have the same meaning:

- (a) “Heteroskedasticity-robust standard errors”
- (b) “heteroskedasticity-consistent standard errors”
- (c) “Robust standard errors”
- (d) “White standard errors”
- (e) “White-washed” standard errors”

2. (Generalized Least Squares (GLS))

For arbitrary Ω matrix, it can be shown that full estimation efficiency requires “generalized least squares” (GLS) estimation. The GLS estimator is:

$$\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y.$$

Under the ideal conditions (but allowing for $\Omega \neq \sigma^2 I$) it is consistent,

MVUE, and normally distributed with covariance matrix $(X'\Omega^{-1}X)^{-1}$:

$$\hat{\beta}_{GLS} \sim N(\beta, (X'\Omega^{-1}X)^{-1}).$$

- (a) Show that when $\Omega = \sigma^2 I$ the GLS estimator is just the standard OLS estimator:

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

- (b) Show that when $\Omega = \sigma^2 I$ the covariance matrix of the GLS estimator is just that of the standard OLS estimator:

$$cov(\hat{\beta}_{GLS}) = cov(\hat{\beta}_{OLS}) = \sigma^2(X'X)^{-1}.$$

3. (GLS for Heteroskedasticity)

- (a) Show that GLS for heteroskedasticity amounts to OLS on data weighted by the inverse disturbance standard deviation $(1/\sigma_i)$, often called “weighted least squares” (WLS). This is “infeasible” WLS since in general we don’t know the σ_i ’s.
- (b) To see why WLS works, consider the heteroskedastic DGP:

$$y_i = x_i'\beta + \varepsilon_i$$

$$\varepsilon_i \sim idN(0, \sigma_i^2).$$

Now weight the data (y_i, x_i) by $1/\sigma_i$:

$$\frac{y_i}{\sigma_i} = \frac{x_i'\beta}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i}.$$

The transformed (but equivalent) DGP is then:

$$y_i^* = x_i^{*\prime}\beta + \varepsilon_i^*$$

$$\varepsilon_i^* \sim iidN(0, 1).$$

The weighted data satisfies the IC and so OLS is MVUE! So GLS is just OLS on appropriately transformed data. In the heteroskedasticity case the appropriate transformation is weighting. We downweight high-variance observations, as is totally natural.

4. (Details of Weighted Least Squares)

Note that weighting the data by $1/\sigma_i$ is the same as weighting the residuals by $1/\sigma_i^2$:

$$\min_{\beta} \sum_{i=1}^N \left(\frac{y_i - x'_i \beta}{\sigma_i} \right)^2 = \min_{\beta} \sum_{i=1}^N \frac{1}{\sigma_i^2} (y_i - x'_i \beta)^2.$$

5. (Feasible Weighted Least Squares)

To make WLS feasible, we need to replace the unknown σ_i^2 's with estimates.

- Good idea: Use weights $w_i = 1/\hat{e}_i^2$, where \hat{e}_i^2 are from the BGP test regression
- Good idea: Use $w_i = 1/\hat{e}_i^2$, where \hat{e}_i^2 are from the White test regression.
- Bad idea: Use $w_i = 1/e_i^2$ is *not* a good idea. e_i^2 is too noisy; we'd like to use not e_i^2 but rather $E(e_i^2|x_i)$. So we use an estimate of $E(e_i^2|x_i)$, namely \hat{e}_i^2 from $e^2 \rightarrow X$

In Figure 8.6 we show regression results with weighting based on the results from the White test regression.

6. (Robustness iteration)

Sometimes, after an OLS regression, people do a second-stage WLS with weights $1/|e_i|$, or something similar. This is not a heteroskedasticity

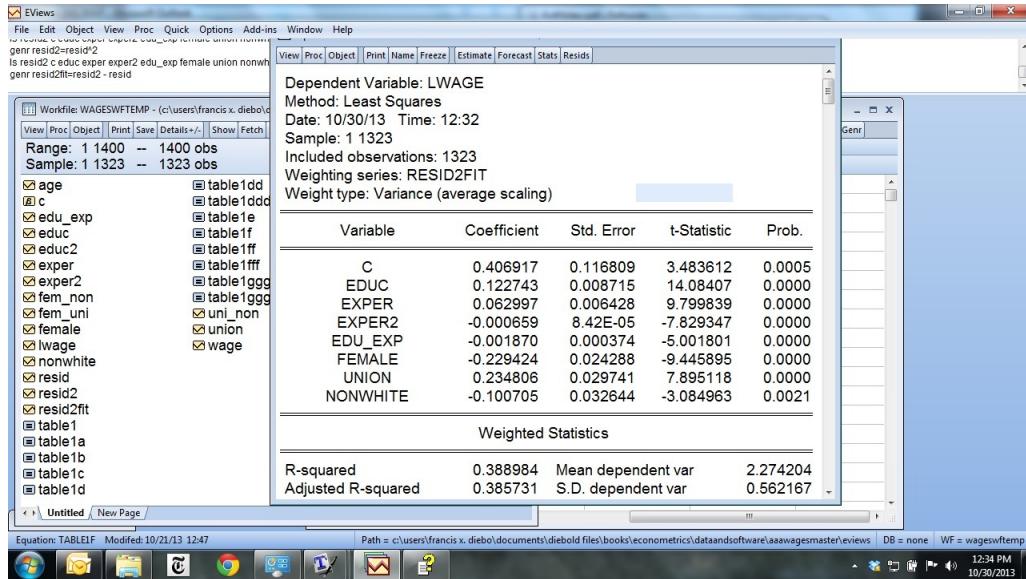


Figure 8.6: Regression Weighted by Fit From White Test Regression

correction, but rather a strategy to downweight outliers. But notice that the two are closely related.

7. (Spatial Correlation)

So far we have studied a heteroskedastic situation (ε_i independent across i but not identically distributed across t). But do we really believe that the disturbances are uncorrelated over space (i)? Spatial correlation in cross sections is another type of violation of the IC. (This time it's “non-zero disturbance covariances” as opposed to “non-constant disturbance variances”.) As always, consider $\varepsilon \sim N(\underline{0}, \Omega)$. Spatial correlation (with possible heteroskedasticity as well) corresponds to:

$$\Omega = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \dots & \sigma_N^2 \end{pmatrix}.$$

8. (“Clustering” in spatial correlation)

Ω could be non-diagonal in cross sections but still sparse in certain ways. A key case is block-diagonal Ω , in which there is nonzero covariance within certain sets of disturbances, but not across sets (“clustering”).

Chapter 9

Limited Dependent Variables

In this chapter we study so-called “discrete-response models”, or “qualitative-response models”, or “limited dependent variable models”, or “classification models”. Terms like “limited dependent variables,” refer to variables that can take only a limited number of values. The classic case is a 0-1 “dummy variable.” The twist is that the dummy variable appears on the left side of a regression, as opposed to the already-discussed use of dummies on the right.

Dummy right-hand side variables (RHS) variables create no problem, and you already understand them. The new issue is Dummy left-hand-side variables (LHS), which do raise special issues.

9.1 Binary Response

Note that the basic regression model,

$$y_i = x'_i \beta + \varepsilon_i$$

immediately implies that

$$E(y_i|x_i) = x'_i \beta.$$

Here we consider left-hand-side variables $y_i = I_i(z)$, where the dummy variable (“**indicator variable**”) $I_i(z)$ indicates whether event z occurs; that

is,

$$I_i(z) = \begin{cases} 1 & \text{if event } z \text{ occurs} \\ 0 & \text{otherwise.} \end{cases}$$

In that case we have

$$E(I_i(z)|x_i) = x'_i \beta.$$

A key insight, however, is that

$$E(I_i(z)|x_i) = P(I_i(z) = 1|x_i),$$

so the model is effectively

$$P(I_i(z) = 1|x_i) = x'_i \beta. \quad (9.1)$$

That is, when the LHS variable is a 0-1 indicator variable, the model is effectively a model relating a conditional probability to the conditioning variables.

There are numerous “events” that fit the 0-1 paradigm. Examples purchasing behavior does a certain consumer buy or not buy a certain product?, hiring behavior (does a certain firm hire or not hire a certain worker?), and loan defaults (does a certain borrower default or not default on a loan?), and recessions (will a certain country have or not have a recession begin during the next year?).

But how should we “fit a line” when the LHS variable is binary? The **linear probability model** does it by brute-force OLS regression $I_i(z) \rightarrow x_i$. There are several econometric problems associated with such regressions, but the one of particular relevance is simply that the linear probability model fails to constrain the fitted values of $E(I_i(z)|x_i) = P(I_i(z) = 1|x_i)$ to lie in the unit interval, in which probabilities must of course lie. We now consider models that impose that constraint by running $x'_i \beta$ through a “squashing function,” $F(\cdot)$, that keeps $P(I_i(z) = 1|x_i)$ in the unit interval. That is, we

move to models with

$$P(I_i(z) = 1|x_i) = F(x'_i \beta),$$

where $F(\cdot)$ is monotone increasing, with $\lim_{w \rightarrow -\infty} F(w) = 0$ and $\lim_{w \rightarrow \infty} F(w) = 1$. Many squashing functions can be entertained, and many *have* been entertained.

9.2 The Logit Model

The most popular and useful squashing function for our purposes is the logistic function, which takes us to the so-called “logit” model. There are several varieties and issues, to which we now turn.

9.2.1 Logit

In the **logit model**, the squashing function $F(\cdot)$ is the **logistic function**,

$$F(w) = \frac{e^w}{1 + e^w} = \frac{1}{1 + e^{-w}},$$

so

$$P(I_i(z) = 1|x_i) = \frac{e^{x'_i \beta}}{1 + e^{x'_i \beta}}.$$

At one level, there’s little more to say; it really *is* that simple. The likelihood function can be derived, and the model can be immediately estimated by numerical maximization of the likelihood function.

But an alternative latent variable formulation yields useful insights. In particular, consider a latent variable, y_t^* , where

$$y_i^* = x'_i \beta + \varepsilon_i$$

$$\varepsilon_i \sim \text{logistic}(0, 1),$$

and let $I_i(z)$ be $I_i(y_i^* > 0)$, or equivalently, $I_i(\varepsilon > -x_i'\beta)$. Interestingly, this is the logit model. To see this, note that

$$\begin{aligned} E(I_i(y_i^* > 0)|x_i) &= P((y_i^* > 0)|x_i) = P(\varepsilon_i > -x_i'\beta) \\ &= P(\varepsilon_i < x_i'\beta) \text{ (by symmetry of the logistic density of } \varepsilon) \\ &= \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}, \end{aligned}$$

where the last equality holds because the logistic density has cdf is $e^w/(1+e^w)$.

This way of thinking about the logit DGP – a continuously-evolving latent variable y_i^* with an observed indicator that turns “on” when $y_i^* > 0$ – is very useful. For example, it helps us to think about consumer choice as a function of continuous underlying utility, business cycle regime as a function of continuous underlying macroeconomic conditions, etc.

The latent-variable approach also leads to natural generalizations like ordered logit, to which we now turn.

9.2.2 Ordered Logit

Here we still imagine a continuously-evolving underlying latent variable, but we have a more-refined indicator, taking not just two values, but several (ordered) values. Examples include financial analyst stocks ratings of “buy,” “hold” and “sell”, and surveys that ask about degree of belief in three or more categories ranging from “strongly disagree” through “strongly agree.”

Suppose that there are N **ordered outcomes**. As before, we have a continuously-evolving latent variable,

$$y_i^* = x_i'\beta + \varepsilon_i$$

$$\varepsilon_i \sim logistic(0, 1).$$

But now we have an indicator with a finer gradation:

$$I_i(y_i^*) = \begin{cases} 0 & \text{if } y_i^* < c_1 \\ 1 & \text{if } c_1 < y_i^* < c_2 \\ 2 & \text{if } c_2 < y_i^* < c_3 \\ \vdots & \\ N & \text{if } c_N < y_i^*. \end{cases}$$

We can estimate this **ordered logit** model by maximum likelihood, just as with the standard logit model. Under some assumptions, all interpretation remains the same.

9.2.3 Complications

In logit regression, both the marginal effects and the R^2 are hard to determine and/or interpret directly.

Marginal Effects

Logit marginal effects $\partial E(y|x)/\partial x_i$ are hard to determine directly; in particular, they are not simply given by the β_i 's. Instead we have

$$\frac{\partial E(y|x)}{\partial x_i} = f(x'\beta)\beta_i,$$

where $f(x) = dF(x)/dx$ is the density corresponding the cdf f .¹ So the marginal effect is not simply β_i ; instead it is β_i weighted by $f(x'\beta)$, which depends on all β 's and x 's. However, signs of β 's are the signs of the effects, because f must be positive. In addition, ratios of β 's do give ratios of effects, because the f 's cancel.

$$R^2$$

¹In the leading logit case, $f(x)$ would be the logistic density.

Recall that traditional R^2 for continuous LHS variables is

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}.$$

It's not clear how to define or interpret R^2 when the LHS variable is 0-1, but several variants have been proposed. The two most important are Effron's and McFadden's.

Effron's R^2 is

$$R^2 = 1 - \frac{\sum(y_i - \hat{P}(I_i(z) = 1|x_i))^2}{\sum(y_i - \bar{y}_i)^2}.$$

Effron's R^2 attempts to maintain the R^2 interpretation as variation explained and as correlation between actual and fitted values.

McFadden's R^2 is

$$R^2 = 1 - \frac{\ln \hat{L}_1}{\ln \hat{L}_0},$$

where $\ln \hat{L}_0$ is the maximized restricted log likelihood (only an intercept included) and $\ln \hat{L}_1$ is the maximized unrestricted log likelihood. McFadden's R^2 attempts to maintain the R^2 interpretation as improvement from restricted to unrestricted model.

9.3 Classification and “0-1 Forecasting”

Classification maps probabilities into 0-1 forecasts. The so-called “Bayes classifier” uses a cutoff (“decision boundary”) of .5, which is hardly surprising. That is, we predict 1 when $\text{logit}(x'\beta) > 1/2$. Note, however, that that's the same as predicting 1 when $x'\beta > 0$. If there are 2 RHS variables (potentially plus an intercept), then the condition $x'\beta > 0$ defines a line in \mathbb{R}^2 . Points on one side will be classified as 0, and points on the other side will be classified as 1. That line is the decision boundary.

We can also have non-linear decision boundaries. Suppose for example

that that x vector contains not only x_1 and x_2 , but also x_1^2 and x_2^2 . Now the condition $x'\beta > 0$ defines a circle in \mathbb{R}^2 . Points inside will be classified as 0, and points outside will be classified as 1. The circle is the decision boundary.

9.4 Exercises, Problems and Complements

1. (Logit and Ordered-Logit Situations)

In the chapter we gave several examples where logit or ordered-logit modeling would be appropriate.

- a. Give three additional examples where logit modeling would be appropriate. Why?
- b. Give three additional examples where ordered-logit modeling would be appropriate. Why?

2. (The Logistic Squashing Function)

We used the logistic function throughout this chapter. In particular, it is the foundation on which the logit model is built.

- a. What is the logistic function? Write it down precisely.
- b. From where does the logistic function come?
- c. Verify that the logistic function is a legitimate squashing function. That is, verify that it is monotone increasing, with $\lim_{w \rightarrow \infty} F(w) = 1$ and $\lim_{w \rightarrow -\infty} F(w) = 0$.

3. (The Logit Likelihood Function)

Consider the logit model (9.1). It is more formally called a **binomial logit model**, in reference to its two outcome categories.

- a. Derive the likelihood function. (Hint: Consider the binomial structure.)

- b. Must the likelihood be maximized numerically, or is an analytic formula available?
4. (Logit as a Linear Model for Log Odds)

The **odds** $O(I_i(z) = 1|x_i)$ of an event z are just a simple transformation of its probability

$$O(I_i(z) = 1|x_i) = \frac{P(I_i(z) = 1|x_i)}{1 - P(I_i(z) = 1|x_i)}.$$

Consider a linear model for log odds

$$\ln \left(\frac{P(I_i(z) = 1|x_i)}{1 - P(I_i(z) = 1|x_i)} \right) = x'_i \beta.$$

Solving the log odds for $P(I_i(z) = 1|x_i)$ yields the logit model,

$$P(I_i(z) = 1|x_i) = \frac{1}{1 + e^{-x'_i \beta}} = \frac{e^{x'_i \beta}}{1 + e^{x'_i \beta}}.$$

Hence the logit model is simply a linear regression model for log odds.

A full statement of the model is

$$y_i \sim Bern(p_i)$$

$$\ln \left(\frac{p_i}{1 - p_i} \right) = x'_i \beta.$$

5. (Probit and GLM Squashing Functions)

Other squashing functions are sometimes used for binary-response regression.

- a. In the **probit model**, we simply use a different squashing function to keep probabilities in the unit interval. $F(\cdot)$ is the standard normal

cumulative density function (cdf), so the model is

$$P(I_i(z) = 1|x_i) = \Phi(x'_i\beta),$$

where $\Phi(x) = P(z \leq x)$ for $N(0, 1)$ random variable z .

- b. More exotic, but equally simple, squashing functions have also been used. Almost all (including those used with logit and probit) are special cases of those allowed in the **generalized linear model (GLM)**, a flexible regression framework with uses far beyond just binary-response regression. In the GLM,

$$E(y_i|x_i) = G^{-1}(x'_i\beta),$$

and very wide classes of “**link functions**” G can be entertained.

6. (Multinomial Models)

In contrast to the binomial logit model, we can also have more than two categories (e.g., what transportation method will I choose to get to work: Private transportation, public transportation, or walking?), and use **multinomial logit**.

7. (Other Situations/Mechanisms Producing Limited Dependent Variables)

Situations involving censoring or counts also produce limited dependent variables.

- a. Data can be censored by definition (e.g. purchases can't be negative). For example, we might see only y_i , where $y_i = y_i^*$ if $y_i^* \geq 0$, and 0 otherwise, and where

$$y_i^* = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

This is the framework in which the **Tobit model** works.

- b. Data can be censored due to sample selection, for example if income is forecast using a model fit only to high-income people.
- c. “Counts” (e.g., points scored in hockey games) are automatically censored, as they must be in the natural numbers, 1, 2, 3...

Chapter 10

Causal Estimation

In this chapter we distinguish between the predictive modeling perspective (which we have adopted so far) and what we will call the *causal estimation* perspective. Both are tremendously important in econometrics. We will investigate the properties of OLS from each perspective. It turns out that much hinges on the validity of IC2.1, which, as you recall from Chapter 3, says that X and ε are independent, and which we have so far not discussed.¹ Roughly, it turns out that from the predictive modeling perspective everything remains fine asymptotically even if IC2.1 fails (which is why we have not yet had reason to discuss it): $\hat{\beta}_{OLS}$ is still consistent in an appropriate sense and asymptotically normally distributed. But from the causal estimation perspective disasters occur if IC2.1 fails: $\hat{\beta}_{OLS}$ is not even consistent, so that inference is potentially severely distorted even in arbitrarily large samples.

Of course none of this makes sense yet, as we have yet introduced the causal estimation perspective. We now do so.

¹The independence assumption can be weakened somewhat, but we will not pursue that here.

10.1 Predictive Modeling vs. Causal Estimation

Here we distinguish “predictive modeling” from “causal estimation”, or “non-causal prediction” from “causal prediction”, or “conditional expectation estimation” from “partial derivative estimation”.

A major goal in econometrics, as we have emphasized thus far, is predicting y . In the language of estimation, the question is “If a new person i arrives with covariates x_i , what is my minimum-MSE estimate of her y_i ?”. So we are estimating a conditional mean $E(y|x)$. That is the domain of predictive modeling.

Sometimes another goal in econometrics is predicting the effects of exogenous “treatments” or “interventions” or “policies”. Phrased in the language of estimation, the question is “If I intervene and give someone a certain treatment ∂x , what is my minimum-MSE estimate of her ∂y ?”. So we are estimating the partial derivative $\partial y / \partial x$, which in general is very different from estimating a conditional expectation $E(y|x)$. That is the domain of causal estimation.

Related, it is important to note the distinction between what we will call “consistency for a predictive effect” and “consistency for a treatment effect.” In large samples, under very general conditions, the relationship estimated by running $y \rightarrow c, x$ is useful for predicting y given an observation on x (“consistency for a predictive effect”). But it may or may not be useful for determining the effect on y of an exogenous shift in x (“consistency for a treatment effect”). The two types of consistency coincide under the IC, but they diverge when IC2.1 fails.

10.1.1 Predictive Modeling and P-Consistency

Consider a standard linear regression setting with K regressors. Assuming quadratic loss, the predictive risk of a parameter configuration β is

$$R(\beta) = E(y - x'\beta)^2.$$

Let B be a set of β 's and let $\beta^* \in B$ minimize $R(\beta)$. We will say that $\hat{\beta}$ is *consistent for a predictive effect* (“P-consistent”) if $\text{plim} R(\hat{\beta}) = R(\beta^*)$; that is, if

$$\left(R(\hat{\beta}) - R(\beta^*) \right) \rightarrow_p 0.$$

Hence in large samples $\hat{\beta}$ provides a good way to predict y for any hypothetical x : simply use $x'\hat{\beta}$. *OLS is effectively always P-consistent; we require almost no conditions of any kind!* P-consistency is effectively induced by the minimization problem that defines OLS, as the minimum-MSE predictor is the conditional mean.

10.1.2 Causal Estimation and T-Consistency

Consider a standard linear regression setting with K regressors. We will say that an estimator $\hat{\beta}$ is *consistent for a treatment effect* (“T-consistent”) if $\text{plim} \hat{\beta}_k = \partial E(y|x)/\partial x_k, \forall k = 1, \dots, K$; that is, if

$$\left(\hat{\beta}_k - \frac{\partial E(y|x)}{\partial x_k} \right) \rightarrow_p 0, \quad \forall k = 1, \dots, K.$$

Hence in large samples $\hat{\beta}_k$ provides a good estimate of the effect on y of a one-unit “treatment” or “intervention” performed on x_k . OLS is T-consistent under the IC including IC2.1. *OLS is generally not T-consistent without IC2.1.*

10.1.3 Correlation vs. Causality, and P-Consistency vs. T-Consistency

The distinction between P-consistency and T-consistency is intimately related to the distinction between correlation and causality. As is well known, correlation does not imply causality! As long as x and y are correlated, we can exploit the correlation (as captured in $\hat{\beta}_{LS}$ from the regression $y \rightarrow x$) very generally to predict y given knowledge of x . That is, there will be a nonzero “predictive effect” of x knowledge on y . But nonzero correlation doesn’t necessarily tell us anything about the causal “treatment effect” of x treatments on y . That requires the ideal conditions, and in particular IC2.1. Even if there is a non-zero predictive effect of x on y (as captured by $\hat{\beta}_{LS}$), there may or may not be a nonzero treatment effect of x on y , and even if there is a nonzero treatment effect it will generally not equal the predictive effect.

So, assembling things, we have that:

1. P-consistency is consistency for a non-causal predictive effect. It is almost trivially easy to obtain, by virtue of the objective function that OLS optimizes.
2. T-consistency is consistency for a causal predictive effect. It is quite difficult to obtain reliably, because it requires IC2.1, which may fail for a variety of reasons.

Thus far we have sketched why P-consistency holds very generally, whereas we have simply asserted that T-consistency is much more difficult to obtain and relies critically on IC2.1. We now sketch *why* T-consistency is much more difficult to obtain and relies critically on IC2.1.

Consider the following example. Suppose that y and z are in fact causally *unrelated*, so that the true treatment effect of z on y is 0 by construction. But suppose that z is correlated with an unobserved variable x that *does* cause y . Then y and z will be correlated due to their joint dependence on

x , and that correlation can be used to predict y given z , despite the fact that, by construction, z treatments (interventions) will have no effect on y . Clearly this sort of situation – omission of a relevant variable correlated with an included variable – may happen commonly, and it violates IC2.1. In the next section we sketch several situations that produce violations of IC2.1, beginning with an elaboration on the above-sketched omitted-variables problem.

10.2 Reasons for Failure of IC2.1

IC2.1 can fail for several reasons; we now sketch some of the most important.

10.2.1 Omitted Variables

Omission of relevant variables is a clear violation of the ideal conditions, insofar as the IC explicitly state that the fitted model matches the DGP. But there is a deeper way to see why and when omitted variables cause trouble, and when they don't, and it involves IC2.1.

Suppose that the DGP is

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

with all IC satisfied, but that we incorrectly regress $y \rightarrow z$, where $\text{corr}(x, z) > 0$. Clearly we'll estimate a positive causal effect of z on y , in large as well as small samples, even though it's completely spurious and would vanish if x had been included in the regression. The positive bias arises because in our example $\text{corr}(x, z) > 0$; in general the sign of the bias could go either way, depending on the sign of the correlation. We speak of “**omitted variable bias**”.

In this example the problem is that condition IC2.1 is violated in the regression $y \rightarrow z$, because the disturbance is correlated with the regressor.

$\hat{\beta}_{OLS}$ is P-consistent, as always. But it's not T-consistent, because the omitted variable x is lurking in the disturbance of the fitted regression $y \rightarrow z$, which makes the disturbance correlated with the regressor (i.e., IC2.1 fails in the fitted regression). The fitted OLS regression coefficient on z will be non-zero and may be very large, even asymptotically, despite the fact that the true causal impact of z on y is zero by construction. The OLS estimated coefficient is reliable for predicting y given z , but not for assessing the effects on y of *treatments* in z .

10.2.2 Measurement Error

Suppose as above that the DGP is

$$y = x + \varepsilon,$$

with all IC satisfied, but that we can't measure x accurately. Instead we measure

$$x^m = x + v.$$

Think of v as an *iid* measurement error with variance σ_v^2 . (Assume that it is also independent of ε .) Sometimes x^m is called a “proxy variable”.

Clearly, as σ_v^2 gets larger relative to σ_x^2 , the fitted regression $y \rightarrow x^m$ is progressively less able to identify the true relationship, as the measured regressor is polluted by progressively more noise. In the limit as $\sigma_v^2 / \sigma_x^2 \rightarrow \infty$, it is impossible, and $\hat{\beta}_{OLS} = 0$. In any event, $\hat{\beta}_{OLS}$ is biased toward zero, in small as well as large samples. We speak of the “**errors in variables problem**”, or **measurement-error bias**.

So far we have motivated measurement-error bias intuitively. Formally, it arises from violation of IC2.1. To see this, note that we have

$$y = x + \varepsilon$$

$$\begin{aligned}
&= (x^m - v) + \varepsilon \\
&= x^m + (\varepsilon - v) \\
&= x^m + \nu.
\end{aligned}$$

In the fitted regression $y \rightarrow x^m$, the disturbance ν is clearly negatively correlated with the included regressor x^m .

In more complicated cases involving multiple regression with some variables measured with error, some measured without error, possibly correlated measurement errors, etc., things quickly get very complicated. Bias still exists, but it is difficult to ascertain its direction.

10.2.3 Simultaneity

Suppose that y and x are jointly determined, as for example in simultaneous determination of quantity (Q) and price (P) in market equilibrium. Then we may write

$$Q = \beta P + \varepsilon,$$

but note that the ε shocks affect not only Q but also P . (Any shock to Q is also a shock to P , since Q and P are determined jointly by supply and demand!) That is, ε is correlated with P , violating IC2.1.

10.3 Confronting Failures of IC2.1

10.3.1 Controlling for Omitted Variables

The remedy for omitted relevant variables is simple in principle: start including them!² Let's continue with our earlier example. The DGP is

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

²In the lingo, the problem is that we failed to “control” for, or include, the omitted variable.

with all IC satisfied, but we incorrectly regress $y \rightarrow z$, where $\text{corr}(x, z) > 0$. We saw that we estimate a positive causal effect of z on y , in large as well as small samples, even though it's completely spurious, due to the failure of IC2.1 in the fitted model. Now consider instead controlling for x as well. In the OLS regression $y \rightarrow x, z$, all IC are satisfied, so z will get a zero coefficient asymptotically, but x will get a β coefficient, which will be accurate for the predictive effect of x on y (of course – OLS is always consistent for predictive effects) *and* the treatment effect of x on y . (Remember, the predictive and treatment effects coincide under the IC.)

Of course the recipe “start including omitted variables” is easier said than done. We simply may never know about various omitted variables, or we may suspect them but be unable to measure them. In a wage equation, for example, in addition to the usual regressors like education and experience, we might want to include “ability” – but how? In any event it’s important to use all devices available, from simple introspection to formal theory, in an attempt to assemble an adequate set of controls.

10.3.2 Instrumental Variables

Consider the simple regression

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

Following standard usage, let us call a regressor x that satisfies IC2.1 “exogenous” (i.e., x is uncorrelated with ε), and a regressor x that fails IC2.1 “endogenous”.

If x is endogenous it means that IC2.1 fails, so we need to do something about it. One solution is to find an acceptable “instrument” for x . An instrument $inst$ is a new regressor that is both exogenous (uncorrelated with ε) and “strong” or “relevant” (highly-correlated with x). IV estimation proceeds as follows: (0) Find an exogenous and relevant $inst$, (1) In a first-stage regres-

sion run $x \rightarrow inst$ and get the fitted values $\hat{x}(inst)$, and (2) in a second-stage regression run $y \rightarrow \hat{x}(inst)$. So in the second-stage regression we replace the endogenous x with our best linear approximation to x based on $inst$, namely $\hat{x}(inst)$. This second-stage regression does not violate IC2.1 – $inst$ is exogenous so $\hat{x}(inst)$ must be as well.

In closing this section, we note that just as the prescription “start including omitted variables” for a specific violation of IC2.1 (omitted variables) may at first appear vapid, so too might the general prescription for violations of IC2.1, “find a good instrument”. But plausible, if not unambiguously “good”, instruments can often be found. Economists generally rely on a blend of introspection and formal economic theory. Economic theory requires many assumptions, but if the assumptions are plausible, then theory can be used to suggest plausible instruments.

10.3.3 Randomized Controlled Trials (RCT’s) and Their Approximation

Randomized controlled trials (RCT’s), or randomized experiments, are the gold standard in terms of assessing causal effects. The basic idea is to randomly select some people into a treatment group, and some into a non-treatment group, and to estimate the mean difference (the “average treatment effect.”)

In a development economics context, for example, you might be interested in whether adoption of a new fertilizer enhances crop yield. You can’t just introduce it, see who adopts and who doesn’t, and then regress yield on an adoption choice dummy,

$$yield \rightarrow c, choice, \quad (10.1)$$

because those farms that chose to use the new fertilizer may have done so because their characteristics made them particularly likely to benefit from it. Instead you’d need an instrument for adoption.

An RCT effectively creates an instrument for *choice* in regression (10.1), by randomizing. You randomly select some farms for adoption (treatment) and some not (control), and you inspect the difference between yields for the two groups. More formally, for the firms in the experiment you could run

$$\text{yield} \rightarrow c, \text{treatment}. \quad (10.2)$$

The OLS estimate of c is the mean yield for the non-adoption (control) group, and the OLS estimate of the coefficient on the treatment dummy is the mean enhancement from adoption (treatment). You can test its significance with the usual t test. The randomization guarantees that the regressor in (10.2) is exogenous, so that IC2.1 is satisfied.

The key insight bears repeating: RCT's, if successfully implemented, guarantee that IC2.1 is satisfied.

But of course there's no free lunch, and there are various issues and potential problems with "successful implementation of an RCT" in all but the simplest cases (like the example above), just as there are many issues and potential problems with "finding a strong and exogenous instrument". In what follows we sketch several RCT variations, extensions, and issues.

Regression Discontinuity Designs (RDD's)

RCT's can be expensive and wasteful when estimating the efficacy of a treatment, as many people who don't need treatment will be randomly assigned treatment anyway. Hence alternative experimental designs are often entertained. A leading example is the "regression discontinuity design" (RDD).

To understand the RDD, consider a famous scholarship example. You want to know whether receipt of an academic scholarship causes enhanced academic performance among top academic performers. You can't just regress academic performance on a scholarship receipt dummy, because recipients are likely to be stong academic performers even without the scholarship.

The question is whether scholarship receipt causes *enhanced* performance for already-strong performers.

You could do an RCT, but in an RCT approach you're going to give lots of academic scholarships to lots of randomly-selected people, many of whom are not strong performers. That's wasteful.

An RDD design is an attractive alternative. You give scholarships only to those who score above some threshold in the scholarship exam, and compare the performances of students who scored just above and below the threshold. In the RDD you don't give any scholarships to weak performers. So it's not wasteful.

Notice how the RDD effectively attempts to approximate a controlled experiment. People just above and below the scholarship threshold are basically the same in terms of academic talent – the only difference is that one group gets the scholarship and one doesn't. The RCT does the controlled experiment directly; it's statistically efficient but can be wasteful. The RDD does the controlled experiment indirectly; it's statistically less efficient but also less wasteful.

Propensity-Score Matching

In the scholarship exam example, you could do an RCT, but it's not attractive for certain reasons. Sometimes it's even worse – an RCT is simply infeasible. Consider, for example, estimating the causal effect of college education on subsequent earnings. As usual, we don't just want to compare earnings of college grads and non college grads (regress earnings on a college dummy). College grads may earn more for many reasons other than college – perhaps higher intelligence, more supportive family, etc. We need to control for such things. An RCT works in principle but is infeasible in practice – you'd have to randomly select a large group of high school students, randomly send part to college, and prohibit the rest from attending college, and follow their

outcomes for decades.

The propensity score approach attempts to approximate an RCT by estimating a logit or similar model for the probability of attending college. The idea is to compare people with similar propensity scores, some of whom went to college and some of whom didn't, to control for everything other than college vs. non-college.

Causal effect estimation based on propensity scores is one example of the general idea of *matching estimation*.

Event Studies (“Synthetic Controls”)

In many time series situations we never intervene and do an experiment; instead we are in the world of “observational studies”, trying to make causal inferences from the historical record. One would of course like to watch the realizations of two universes, the one that actually occurred with some treatment applied, and a parallel counterfactual universe without the treatment applied. That's not possible in general, but we can approximate the comparison by estimating a model on pre-treatment data and using it to predict what would have happened in the absence of the treatment, and comparing it to what happened in the real data with the treatment.

“Treatment” sounds like active intervention, but again, the treatment is usually passive in event study contexts. Consider the following example. We want to know the effect of a new gold discovery on stock returns of a certain gold mining firm. We can't just look at the firm's returns on the announcement day, because daily stock returns vary greatly for lots of reasons. Event studies proceed by (1) specifying and estimating a model for the object of interest (in this case a firm's daily stock returns) over the pre-event period, using only pre-event data, $1, \dots, T$ (in this case pre-announcement data), (2) using the model to predict into the post-event period $T+1, T+2, \dots$, and (3) comparing the post-event forecast to the post-event realization.

Internal Validity and its Problems

Successfully-implemented RCT's are generally "internally valid" for *something*; that is, they produce credible estimates of treatment effect for the precise experiment performed and situation studied. But lots of things can go wrong, casting doubt on internal validity. A short list, with many items inter-related, includes:

- Is the randomization really credible?
- Are the sample sizes large enough?
- Are there placebo effects?
- Who knows what about who is treated and who is not? (Single-blind RCT, double blind RCT, open RCT, ...)
- Are the behaviors of the groups evolving over time, perhaps due to interaction with each other ("spillovers")?
- Are people entering and/or leaving the study at different times? If so, when and why?
- Might experimenters coddle the treatment group in various ways in attempts to raise the likelihood of "significant" effects of their treatments?
- Might "negative" or "insignificant" results be discarded, and hence only positive and significant results published? (The "file-drawer problem".)

External Validity and its Problems

Even if an RCT study is internally valid, its results may not generalize to other populations or situations. That is, even if internally valid, it may not be *externally* valid.

Consider, for example, a study of the effects of fertilizer on crop yield done for region X in India during a heat wave. Even if successfully randomized, and hence internally valid, the estimated treatment effect is for the effects of fertilizer on crop yield in region X *during a heat wave*. The results do not necessarily generalize – and in this example surely do not generalize – to

times of “normal” weather, even in region X. And of course, for a variety of reasons, they may not generalize to regions other than X, even in heat waves.

Hence, even if an RCT is internally valid, there is no guarantee that it is externally valid, or “extensible”. That is, there is no guarantee that its results will hold in other cross sections and/or time periods.

10.4 Exercises, Problems and Complements

1. Omitted variable bias.

How might you assess whether a regression suffers from omitted variable bias?

2. Included irrelevant variables.

Another violation of the full ideal conditions is **inclusion of irrelevant variables**. Fortunately the effects are minor; some degrees of freedom are wasted, but otherwise there’s no problem. How would you assess whether an included variable in a regression is irrelevant? Whether a set of included variables is irrelevant?

3. Tradeoffs between instrument exogeneity and relevance.

We want instruments that are both exogenous (uncorrelated with ε) and “strong” or “relevant” (highly-correlated with x). There is a trade-off. For example, an exogenous but weakly-relevant instrument might nevertheless be valuable, as might a relevant but “slightly-endogenous” instrument. The former instrument may produce an IV estimator that is consistent but high-variance, whereas the latter may produce an estimator that is (slightly) inconsistent but low-variance.

4. More on IC2.1 in cross-section and the time-series cases.

In cross sections we wrote IC2 as “ ε_i independent of x_i ”. We did not yet have occasion to state IC2 in time series, since we never returned

to IC2 until this chapter. In time series it becomes “ ε_t independent of x_t, x_{t-1}, \dots ”.

5. Instruments in time-series environments.

In time-series contexts with x_t serially correlated, an obvious instrument for x_t is its *lag*, x_{t-1} . Due to the serial correlation in x , x_{t-1} is correlated with x_t , yet x_{t-1} *can't* be correlated with ε_t , which is independent over time and hence uncorrelated with x_{t-1} .

Part III

Time Series

Chapter 11

Trend and Seasonality

The time series that we want to model vary over time, and we often mentally attribute that variation to unobserved underlying components related to **trend** and **seasonality**.

11.1 Linear Trend

Trend involves slow, long-run, evolution in the variables that we want to model and forecast. In business, finance, and economics, for example, trend is produced by slowly evolving preferences, technologies, institutions, and demographics. We'll focus here on models of **deterministic trend**, in which the trend evolves in a perfectly predictable way. Deterministic trend models are tremendously useful in practice.

Linear trend is a simple linear function of time,

$$Trend_t = \beta_1 + \beta_2 TIME_t.$$

The indicator variable *TIME* is constructed artificially and is called a “time trend”, or “time indicator”, or “**time dummy**.¹ *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, $TIME_t = t$, so that the *TIME* variable simply indicates the time. β_1 is the **intercept**; it's the

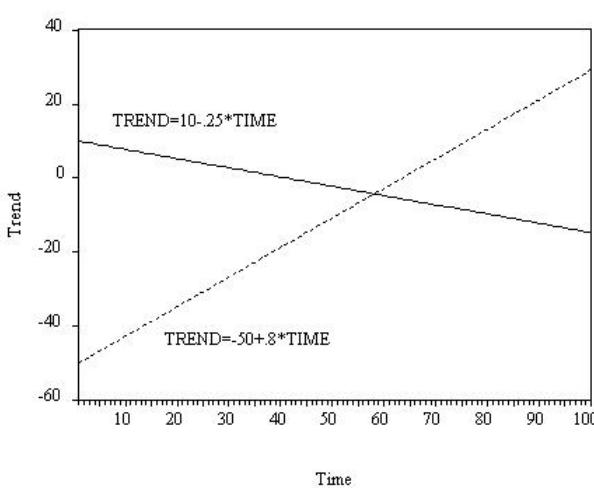


Figure 11.1: Various Linear Trends

value of the trend at time $t=0$. β_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 β_1 , the steeper the trend's slope. In Figure 11.1, for example, we show two linear trends, one increasing and one decreasing. The increasing trend has an intercept of $\beta_1 = -50$ and a slope of $\beta_2 = .8$, whereas the decreasing trend has an intercept of $\beta_1 = 10$ and a gentler absolute slope of $\beta_2 = -.25$.

In business, finance, and economics, linear trends are typically increasing, corresponding to growth, but they don't have to increase. In recent decades, for example, male labor force participation rates have been falling, as have the times between trades on stock exchanges. Moreover, in some cases, such as records (e.g., world records in the marathon), trends are decreasing by definition.

Estimation of a linear trend model (for a series y , say) is easy. First we need to create and store on the computer the variable $TIME$. Fortunately we don't have to type the $TIME$ values (1, 2, 3, 4, ...) in by hand; in most good software environments, a command exists to create the trend automatically. Then we simply run the least squares regression $y \rightarrow c, TIME$.

11.2 Non-Linear Trend

Nonlinearity can be important in time series just as in cross sections, but there is a special case of key importance in time series: nonlinear *trend*. Here we introduce it.

11.2.1 Quadratic Trend

Sometimes trend appears non-linear, 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.

We can allow for gentle curvature by including not only $TIME$, but also $TIME^2$,

$$Trend_t = \beta_1 + \beta_2 TIME_t + \beta_3 TIME_t^2.$$

This is called **quadratic trend**, because the trend is a quadratic function of $TIME$.¹ Linear trend emerges as a special (and potentially restrictive) case when $\beta_3 = 0$.

A variety of different non-linear quadratic trend shapes are possible, depending on the signs and sizes of the coefficients; we show several in Figure 11.2. In particular, if $\beta_2 > 0$ and $\beta_3 > 0$ as in the upper-left panel, the trend is monotonically, but non-linearly, increasing. Conversely, if $\beta_2 < 0$ and $\beta_3 < 0$, the trend is monotonically decreasing. If $\beta_2 < 0$ and $\beta_3 > 0$ the trend has a U shape, and if $\beta_2 > 0$ and $\beta_3 < 0$ the trend has an inverted U shape. 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”.

Estimating quadratic trend models is no harder than estimating linear trend models. We first create $TIME$ and its square; call it $TIME2$, where

¹Higher-order **polynomial trends** are sometimes entertained, but it's important to use low-order polynomials to maintain smoothness.

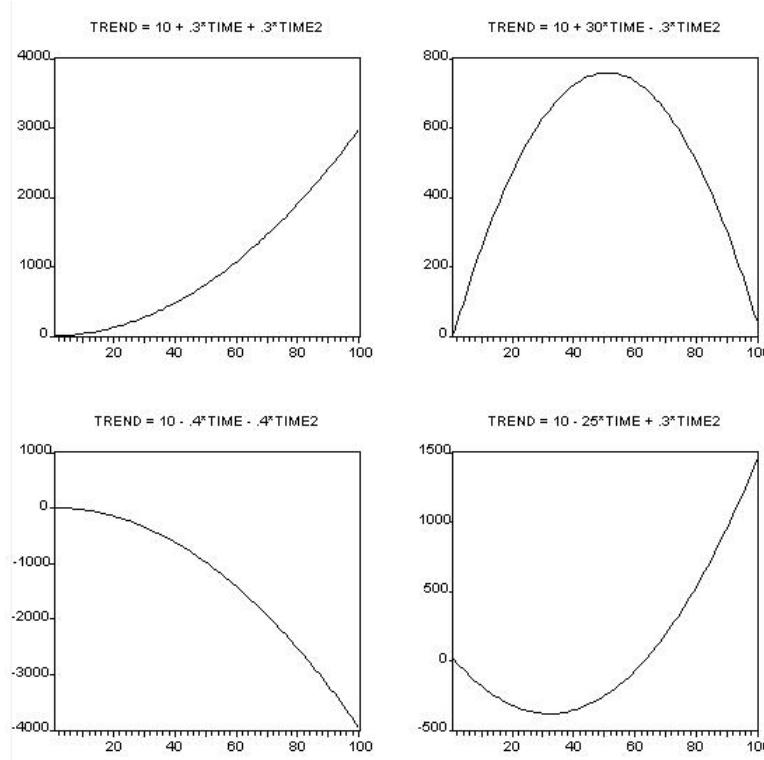


Figure 11.2: Various Quadratic Trends

$TIME2_t = TIME_t^2$. Because $TIME = (1, 2, \dots, T)$, $TIME2 = (1, 4, \dots, T^2)$. Then we simply run the least squares regression $y \rightarrow c, TIME, TIME2$. Note in particular that although the quadratic is a non-linear function, it is linear in the variables $TIME$ and $TIME2$.

11.2.2 Exponential Trend

The insight that exponential growth is non-linear in levels but linear in logarithms takes us to the idea of **exponential trend**, or **log-linear trend**, which is very common in business, finance and economics.²

Exponential trend is common because economic variables often display roughly constant real growth rates (e.g., two percent per year). If trend is

²Throughout this book, logarithms are *natural* (base e) logarithms.

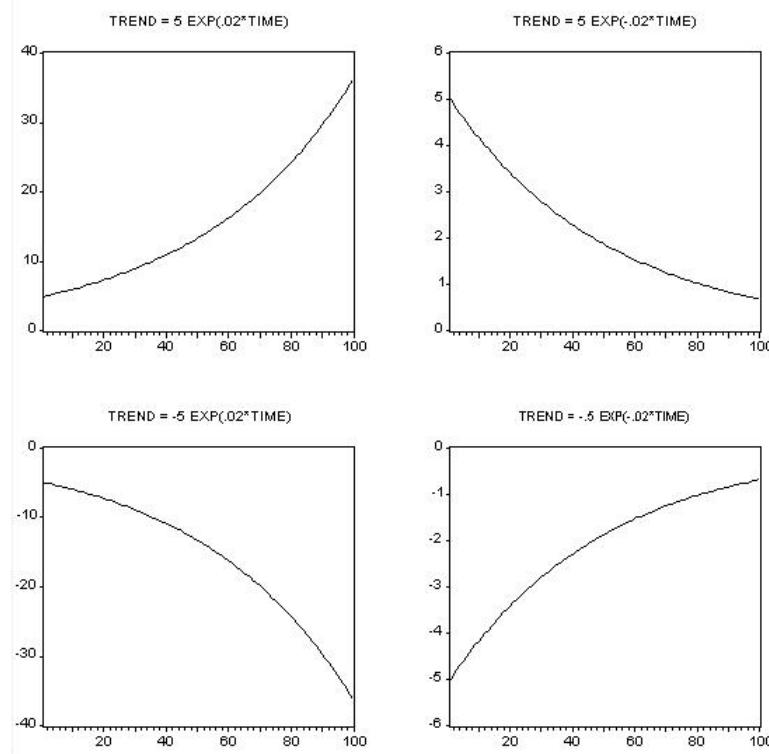


Figure 11.3: Various Exponential Trends

characterized by constant growth at rate β_2 , then we can write

$$Trend_t = \beta_1 e^{\beta_2 TIME_t}.$$

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

$$\ln(Trend_t) = \ln(\beta_1) + \beta_2 TIME_t. \quad (11.1)$$

Thus, $\ln(Trend_t)$ is a linear function of time.

In Figure 11.3 we show the variety of exponential trend shapes that can be obtained depending on the parameters. Depending on the signs and sizes of the parameter values, exponential trend can achieve a variety of patterns, increasing or decreasing at increasing or decreasing rates.

Although the exponential trend model is non-linear, we can estimate it by

simple least squares regression, because it is linear in logs. We simply run the least squares regression, $\ln y \rightarrow c, TIME$. Note that because the intercept in equation (11.1) is *not* β_1 , but rather $\ln(\beta_1)$, we need to exponentiate the estimated intercept to get an estimate of β_1 . 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 . This is necessary, for example, for appropriately comparing fitted values or residuals (or statistics based on residuals, like *AIC* and *SIC*) from estimated exponential trend models to those from other trend models.

It's important to note that, although the same sorts of qualitative trend shapes can sometimes be achieved with quadratic and exponential trend, there are subtle differences between them. The non-linear trends in some series are well approximated by quadratic trend, while the trends in other series are better approximated by exponential trend. Ultimately it's an empirical matter as to which is best in any particular application.

11.2.3 Non-Linearity in Liquor Sales Trend

We already fit a non-linear (exponential) trend to liquor sales, when we fit a linear trend to log liquor sales. But it still didn't fit so well.

We now examine quadratic trend model (again in logs). The log-quadratic trend estimation results appear in Figure 11.4. Both *TIME* and *TIME2* are highly significant. The adjusted R^2 for the log-quadratic trend model is 89%, higher than for the the log-linear trend model. As with the log-linear trend model, the Durbin-Watson statistic provides no evidence against the hypothesis that the regression disturbance is white noise. The residual plot (Figure 11.5) shows that the fitted quadratic trend appears adequate, and that it increases at a decreasing rate. The residual plot also continues to indicate obvious residual seasonality. (Why does the Durbin-Watson not detect it?)

Dependent Variable: LSALES				
Method: Least Squares				
Date: 08/08/13 Time: 08:53				
Sample: 1987M01 2014M12				
Included observations: 336				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.231269	0.020653	301.7187	0.0000
TIME	0.007768	0.000283	27.44987	0.0000
TIME2	-1.17E-05	8.13E-07	-14.44511	0.0000
R-squared	0.903676	Mean dependent var	7.096188	
Adjusted R-squared	0.903097	S.D. dependent var	0.402962	
S.E. of regression	0.125439	Akaike info criterion	-1.305106	
Sum squared resid	5.239733	Schwarz criterion	-1.271025	
Log likelihood	222.2579	Hannan-Quinn criter.	-1.291521	
F-statistic	1562.036	Durbin-Watson stat	1.754412	
Prob(F-statistic)	0.000000			

Figure 11.4: Log-Quadratic Trend Estimation

In Figure 11.6 we show the results of regression on quadratic trend and a full set of seasonal dummies. The trend remains highly significant, and the coefficients on the seasonal dummies vary significantly. The adjusted R^2 rises to 99%. The Durbin-Watson statistic, moreover, has greater ability to detect residual serial correlation now that we have accounted for seasonality, and it sounds a loud alarm. The residual plot of Figure 11.7 shows no seasonality, as the model now accounts for seasonality, but it confirms the Durbin-Watson statistic's warning of serial correlation. The residuals appear highly persistent.

There remains one model as yet unexplored, exponential trend fit to $LSALES$. We do it by NLS (why?) and present the results in Figure ***. Among the linear, quadratic and exponential trend models for $LSALES$, both SIC and AIC clearly favor the quadratic.

- Exogenously-specified break in log-linear trend model
- Endogenously-selected break in log-linear trend model
- SIC for best broken log-linear trend model vs. log-quadratic trend model

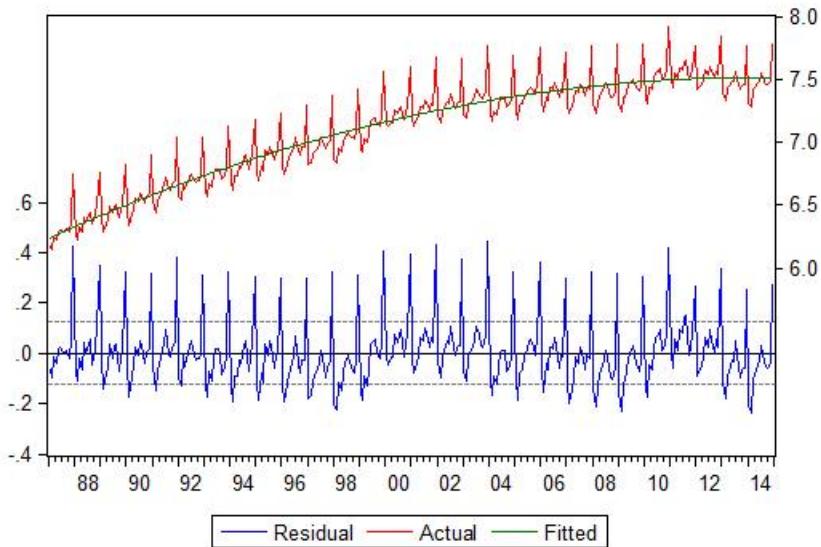


Figure 11.5: Residual Plot, Log-Quadratic Trend Estimation

11.3 Seasonality

In the last section we focused on the trends; now we'll focus on **seasonality**. A seasonal pattern is one that repeats itself every year.³ The annual seasonal repetition can be exact, in which case we speak of **deterministic seasonality**. Here we focus exclusively on deterministic seasonality models.

Seasonality arises from links of technologies, preferences and institutions to the calendar. The weather (e.g., daily high temperature) is a trivial but very important seasonal series, as it's always hotter in the summer than in the winter. Any technology that involves the weather, such as production of agricultural commodities, is likely to be seasonal as well.

Preferences may also be linked to the calendar. Consider, for example, gasoline sales. People want to do more vacation travel in the summer, which tends to increase both the price and quantity of summertime gasoline sales, both of which feed into higher current-dollar sales.

³Note therefore that seasonality is impossible, and therefore not an issue, in data recorded once per year, or less often than once per year.

Dependent Variable: LSALES
 Method: Least Squares
 Date: 08/08/13 Time: 08:53
 Sample: 1987M01 2014M12
 Included observations: 336

Variable	Coefficient	Std. Error	t-Statistic	Prob.
TIME	0.007739	0.000104	74.49828	0.0000
TIME2	-1.18E-05	2.98E-07	-39.36756	0.0000
D1	6.138362	0.011207	547.7315	0.0000
D2	6.081424	0.011218	542.1044	0.0000
D3	6.168571	0.011229	549.3318	0.0000
D4	6.169584	0.011240	548.8944	0.0000
D5	6.238568	0.011251	554.5117	0.0000
D6	6.243596	0.011261	554.4513	0.0000
D7	6.287566	0.011271	557.8584	0.0000
D8	6.259257	0.011281	554.8647	0.0000
D9	6.199399	0.011290	549.0938	0.0000
D10	6.221507	0.011300	550.5987	0.0000
D11	6.253515	0.011309	552.9885	0.0000
D12	6.575648	0.011317	581.0220	0.0000
R-squared	0.987452	Mean dependent var	7.096188	
Adjusted R-squared	0.986946	S.D. dependent var	0.402962	
S.E. of regression	0.046041	Akaike info criterion	-3.277812	
Sum squared resid	0.682555	Schwarz criterion	-3.118766	
Log likelihood	564.6725	Hannan-Quinn criter.	-3.214412	
Durbin-Watson stat	0.581383			

Figure 11.6: Liquor Sales Log-Quadratic Trend Estimation with Seasonal Dummies

Finally, social institutions that are linked to the calendar, such as holidays, are responsible for seasonal variation in a variety of series. In Western countries, for example, sales of retail goods skyrocket every December, Christmas season. In contrast, sales of durable goods fall in December, as Christmas purchases tend to be nondurables. (You don't buy someone a refrigerator for Christmas.)

You might imagine that, although certain series are seasonal for the reasons described above, seasonality is nevertheless uncommon. On the contrary, and perhaps surprisingly, seasonality is pervasive in business and economics. Many industrialized economies, for example, expand briskly every fourth quarter and contract every first quarter.

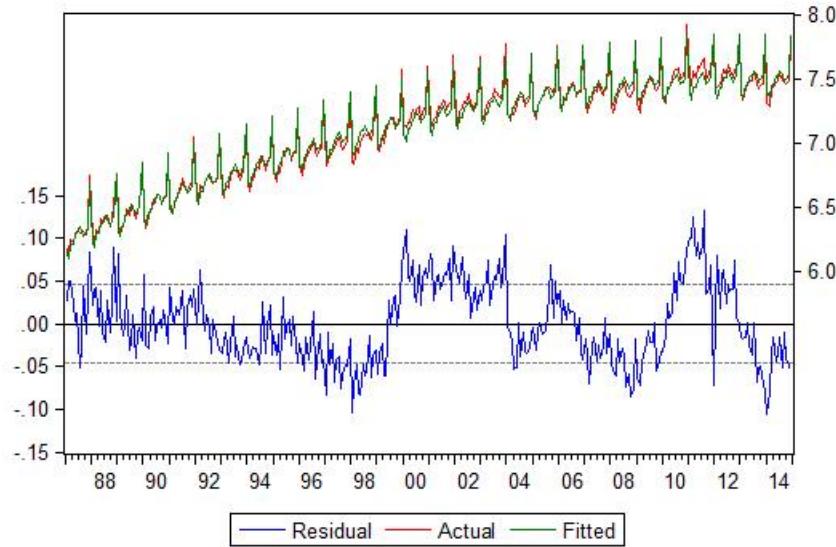


Figure 11.7: Residual Plot, Liquor Sales Log-Quadratic Trend Estimation With Seasonal Dummies

11.3.1 Seasonal Dummies

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, and so on.

The pure seasonal dummy model is

$$\text{Seasonal}_t = \sum_{s=1}^S \gamma_s \text{SEAS}_{ts}$$

$$\text{where } \text{SEAS}_{ts} = \begin{cases} 1 & \text{if observation } t \text{ falls in season } s \\ 0 & \text{otherwise} \end{cases}$$

The SEAS_{ts} variables are called **seasonal dummy variables**. They simply indicate which season we're in.

Operationalizing the model is simple. Suppose, for example, that we have

quarterly data, so that $S = 4$. Then we create four variables⁴:

$$SEAS_1 = (1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, \dots, 0)'$$

$$SEAS_2 = (0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, \dots, 0)'$$

$$SEAS_3 = (0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, \dots, 0)'$$

$$SEAS_4 = (0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, \dots, 1)'.$$

$SEAS_1$ indicates whether we're in the first quarter (it's 1 in the first quarter and zero otherwise), $SEAS_2$ 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 one seasonal dummy is 1, and all others are zero.

To estimate the model for a series y , we simply run the least squares regression,

$$y \rightarrow SEAS_1, \dots, SEAS_S.$$

Effectively, we're just regressing on an intercept, but we allow for a different intercept in each season. Those different intercepts (that is γ_s 's) are called the seasonal factors; they summarize the seasonal pattern over the year, and we often may want to examine them and plot them. In the absence of seasonality, those intercepts are all the same, so we can drop all the seasonal dummies and instead simply include an intercept in the usual way.

In time-series contexts it's often most natural to include a full set of seasonal dummies, without an intercept. But of course we could instead include any $S - 1$ seasonal dummies and an intercept. Then the constant term is the intercept for the omitted season, and the coefficients on the seasonal dummies give the seasonal increase or decrease relative to the omitted season. In no case, however, should we include S seasonal dummies *and* an intercept. Including an intercept is equivalent to including a variable in the regression whose value is always one, but note that the full set of S seasonal dummies sums to a variable whose value is always one, so it is completely redundant.

⁴For illustrative purposes, assume that the data sample begins in Q1 and ends in Q4.

Trend may be included as well. For example, we can account for seasonality and linear trend by running⁵

$$y \rightarrow TIME, SEAS_1, \dots, SEAS_S.$$

In fact, you can think of what we’re doing in this section as a generalization of what we did in the last, in which we focused exclusively on trend. We *still* want to account for trend, if it’s present, but we want to expand the model so that we can account for seasonality as well.

11.3.2 More General Calendar Effects

The idea of seasonality may be extended to allow for more general **calendar effects**. “Standard” seasonality is just one type of calendar effect. Two additional important calendar effects are **holiday variation** and **trading-day variation**.

Holiday variation refers to the fact that some holidays’ dates change over time. That is, although they arrive at approximately the same time each year, the exact dates differ. Easter is a common example. Because the behavior of many series, such as sales, shipments, inventories, hours worked, and so on, depends in part on the timing of such holidays, we may want to keep track of them in our forecasting models. As with seasonality, holiday effects may be handled with dummy variables. In a monthly model, for example, in addition to a full set of seasonal dummies, we might include an “Easter dummy,” which is 1 if the month contains Easter and 0 otherwise.

Trading-day variation refers to the fact that different months contain different numbers of trading days or business days, which is an important consideration when modeling and forecasting certain series. For example, in a monthly forecasting model of volume traded on the London Stock Exchange, in addition to a full set of seasonal dummies, we might include a trading day

⁵Note well that we drop the intercept!

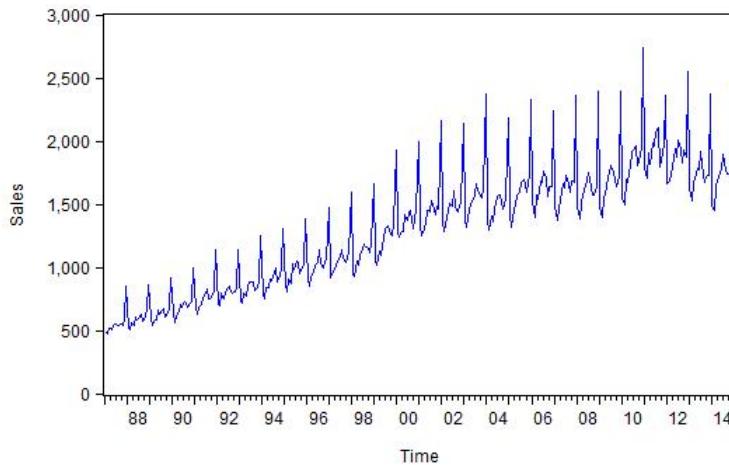


Figure 11.8: Liquor Sales

variable, whose value each month is the number of trading days that month.

More generally, you can model any type of calendar effect that may arise, by constructing and including one or more appropriate dummy variables.

11.4 Trend and Seasonality in Liquor Sales

We'll illustrate trend and seasonal modeling with an application to liquor sales. The data are measured monthly.

We show the time series of liquor sales in Figure 11.8, which displays clear trend (sales are increasing) and seasonality (sales skyrocket during the Christmas season, among other things).

We show log liquor sales in Figure 11.9 ; we take logs to stabilize the variance, which grows over time.⁶ Log liquor sales has a more stable variance, and it's the series for which we'll build models.⁷

Linear trend estimation results appear in Table 11.10. The trend is increasing and highly significant. The adjusted R^2 is 84%, reflecting the fact

⁶The 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.

⁷From this point onward, for brevity we'll simply refer to “liquor sales,” but remember that we've taken logs.

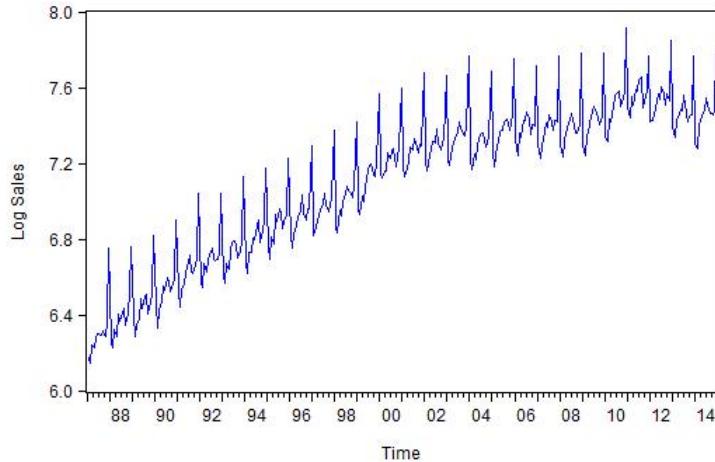


Figure 11.9: Log Liquor Sales

that trend is responsible for a large part of the variation in liquor sales.

The residual plot (Figure 11.11) suggests, however, that linear trend is inadequate. Instead, the trend in log liquor sales appears nonlinear, and the neglected nonlinearity gets dumped in the residual. (We'll introduce nonlinear trend later.) The residual plot also reveals obvious residual seasonality.

In Figure 11.12 we show estimation results for a model with linear trend and seasonal dummies. All seasonal dummies are of course highly significant (no month has average sales of 0), and importantly the various seasonal coefficients in many cases are significantly different from each other (that's the seasonality). R^2 is higher.

In Figure 11.13 we show the corresponding residual plot. The model now picks up much of the seasonality, as reflected in the seasonal fitted series and the non-seasonal residuals. However, it clearly misses nonlinearity in the trend, which therefore appears in the residuals.

In Figure 11.14 we plot the estimated seasonal pattern (the set of 12 estimated seasonal coefficients), which peaks during the winter holidays.

All of these results are crude approximations, because the linear trend is clearly inadequate. We will subsequently allow for more sophisticated

Dependent Variable: LSALES				
Method: Least Squares				
Date: 08/08/13 Time: 08:53				
Sample: 1987M01 2014M12				
Included observations: 336				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.454290	0.017468	369.4834	0.0000
TIME	0.003809	8.98E-05	42.39935	0.0000
R-squared	0.843318	Mean dependent var	7.096188	
Adjusted R-squared	0.842849	S.D. dependent var	0.402962	
S.E. of regression	0.159743	Akaike info criterion	-0.824561	
Sum squared resid	8.523001	Schwarz criterion	-0.801840	
Log likelihood	140.5262	Hannan-Quinn criter.	-0.815504	
F-statistic	1797.705	Durbin-Watson stat	1.078573	
Prob(F-statistic)	0.000000			

Figure 11.10: Linear Trend Estimation

(nonlinear) trends.

11.5 Exercises, Problems and Complements

1. (Mechanics of trend estimation and detrending)

Obtain from the web a quarterly time series of U.S. real GDP in levels, spanning the last forty years, and ending in Q4.

- a. Produce a time series plot and discuss.
- b. Fit a linear trend. Discuss both the estimation results and the residual plot.
- c. Is there any evidence of seasonality in the residuals? Why or why not?
- d. The *residuals* from your fitted model are effectively a linearly **detrended** version of your original series. Why? Discuss.
- 2. (Using model selection criteria to select a trend model)

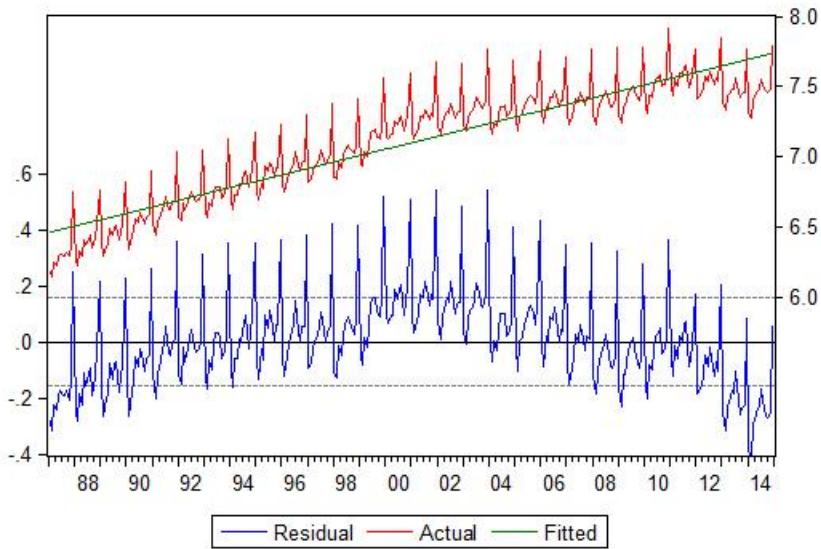


Figure 11.11: Residual Plot, Linear Trend Estimation

You are tracking and forecasting the earnings of a new company developing and applying proprietary nano-technology. The earnings are trending upward. You fit linear, quadratic, and exponential trend models, yielding sums of squared residuals of 4352, 2791, and 2749, respectively. Which trend model would you select, and why?

3. (Seasonal adjustment)

Just as we sometimes want to remove the trend from a series, sometimes we want to seasonally adjust a series before modeling it. **Seasonal adjustment** may be done using a variety of methods.

- Discuss in detail how you'd use a linear trend plus seasonal dummies model to seasonally adjust a series.
- Seasonally adjust the log liquor sales data using a linear trend plus seasonal dummy model. Discuss the patterns present and absent from the seasonally adjusted series.
- Search the Web (or the library) for information on the latest U.S.

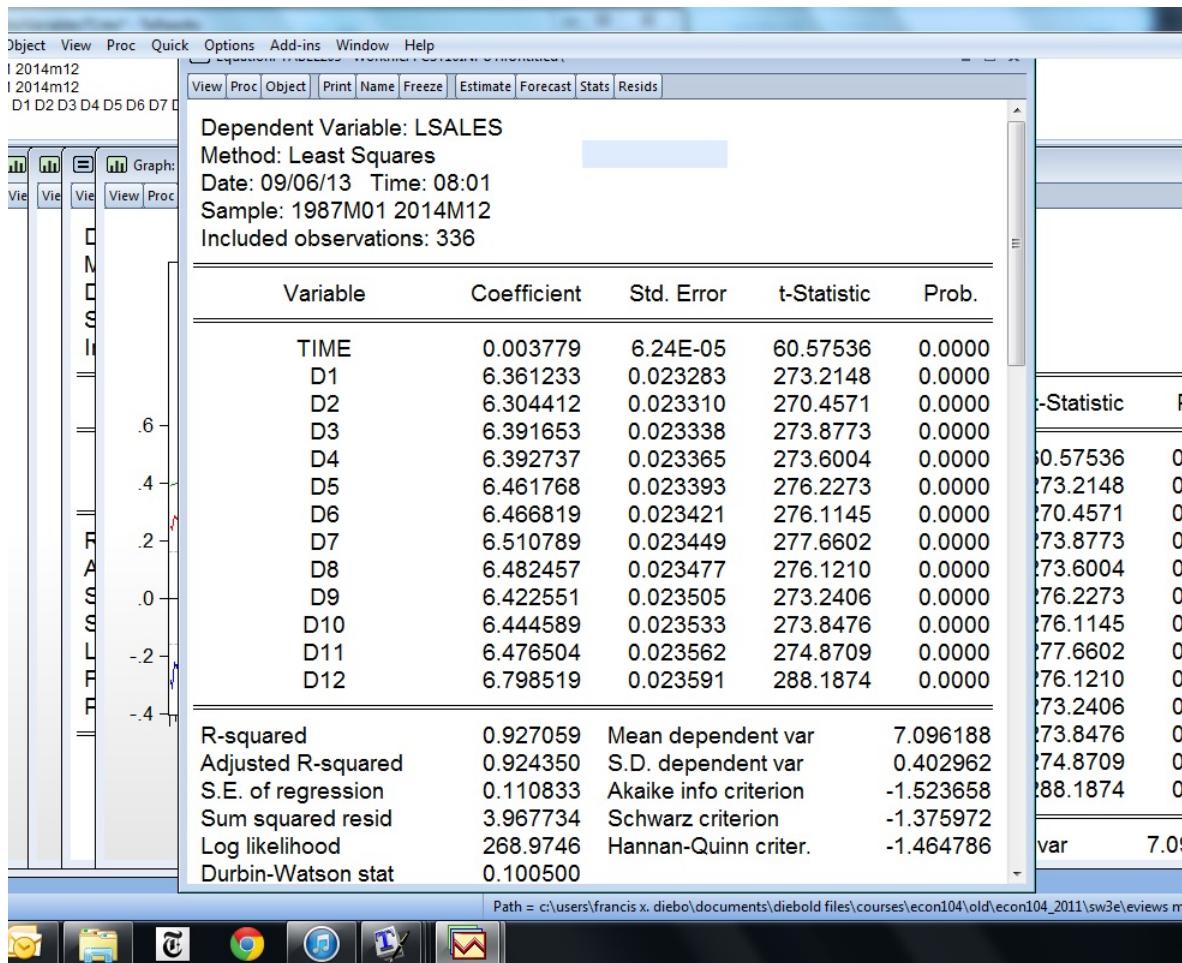


Figure 11.12: Estimation Results, Linear Trend with Seasonal Dummies

Census Bureau seasonal adjustment procedure, and report what you learned.

4. (Handling sophisticated calendar effects)

Describe how you would construct a purely seasonal model for the following monthly series. In particular, what dummy variable(s) would you use to capture the relevant effects?

- A sporting goods store suspects that detrended monthly sales are roughly the same for each month in a given three-month season. For example, sales are similar in the winter months of January, February

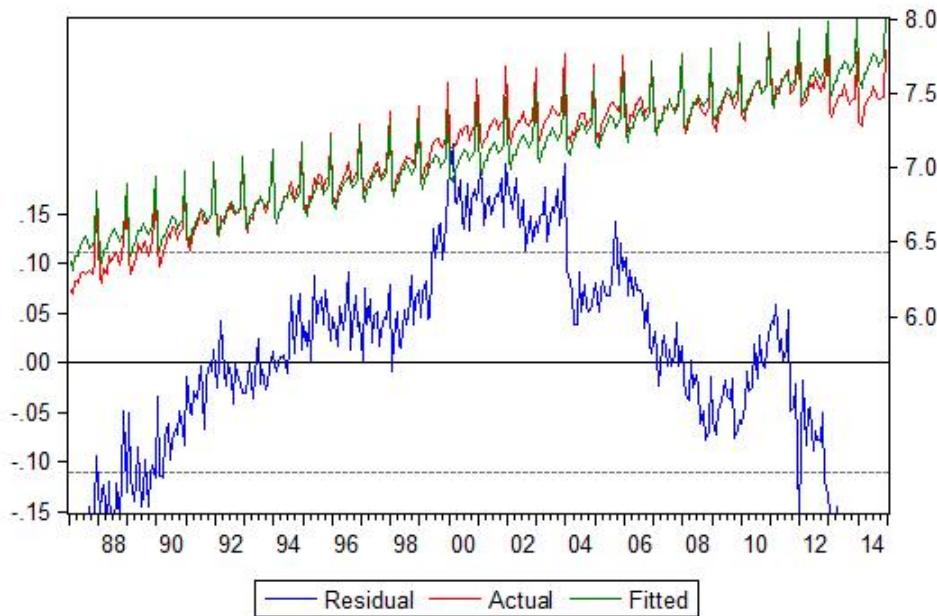


Figure 11.13: Residual Plot, Linear Trend with Seasonal Dummies

- and March, in the spring months of April, May and June, and so on.
- A campus bookstore suspects that detrended sales are roughly the same for all first, all second, all third, and all fourth months of each trimester. For example, sales are similar in January, May, and September, the first months of the first, second, and third trimesters, respectively.
 - (Trading-day effects) A financial-markets trader suspects that detrended trading volume depends on the number of trading days in the month, which differs across months.
 - (Time-varying holiday effects) A candy manufacturer suspects that detrended candy sales tend to rise at Easter.
 - (Testing for seasonality)

Using the log liquor sales data:

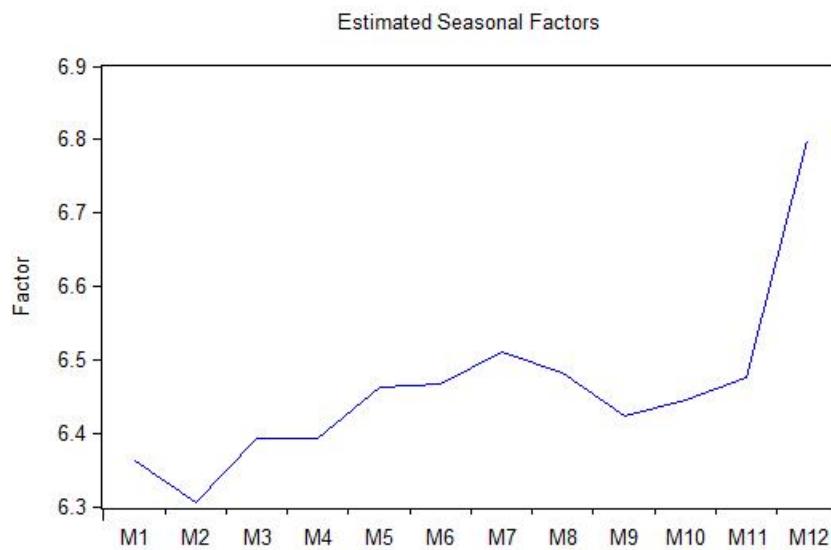


Figure 11.14: Seasonal Pattern

- a. As in the chapter, construct and estimate a model with a full set of seasonal dummies.
- b. Test the hypothesis of no seasonal variation. Discuss.
- c. Test for the equality of the January through April seasonal factors. Discuss.
- d. Test for equality of the May through November seasonal factors. Discuss.
- e. Estimate a suitable “pruned” model with fewer than twelve seasonal dummies that nevertheless adequately captures the seasonal pattern.
6. Specifying and testing nonlinear trend models.

In 1965, Intel co-founder Gordon Moore predicted that the number of transistors that one could place on a square-inch integrated circuit would double every twelve months.

- a. What sort of trend is this?

- b. Given a monthly series containing the number of transistors per square inch for the latest integrated circuit, how would you test Moore's prediction? How would you test the currently accepted form of "Moore's Law," namely that the number of transistors actually doubles every eighteen months?

7. (Properties of polynomial trends)

Consider a sixth-order deterministic polynomial trend:

$$T_t = \beta_1 + \beta_2 TIME_t + \beta_3 TIME_t^2 + \dots + \beta_7 TIME_t^6.$$

- 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. Fit the sixth-order polynomial trend model to a trending series that interests you, and discuss your results.
8. (Selecting non-linear trend models)

Using AIC and SIC, perform a detailed comparison of polynomial vs. exponential trend in LSALES. Do you agree with our use of quadratic trend in the text?

9. (Difficulties with non-linear optimization)

Non-linear optimization can be a tricky business, fraught with problems. Some problems are generic. It's relatively easy to find a local optimum, for example, but much harder to be confident that the local optimum is global. Simple checks such as trying a variety of startup values and checking the optimum to which convergence occurs are used routinely, but the problem nevertheless remains. Other problems may be software

specific. For example, some software may use highly accurate analytic derivatives whereas other software uses approximate numerical derivatives. Even the same software package may change algorithms or details of implementation across versions, leading to different results.

10. (Direct estimation of exponential trend in levels)

We can estimate an exponential trend in two ways. First, as we have emphasized, we can take logs and then use OLS to fit a linear trend. Alternatively we can use NLS, proceeding directly from the exponential representation and letting the computer find

$$(\hat{\beta}_1, \hat{\beta}_2) = \operatorname{argmin}_{\beta_1, \beta_2} \sum_{t=1}^T [y_t - \beta_1 e^{\beta_2 \text{TIME}_t}]^2.$$

- a. The NLS approach is more tedious? Why?
- b. The NLS approach is less thoroughly numerically trustworthy? Why?
- c. Nevertheless the NLS approach can be very useful? Why? (Hint: Consider comparing SIC values for quadratic vs. exponential trend.)

11. (Logistic trend)

In the main text we introduced the logistic functional form. A key example is **logistic trend**, which is

$$\text{Trend}_t = \frac{1}{a + br^{\text{TIME}_t}},$$

with $0 < r < 1$.

- a. Graph the trend shape for various combinations of a and b values. When might such a trend shape be useful?
- b. Can you think of other specialized situations in which other specialized trend shapes might be useful? Produce mathematical formulas for the additional specialized trend shapes you suggest.

12. (Modeling Liquor Sales Trend and Seasonality)

Consider the liquor sales data. Never include an intercept. Discuss all results in detail.

- (a) Fit a linear trend plus seasonal dummy model to log liquor sales ($LSALES$), using a full set of seasonal dummies.
- (b) Find a “best” linear trend plus seasonal dummy $LSALES$ model. That is, consider tightening the seasonal specification to include fewer than 12 seasonal dummies, and decide what’s best.
- (c) Keeping the same seasonality specification as in (12b), re-estimate the model in levels (that is, the LHS variable is now $SALES$ rather than $LSALES$) using exponential trend and nonlinear least squares. Do your coefficient estimates match those from (12b)? Does the SIC match that from (12b)?
- (d) Repeat (12c), again using $SALES$ and again leaving intact your seasonal specification from (12b), but try linear and quadratic trend instead of the exponential trend in (12c). What is your “final” $SALES$ model?
- (e) Critique your final $SALES$ model from (12d). In what ways is it likely still deficient? You will of course want to discuss its residual plot (actual values, fitted values, residuals), as well as any other diagnostic plots or statistics that you deem relevant.
- (f) Take your final estimated $SALES$ model from (12d), and include as regressors three lags of $SALES$ (i.e., $SALES_{t-1}$, $SALES_{t-2}$ and $SALES_{t-3}$). What role do the lags of $SALES$ play? Consider this new model to be your “final, final” $SALES$ model, and repeat (12e).

13. Moving-Average Trend and De-Trending

The trend regression technique is one way to estimate trend. An additional way involves model-free **smoothing** techniques. A leading case is “moving-average smoothing”. We’ll focus on three moving-average smoothers: two-sided moving averages, one-sided moving averages, and one-sided weighted moving averages. Denote the original data by $\{y_t\}_{t=1}^T$ and the smoothed data by $\{z_t\}_{t=1}^T$. Then the **two-sided moving average** is

$$z_t = (2m + 1)^{-1} \sum_{i=-m}^m y_{t-i},$$

the **one-sided moving average** is

$$z_t = (m + 1)^{-1} \sum_{i=0}^m y_{t-i},$$

and the **one-sided weighted moving average** is

$$z_t = \sum_{i=0}^m w_i y_{t-i},$$

where the w_i are weights and m is an integer chosen by the user. The “standard” one-sided moving average corresponds to a one-sided weighted moving average with all weights equal to $(m + 1)^{-1}$.

- a. For each of the smoothing techniques, discuss the role played by m . What happens as m gets very large? Very small? In what sense does m play a role similar to p , the order of a polynomial trend?
- b. If the original data runs from time 1 to time T , over what range can smoothed values be produced using each of the three smoothing methods? What are the implications for “real-time” smoothing or “on-line” smoothing versus “ex post” smoothing or “off-line” smoothing?

Another model-free approach to trend fitting and de-trending is known as **Hodrick-Prescott filtering**. The “HP trend” solves:

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

- a. λ is often called the “penalty parameter.” What does λ govern?
- b. What happens as $\lambda \rightarrow 0$?
- c. What happens as $\lambda \rightarrow \infty$?
- d. People routinely use bigger λ for higher-frequency data. Why? (Common values are $\lambda = 100, 1600$ and $14,400$ for annual, quarterly, and monthly data, respectively.)

15. Regime Switching I: Observed-Regime Threshold Model

$$y_t = \begin{cases} c^{(u)} + \phi^{(u)} y_{t-1} + \varepsilon_t^{(u)}, & \theta^{(u)} < y_{t-d} \\ c^{(m)} + \phi^{(m)} y_{t-1} + \varepsilon_t^{(m)}, & \theta^{(l)} < y_{t-d} < \theta^{(u)} \\ c^{(l)} + \phi^{(l)} y_{t-1} + \varepsilon_t^{(l)}, & \theta^{(l)} > y_{t-d} \end{cases}$$

16. Regime Switching II: Markov-Switching Model

Regime governed by latent 2-state Markov process:

$$M = \begin{pmatrix} p_{00} & 1 - p_{00} \\ 1 - p_{11} & p_{11} \end{pmatrix}$$

Switching mean:

$$f(y_t | s_t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y_t - \mu_{s_t})^2}{2\sigma^2}\right).$$

Switching regression:

$$f(y_t|s_t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y_t - x'_t\beta_{s_t})^2}{2\sigma^2}\right).$$

11.6 Notes

Nerlove et al. (1979) and Harvey (1991) discuss a variety of models of trend and seasonality.

The two most common and important “official” seasonal adjustment methods are X-12-ARIMA from the U.S. Census Bureau, and TRAMO-SEATS from the Bank of Spain.

Chapter 12

Serial Correlation

In this chapter we consider serially correlated regression disturbances, a new type of violation of the IC of crucial importance in time series. Disturbance serial correlation, or autocorrelation, means *correlation over time*. That is, the current disturbance is correlated with one or more past disturbances.

Under the IC we have:

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

Now, with serial correlation, we have:

$$\varepsilon \sim N(\underline{0}, \Omega),$$

where Ω is not diagonal. A key cause is omission of serially-correlated x 's in the regression specification, which results in serially-correlated ε . Hence the “omitted variables problem” and the “serial correlation problem” are closely related.

A leading example is “first-order autoregressive” or “ $AR(1)$ ” disturbance serial correlation:

$$y_t = x'_t \beta + \varepsilon_t$$

$$\varepsilon_t = \phi \varepsilon_{t-1} + v_t, \quad |\phi| < 1$$

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

Extension to “ $AR(p)$ ” disturbance serial correlation is immediate.

Serial correlation has important consequences for β estimation and inference. As with Heteroskedasticity, point remains OK (OLS parameter estimates remain consistent and asymptotically normal), but inference is damaged (OLS standard errors are biased and inconsistent).

Serial correlation also has important consequences for y prediction. Unlike With heteroskedasticity, even *point* predictions need re-thinking. Hence serial correlation is a bigger problem for prediction than heteroskedasticity. Here’s the intuition. Serial correlation in disturbances / residuals implies that the included “ x variables” have missed something that could be exploited for improved *point* forecasting of y (and hence also improved interval and density forecasting). That is, *all* types of forecasts are sub-optimal when serial correlation is neglected. Put differently, serial correlation in forecast errors means that you can forecast your forecast errors! So something is wrong and can be improved...

12.1 Characterizing Serial Correlation (in Population, Mostly)

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 might think of a rigid up-and-down pattern, as for example with a cos or sin function, 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, all-encompassing, notion of cyclical: any sort of dynamics not captured by trends or seasonals.

Cycles, according to our broad interpretation, simply have to have some dynamics, some persistence, some way in which the present is linked to the past, and 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 deterministic models. Cyclical dynamics, however, are more complicated. Because of the wide variety of cyclical patterns, the sorts of models we need are substantially more involved. The material is also a bit difficult the first time around because it's unavoidably rather mathematical, so careful, systematic study is required.

12.1.1 Covariance Stationary Time Series

Now we introduce the idea of a covariance stationary time series. We will generally use y_t to denote a time series. That series could be unobserved (like the disturbance in a regression, which we called ε_t in the motivational discussion above), or it could be observed (like U.S. GDP). Everything we say below is valid either way.

Formally, a time series is an ordered infinite-dimensional random variable. 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 τ is just the covariance between y_t and $y_{t-\tau}$. It will of course depend on τ , 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, τ , 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 τ . Note that

the autocovariance function is symmetric; that is, $\gamma(\tau) = \gamma(-\tau)$, for all τ . Typically, we'll consider only non-negative values of τ . 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.¹ The upshot

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

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.

At the beginning of this chapter we noted that autocorrelation corresponds to non-diagonal Ω . Now, having introduced the autocovariance function, we can display the precise form of Ω under serial correlation:

$$\Omega = \begin{pmatrix} \gamma_\varepsilon(0) & \gamma_\varepsilon(1) & \dots & \gamma_\varepsilon(T-1) \\ \gamma_\varepsilon(1) & \gamma_\varepsilon(0) & \dots & \gamma_\varepsilon(T-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_\varepsilon(T-1) & \gamma_\varepsilon(T-2) & \dots & \gamma_\varepsilon(0) \end{pmatrix}$$

Note the "band symmetric" structure, illustrated here for $T = 4$:

$$\Omega = \begin{pmatrix} a & b & c & d \\ b & a & b & c \\ c & b & a & b \\ d & c & b & a \end{pmatrix}$$

Now we introduce the closely-related autocorrelation function. 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**, $\rho(\tau)$, 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 contemporaneously 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 $y_{t-1}, \dots, y_{t-\tau}$.² We call such regressions **autoregressions**, because the variable is regressed on lagged values of itself. It's easy to see that the

²To 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.

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} , ..., $y_{t-\tau+1}$; that is, they measure the partial correlation between y_t and $y_{t-\tau}$.

As with the autocorrelations, we often graph the partial autocorrelations as a function of τ 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.³

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 12.1 we show an autocorrelation function that displays gradual one-sided damping. The precise decay patterns of autocorrelations and partial autocorrelations of a covariance stationary series, however, depend on the specifics of the series. In Figure 12.2, for example, we show an autocorrelation function that that differs in the way it approaches zero – the autocorrelations drop abruptly to zero beyond a certain displacement.

12.1.2 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

³Also 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.

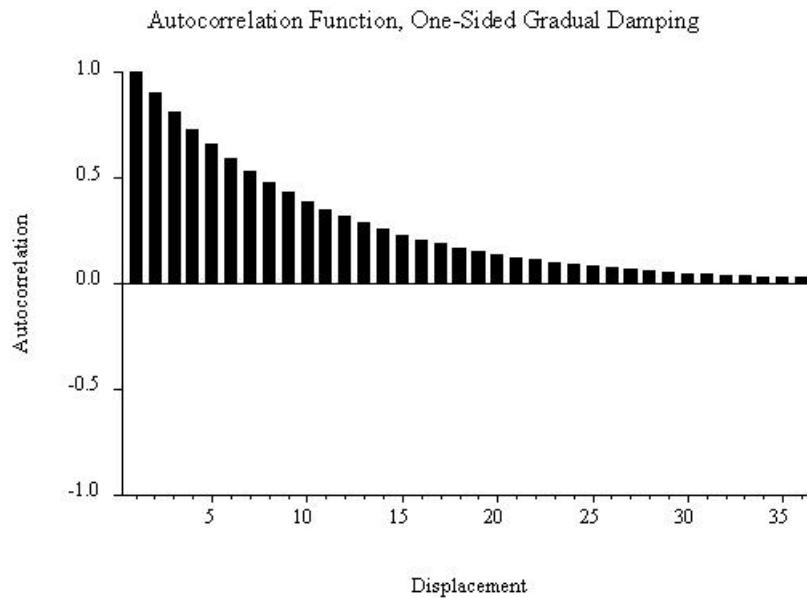


Figure 12.1

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.

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.

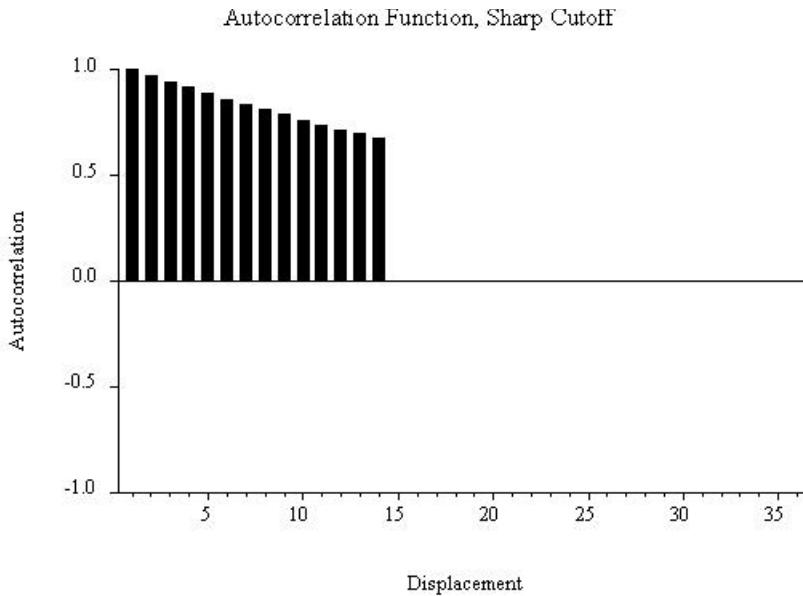


Figure 12.2

Sample Autocorrelations

The autocorrelation at displacement τ 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 τ , 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 τ (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 yields⁴

$$T\hat{\rho}^2(\tau) \sim \chi_1^2.$$

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 χ^2 variables is also χ^2 with degrees of freedom equal to the sum of the degrees 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 χ_m^2 random variable under the null hypothesis that y is white noise.⁵ A slight modification of this, designed to follow more closely the χ^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 χ_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

⁴Recall that the square of a standard normal random variable is a χ^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 τ , as we will soon do.

⁵ m is a maximum displacement selected by the user. Shortly we'll discuss how to choose it.

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.

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 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 τ 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.

12.2 Modeling Serial Correlation (in Population)

12.2.1 White Noise

In this section 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,” ε_t , is uncorrelated over time. We say that ε_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**.⁶ Sometimes for short we write

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

⁶It'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.

and hence

$$y_t \sim WN(0, \sigma^2).$$

Note that, although ε_t and hence y_t are serially uncorrelated, they are not necessarily serially independent, because they are not necessarily normally distributed.⁷ If in addition to being serially uncorrelated, y is serially independent, then we say that y is **independent white noise**.⁸ 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**.⁹ 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 12.3 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 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 distur-

⁷Recall that zero correlation implies independence only in the normal case.

⁸Another name for independent white noise is **strong white noise**, in contrast to standard serially uncorrelated **weak white noise**.

⁹Carl Friedrich Gauss, one of the greatest mathematicians of all time, discovered the normal distribution some 200 years ago; hence the adjective “Gaussian.”

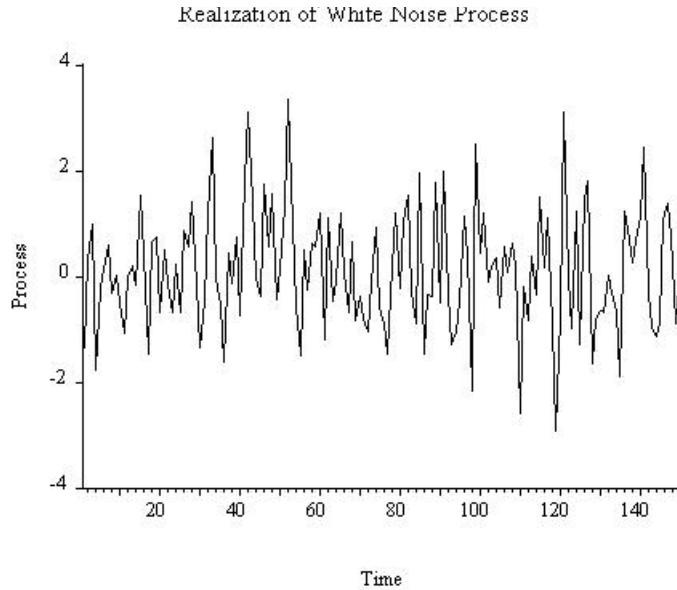


Figure 12.3

bances. 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 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.¹⁰

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,

¹⁰Recall that $\sigma^2 = \gamma(0)$.

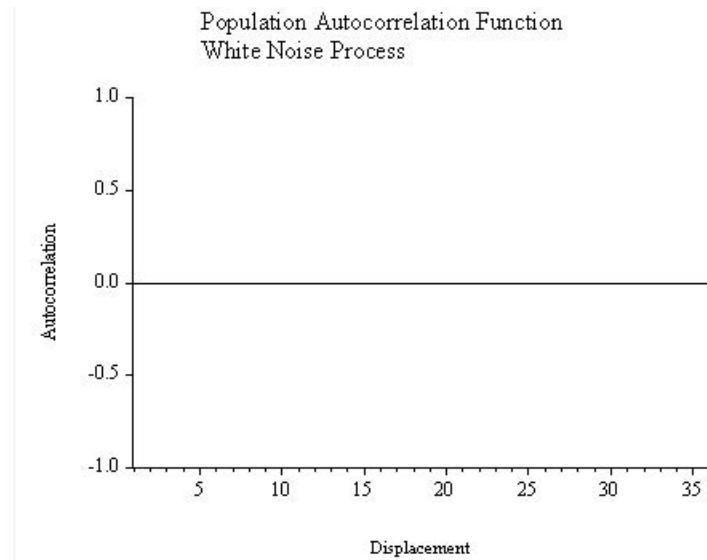


Figure 12.4

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 0.¹¹ Formally, then, the autocovariance function for a white noise process is

$$\gamma(\tau) = \begin{cases} \sigma^2, & \tau = 0 \\ 0, & \tau \geq 1, \end{cases}$$

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 12.4 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

¹¹If the autocovariances are all zero, so are the autocorrelations, because the autocorrelations are proportional to the autocovariances.

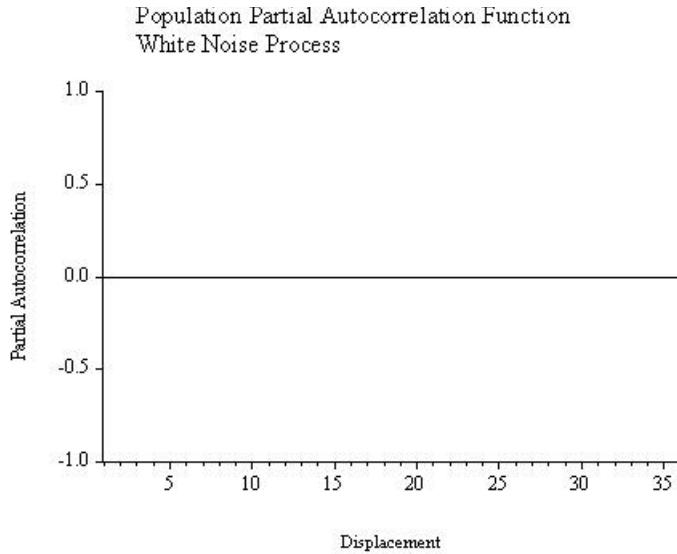


Figure 12.5

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

$$p(\tau) = \begin{cases} 1, & \tau = 0 \\ 0, & \tau \geq 1. \end{cases}$$

We show the partial autocorrelation function of a white noise process in Figure 12.5 . 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.¹² 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 σ^2 . Now consider the conditional mean and variance, where the information set Ω_{t-1} upon which we condition contains either the

¹²If you need to refresh your memory on conditional means, consult any good introductory statistics book, such as Wonnacott and Wonnacott (1990).

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

$$E(y_t | \Omega_{t-1}) = 0,$$

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.

12.2.2 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 Δ , 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

$$B(L) = b_0 + b_1L + b_2L^2 + \dots = \sum_{i=0}^{\infty} b_iL^i.$$

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. =’s theorem, to which we now turn, establishes that centrality.

12.2.3 Autoregression

When building 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 more complex reality. That’s the modern view, and it has important implications for time-series modeling. In particular, the key to successful time series modeling is parsimonious, yet accurate, approximations. Here we emphasize a very important class of approximations, the **autoregressive (AR) model**.

We begin by characterizing the autocorrelation function and related quantities under the assumption that the *AR* model is “true.”¹³ 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. 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.

¹³Sometimes, 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**.

The autoregressive process is a natural approximation to time-series dynamics. 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

$$y_t = \phi y_{t-1} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

In lag operator form, we write

$$(1 - \phi L)y_t = \varepsilon_t.$$

In Figure 12.6 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

$$\varepsilon_t \sim iidN(0, 1),$$

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.

Certain conditions must be satisfied for an autoregressive process to be

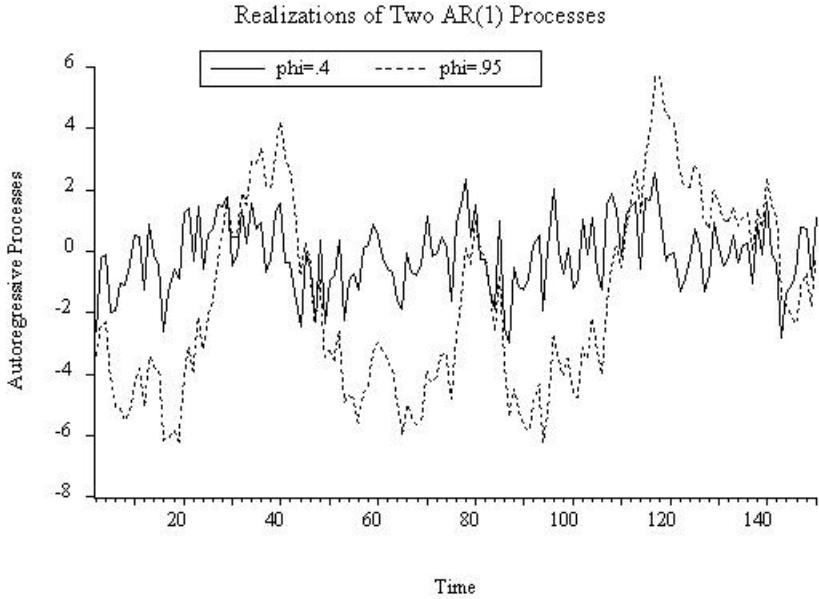


Figure 12.6

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

$$y_t = \varepsilon_t + \phi \varepsilon_{t-1} + \phi^2 \varepsilon_{t-2} + \dots$$

This representation of y in terms of current and past shocks is called a **moving-average representation**. 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.

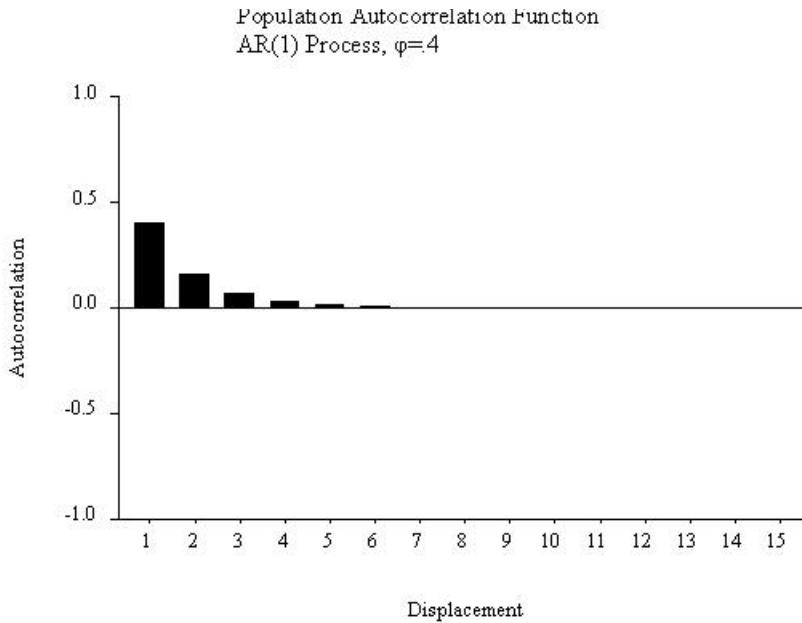


Figure 12.7

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

$$var(y_t) = var(\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \dots)$$

$$= \sigma^2 + \phi^2\sigma^2 + \phi^4\sigma^2 + \dots$$

$$= \sigma^2 \sum_{i=0}^{\infty} \phi^{2i}$$

$$= \frac{\sigma^2}{1-\phi^2}.$$

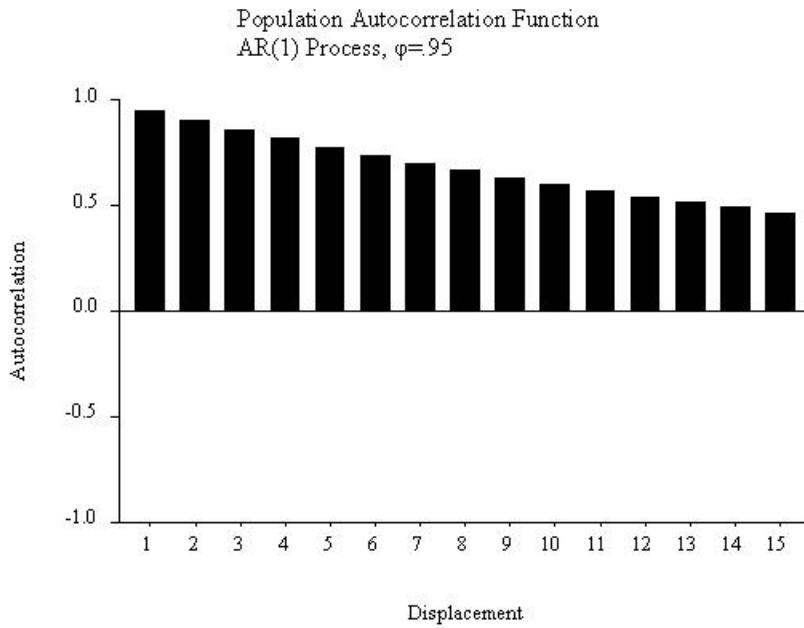


Figure 12.8

The conditional moments, in contrast, are

$$E(y_t|y_{t-1}) = E(\phi y_{t-1} + \varepsilon_t|y_{t-1})$$

$$= \phi E(y_{t-1}|y_{t-1}) + E(\varepsilon_t|y_{t-1})$$

$$= \phi y_{t-1} + 0$$

$$= \phi y_{t-1}$$

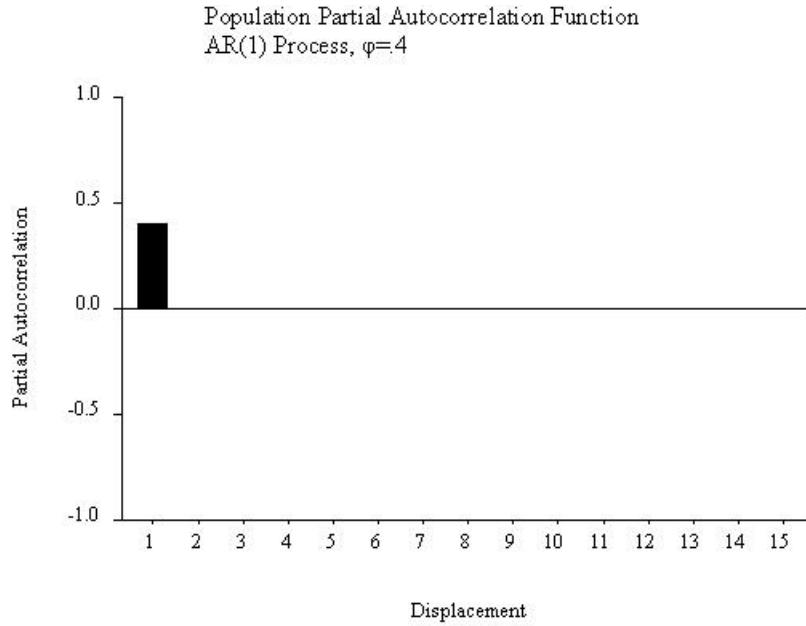


Figure 12.9

and

$$\text{var}(y_t | y_{t-1}) = \text{var}((\phi y_{t-1} + \varepsilon_t) | y_{t-1})$$

$$= \phi^2 \text{var}(y_{t-1} | y_{t-1}) + \text{var}(\varepsilon_t | y_{t-1})$$

$$= 0 + \sigma^2$$

$$= \sigma^2.$$

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

$$y_t y_{t-\tau} = \phi y_{t-1} y_{t-\tau} + \varepsilon_t y_{t-\tau}.$$

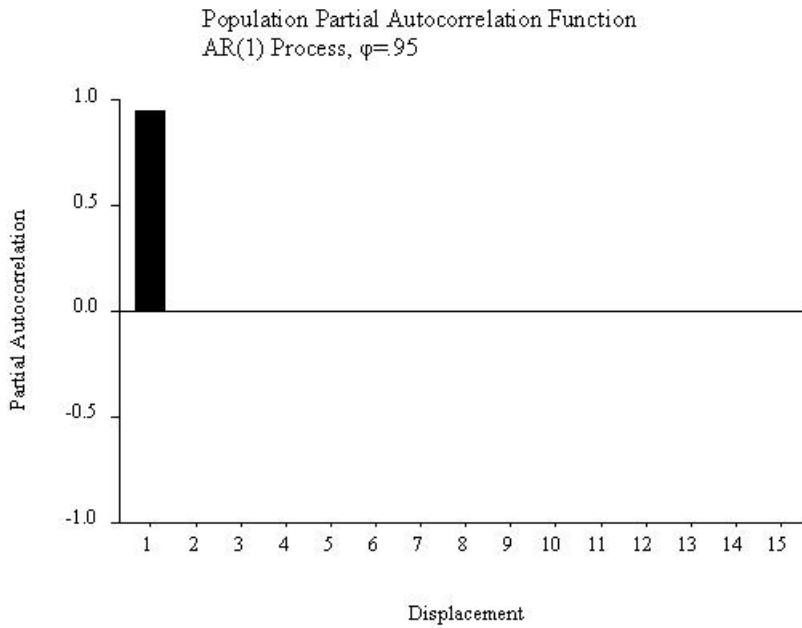


Figure 12.10

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 τ , 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

$$\gamma(0) = \sigma^{\frac{2}{1-\phi^2}}.$$

Thus we have

$$\gamma(0) = \sigma^{\frac{2}{1-\phi^2}}$$

$$\gamma(1) = \phi\sigma^{\frac{2}{1-\phi^2}}$$

$$\gamma(2) = \phi^2\sigma^{\frac{2}{1-\phi^2}},$$

and so on. In general, then,

$$\gamma(\tau) = \phi^\tau \sigma^{\frac{2}{1-\phi^2}}, \tau = 0, 1, 2, \dots$$

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, 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 ϕ is positive, the autocorrelation decay is one-sided. If ϕ 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 Figures 12.7 and 12.8 we show the autocorrelation functions for $AR(1)$ processes with parameters $\phi = .4$ and $\phi = .95$. The persistence is much stronger when $\phi = .95$.

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 12.9 and 12.10 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.

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 the inverses of all roots of the autoregressive lag operator polynomial $\Phi(L)$ are inside the unit circle.¹⁴ 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. 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

¹⁴A 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.

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.¹⁵ 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 process at hand; we plot it in Figure 12.11. Because the roots are complex, the autocorrelation function oscillates, and because the roots are close to the unit circle, the oscillation damps slowly.

¹⁵Note that complex roots can't occur in the $AR(1)$ case.

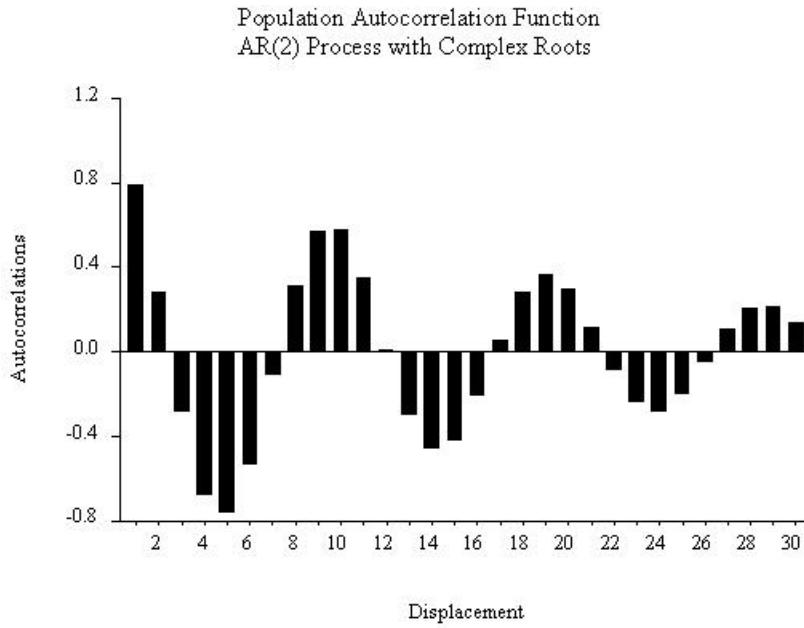


Figure 12.11

12.3 Modeling Serial Correlation (in Sample)

Here we return to regression with $AR(p)$ disturbances, as always using the liquor sales data for illustration.

12.3.1 Detecting Serial Correlation

If a model has extracted all the systematic information from the data, then what's left – the residual – should be *iid* random noise. Hence the usefulness of various residual-based tests of the hypothesis that regression disturbances are white noise (i.e., not serially correlated).

Of course the most obvious thing is simply to inspect the residual plot. For convenience we reproduce our liquor sales residual plot in Figure 12.12. There is clear visual evidence of serial correlation in our liquor sales residuals. Sometimes, however, things are not so visually obvious. Hence we now introduce some additional tools.

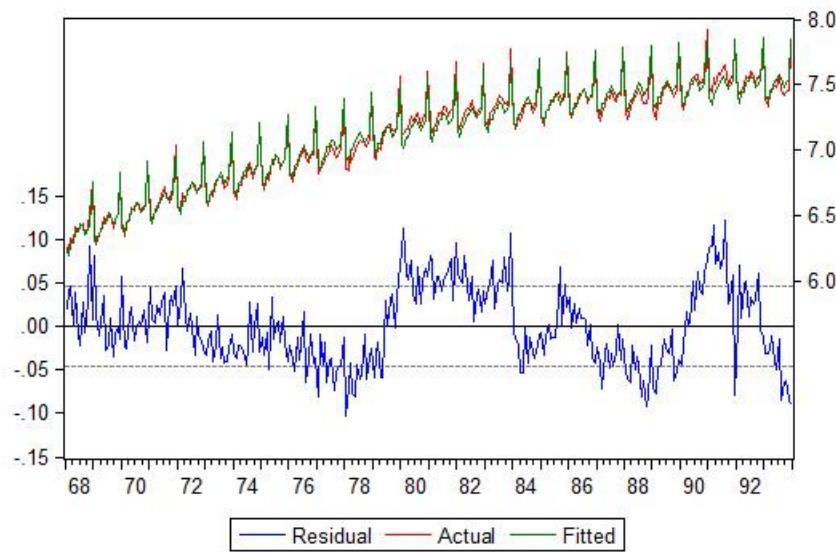


Figure 12.12

Residual Scatterplots

The relevant scatterplot for detecting serial correlation involves plotting e_t against $e_{t-\tau}$. A leading case, corresponding to potential relevance of first-order serial correlation, involves plotting e_t against e_{t-1} . We show this scatterplot for our liquor sales residuals in Figure 12.13. There is an obvious relationship.

Durbin-Watson

The Durbin-Watson is a more formal test. We work in the simple paradigm ($AR(1)$):

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

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

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

The regression disturbance is serially correlated when $\phi \neq 0$. The hypothesis of interest is that $\phi = 0$. When $\phi = 0$, the ideal conditions hold,

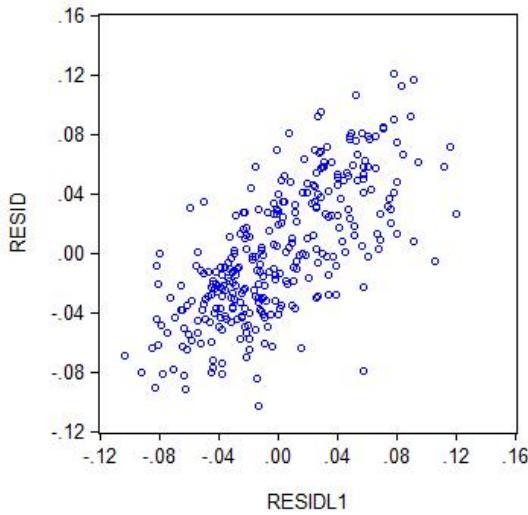


Figure 12.13

but when $\phi \neq 0$, the disturbance is serially correlated. More specifically, when $\phi \neq 0$, ε_t follows an $AR(1)$ processs. 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 economic applications.

Proceeding, we want to test $H_0 : \phi = 0$ against $H_1 : \phi \neq 0$. We regress $y \rightarrow X$ and obtain the residuals e_t

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

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.

Let us attempt to understand DW more thoroughly. We have:

$$\begin{aligned} DW &= \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T e_t^2} = \frac{\frac{1}{T} \sum_{t=2}^T (e_t - e_{t-1})^2}{\frac{1}{T} \sum_{t=1}^T e_t^2} \\ &= \frac{\frac{1}{T} \sum_{t=2}^T e_t^2 + \frac{1}{T} \sum_{t=2}^T e_{t-1}^2 - 2 \frac{1}{T} \sum_{t=2}^T e_t e_{t-1}}{\frac{1}{T} \sum_{t=1}^T e_t^2} \end{aligned}$$

Hence as $T \rightarrow \infty$:

$$DW \approx \frac{\sigma^2 + \sigma^2 - 2\text{cov}(e_t, e_{t-1})}{\sigma^2} = 2(1 - \underbrace{\text{corr}(e_t, e_{t-1})}_{\rho_e(1)})$$

Hence $DW \in [0, 4]$, $DW \rightarrow 2$ as $\phi \rightarrow 0$, and $DW \rightarrow 0$ as $\phi \rightarrow 1$.

Also note 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 if the regressors include lagged dependent variables.¹⁶ (See EPC 6.) On both counts, we'd like more general and flexible approaches for diagnosing serial correlation.

For liquor sales, $DW = .59$ – clear evidence of residual serial correlation!

The Breusch-Godfrey Test

The **Breusch-Godfrey test** is an 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.

¹⁶Following 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.

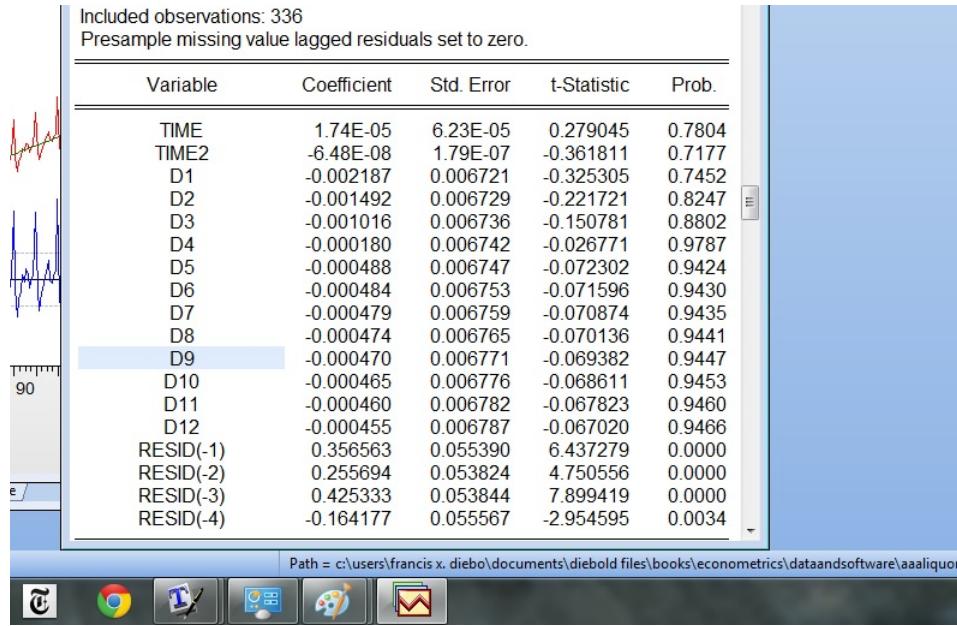


Figure 12.14: BG Test Equation, 4 Lags

We work in a general $AR(p)$ environment:

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

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \dots + \phi_p \varepsilon_{t-p} + v_t$$

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

We want to test $H_0 : (\phi_1, \dots, \phi_p) = \underline{0}$ against $H_1 : (\phi_1, \dots, \phi_p) \neq \underline{0}$. We proceed as follows:

1. Regress $y_t \rightarrow x_t$ and obtain the residuals e_t
2. Regress $e_t \rightarrow x_t, e_{t-1}, \dots, e_{t-p}$
3. Examine TR^2 . In large samples $TR^2 \sim \chi_p^2$ under the null.

This should sound familiar, as it precisely parallels the BGP heteroskedasticity test that we studied earlier in Chapter 8.

Some test regression results appear in Figures 12.14-12.15. In particular we have a BG for $AR(4)$ disturbances of $TR^2 = 216.7$, ($p = 0.0000$), and

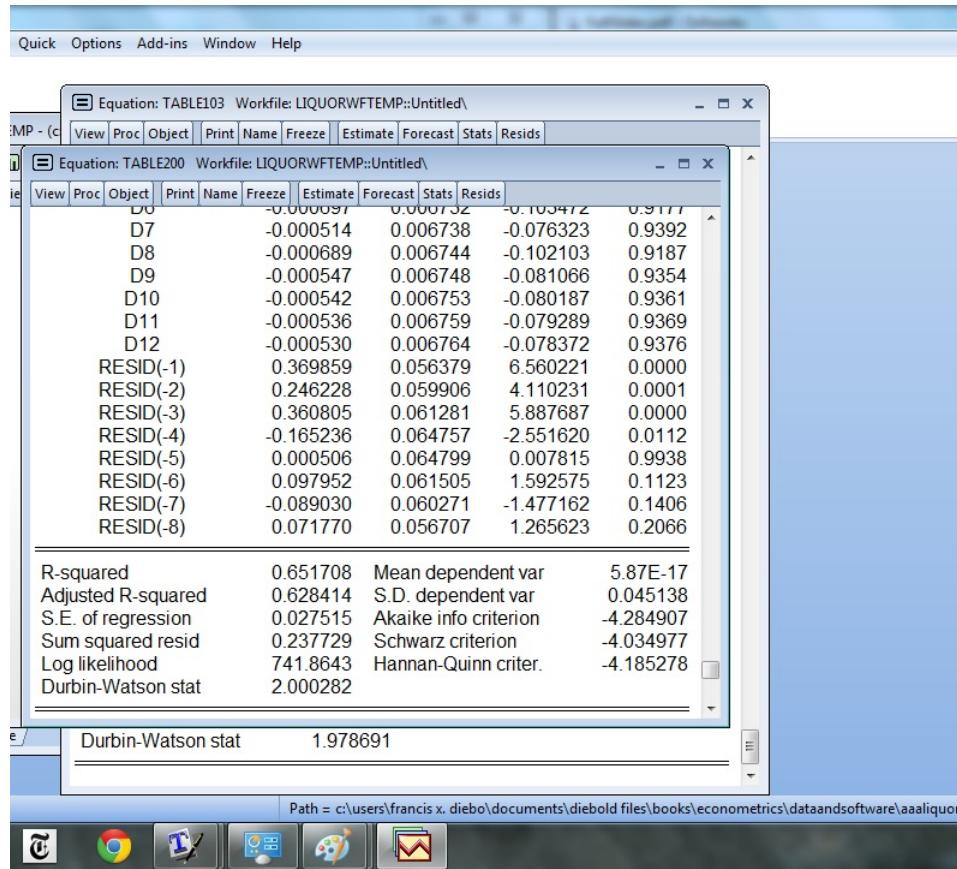


Figure 12.15: BG Test Equation, 8 Lags

a BG for $AR(8)$ Disturbances of $TR^2 = 219.0$ ($p = 0.0000$). There is strong evidence of autoregressive dynamics through lag 4, but not after lag 4, suggesting $AR(4)$.

12.3.2 The Residual Correlogram

The residual sample autocorrelations are:

$$\hat{\rho}_e(\tau) = \frac{\widehat{cov}(e_t, e_{t-\tau})}{\widehat{var}(e_t)} = \frac{\frac{1}{T} \sum_t e_t e_{t-\tau}}{\frac{1}{T} \sum_t e_t^2}.$$

The residual sample partial autocorrelation at displacement τ s, $\hat{p}_e(\tau)$, is

the coefficient on $e_{t-\tau}$ in the regression

$$e_t \rightarrow c, e_{t-1}, \dots, e_{t-(\tau-1)}, e_{t-\tau}.$$

The approximate 95% “Bartlett bands” remain $0 \pm \frac{2}{\sqrt{T}}$. The Q statistics also remain unchanged:

$$\begin{aligned} Q_{BP} &= T \sum_{\tau=1}^m \hat{\rho}_e^2(\tau) \sim \chi_{m-K}^2 \\ Q_{LB} &= T(T+2) \sum_{\tau=1}^m \left(\frac{1}{T-\tau} \right) \hat{\rho}_e^2(\tau) \sim \chi_{m-K}^2. \end{aligned}$$

The only wrinkle is that, when we earlier introduced the correlogram, we focused on the case of an observed time series, in which case we showed that the Q statistics are distributed as χ_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 χ_{m-K}^2 random variable, where k is the number of parameters estimated.

We show the residual correlogram for the trend + seasonal model in Figure 12.16. It strongly supports the $AR(4)$ specification. The sample autocorrelations decay gradually, and the sample partial autocorrelations cut off sharply at displacement 4.

12.3.3 Estimating Serial Correlation

The remaining issue is how to estimate a regression model with serially correlated disturbances. Let us illustrate with the $AR(1)$ case. The model is:

$$y_t = x'_t \beta + \varepsilon_t \quad (1a)$$

Included observations: 312

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
0.700	0.700	1	0.700	154.34	0.000
0.686	0.383	2	0.686	302.86	0.000
0.725	0.369	3	0.725	469.36	0.000
0.569	-0.141	4	0.569	572.36	0.000
0.569	0.017	5	0.569	675.58	0.000
0.577	0.093	6	0.577	782.19	0.000
0.460	-0.078	7	0.460	850.06	0.000
0.480	0.043	8	0.480	924.38	0.000
0.466	0.030	9	0.466	994.46	0.000
0.327	-0.188	10	0.327	1029.1	0.000

Figure 12.16: Residual Correlogram From Trend + Seasonal Model

$$\varepsilon_t = \phi \varepsilon_{t-1} + v_t \quad (1b)$$

$$v_t \sim iid N(0, \sigma^2) \quad (1c)$$

It is possible to work out the likelihood function for this model and maximize it. It is also possible to work out so-called “generalized least squares” procedures, which are different from OLS and which account for serial correlation. Fortunately, however, there is no need for any of that. Instead, let us simply manipulate the above equations a bit. We have:

$$\phi y_{t-1} = \phi x'_{t-1} \beta + \phi \varepsilon_{t-1} \quad (1a*) \text{ (by multiplying (1a) through by } \phi)$$

$$\implies (y_t - \phi y_{t-1}) = (x'_t - \phi x'_{t-1}) \beta + (\varepsilon_t - \phi \varepsilon_{t-1}) \text{ (just (1a) - (1a*))}$$

$$\implies y_t = \phi y_{t-1} + x'_t \beta - x'_{t-1}(\phi \beta) + v_t$$

This “new” model satisfies the IC!

So dealing with autocorrelated disturbances amounts to nothing more than including some extra lags in the regression. The IC are satisfied so OLS is fine. $AR(1)$ disturbances require 1 lag, as we just showed. General $AR(p)$ disturbances require p lags.

For liquor sales, everything points to $AR(4)$ dynamics

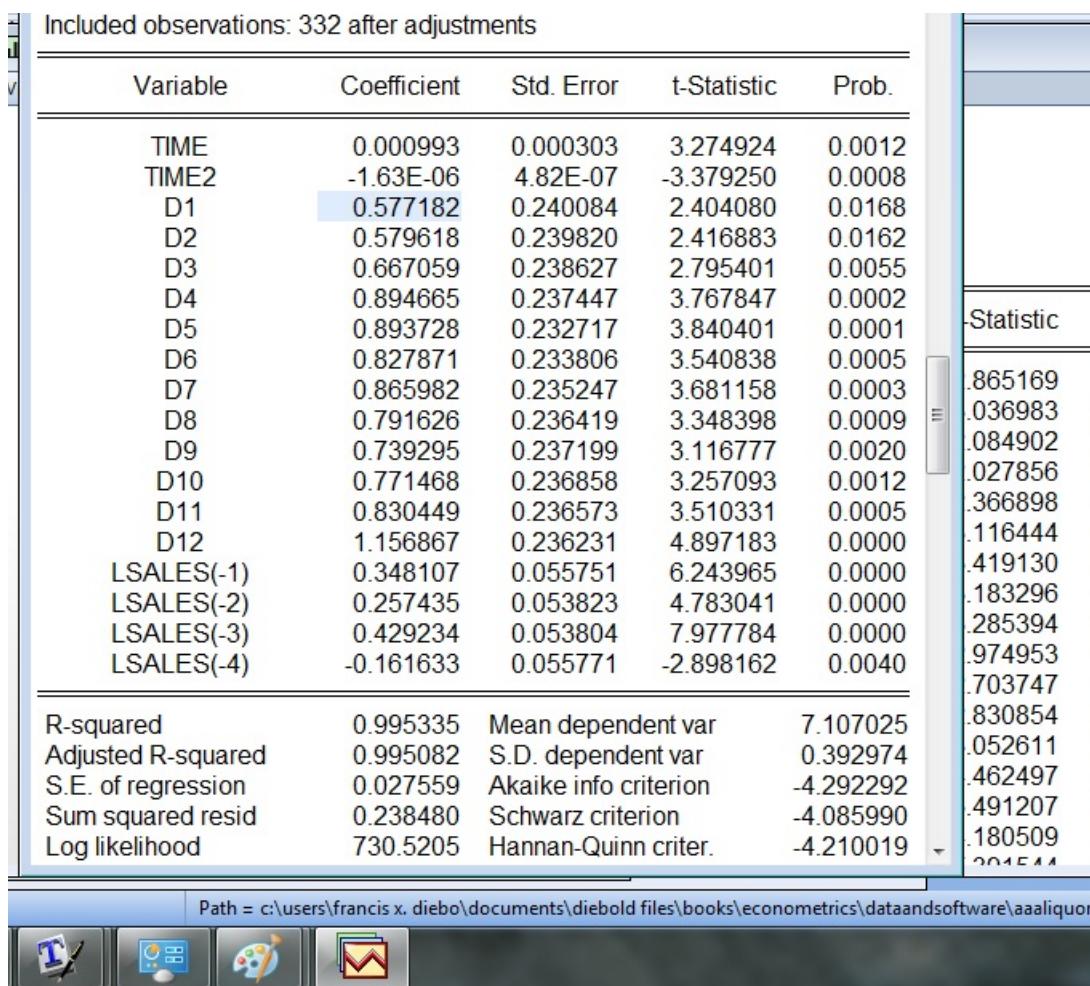


Figure 12.17: Trend + Seasonal Model with Four Autoregressive Lags

- Supported by original trend + seasonal residual correlogram
- Supported by *DW* (designed to detect $AR(1)$ but of course it can also reject against higher-order autoregressive alternatives)
 - Supported by *BG*
 - Supported by *SIC* pattern ($AR(1) = -3.797$, $AR(2) = -3.941$, $AR(3) = -4.080$, $AR(4) = -4.086$, $AR(5) = -4.071$, $AR(6) = -4.058$, $AR(7) = -4.057$, $AR(8) = -4.040$)

In Figures 12.17-12.20 we show the “final” model estimation results, the corresponding residual plot, and the residual histogram and normality test.

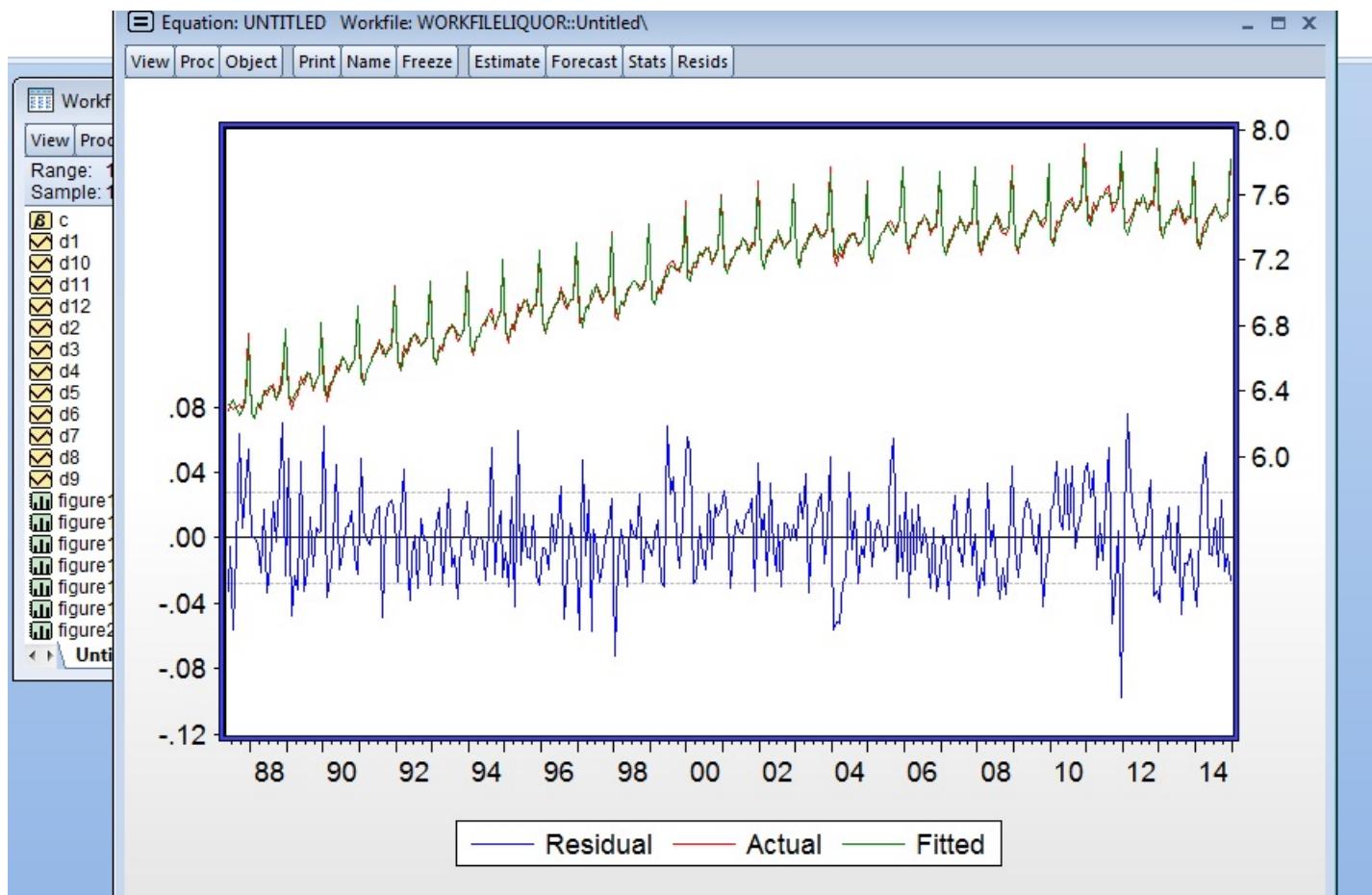


Figure 12.18: Trend + Seasonal Model with Four Lags of y , Residual Plot

12.4 Exercises, Problems and Complements

1. (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 time series. 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. $\gamma(t, \tau) = \alpha\tau$

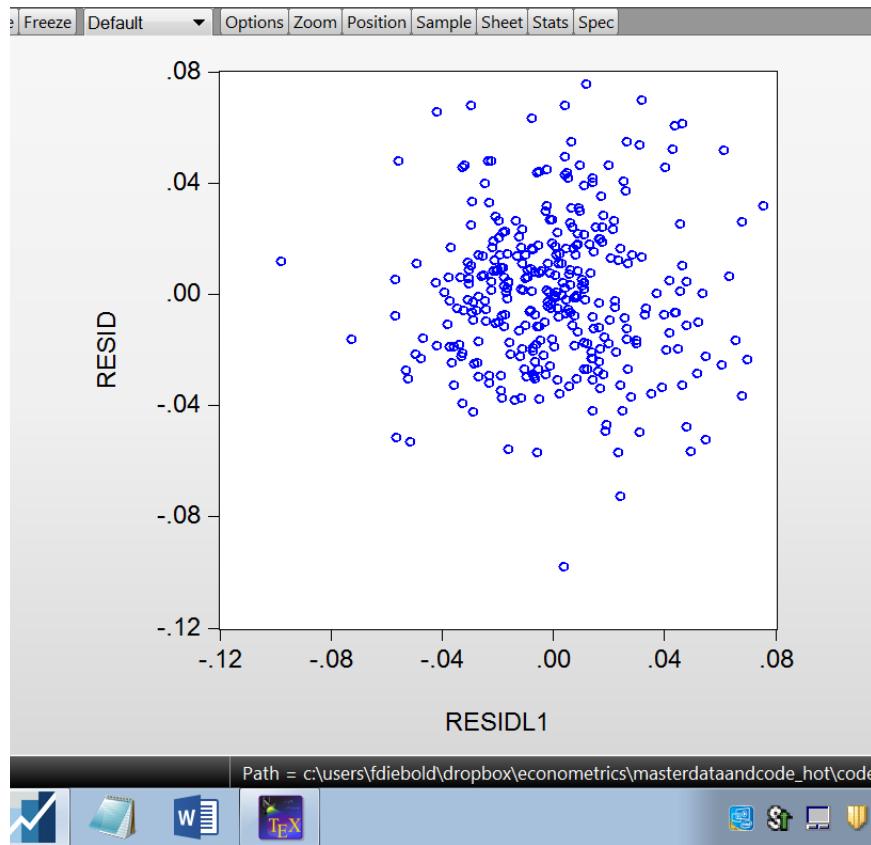


Figure 12.19: Trend + Seasonal Model with Four Autoregressive Lags, Residual Scatterplot

d. $\gamma(t, \tau) = \frac{\alpha}{\tau}$, where α is a positive constant.

Which autocovariance function(s) are consistent with covariance stationarity, and which are not? Why?

2. (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?

3. (Simulating time series processes)

Many cutting-edge estimation techniques involve simulation. Moreover, simulation is often a good way to get a feel for a model and its behavior.

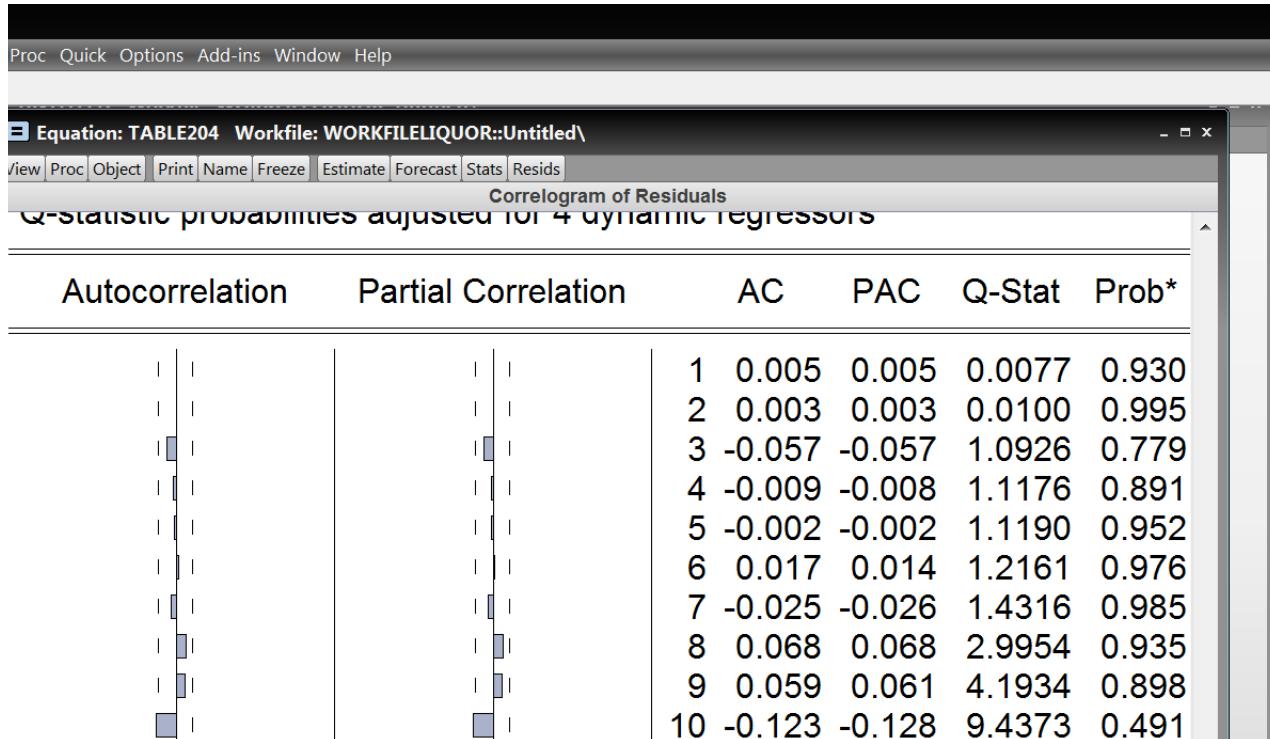


Figure 12.20: Trend + Seasonal Model with Four Autoregressive Lags, Residual Autocorrelations

White noise can be simulated on a computer using **random number generators**, which are available in most statistics, econometrics and forecasting packages.

- Simulate a Gaussian white noise realization of length 200. Call the white noise ε_t . Compute the correlogram. Discuss.
- 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.
- 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.
- (Outliers in Time Series)

Outliers can arise for a number of reasons. Perhaps the outlier is simply a mistake due to a clerical recording error, in which case you'd want to

replace the incorrect data with the correct data. We'll call such outliers **measurement outliers**, because they simply reflect measurement errors. In a time-series context, if a particular value of a recorded series is plagued by a measurement outlier, there's no reason why observations at other times should necessarily be affected.

Alternatively, outliers in time series may be associated with large unanticipated shocks, the effects of which may certainly linger. If, for example, an adverse shock hits the U.S. economy this quarter (e.g., the price of oil on the world market triples) and the U.S. plunges into a severe depression, then it's likely that the depression will persist for some time. Such outliers are called **innovation outliers**, because they're driven by shocks, or "innovations," whose effects naturally last more than one period due to the dynamics operative in business, economic, and financial series.

5. (DW from a pure trend model)

Fit a quadratic trend to the liquor sales data, and check the DW statistic. It looks fine. Why? Are things really fine?

6. (Diagnostic checking of model residuals)

The Durbin-Watson test is invalid in the presence of lagged dependent variables. Breusch-Godfrey remains valid.

a. **Durbin's h test** is an alternative to the Durbin-Watson test. As 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.

b. Which do you think is likely to be most useful to you in assessing the properties of residuals from time-series models: the residual correlo-

gram, Durbin's h test, or the Breusch-Godfrey test? Why?

7. Dynamic logit.

Note that, in a logit regression, one or more of the RHS variables could be lagged dependent variables, $I_{t-i}(z)$, $i = 1, 2, \dots$

Chapter 13

Structural Change

Recall the full ideal conditions, one of which was that the model coefficients are fixed. Violations of that condition are of great concern in time series. The cross-section dummy variables that we already studied effectively allow for structural change in the cross section (heterogeneity across groups). But structural change is of special relevance in time series. It can be gradual (Lucas critique, learning, evolution of tastes, ...) or abrupt (e.g., new legislation).

Structural change is related to nonlinearity, because structural change is actually a *type* of nonlinearity. Structural change is also related to outliers, because outliers can sometimes be viewed as a kind of structural change – a quick intercept break and return.

For notational simplicity we consider the case of simple regression throughout, but the ideas extend immediately to multiple regression.

13.1 Gradual Parameter Evolution

In many cases, parameters may evolve gradually rather than breaking abruptly. Suppose, for example, that

$$y_t = \beta_{1t} + \beta_{2t}x_t + \varepsilon_t$$

where

$$\beta_{1t} = \gamma_1 + \gamma_2 TIME_t$$

$$\beta_{2t} = \delta_1 + \delta_2 TIME_t.$$

Then we have:

$$y_t = (\gamma_1 + \gamma_2 TIME_t) + (\delta_1 + \delta_2 TIME_t)x_t + \varepsilon_t.$$

We simply run:

$$y_t \rightarrow c, , TIME_t, x_t, TIME_t \cdot x_t.$$

This is yet another important use of dummies. The regression can be used both to test for structural change (F test of $\gamma_2 = \delta_2 = 0$), and to accommodate it if present.

13.2 Abrupt Parameter Breaks

13.2.1 Exogenously-Specified Breaks

Suppose that we don't know whether a break occurred, but we know that if it *did* occur, it occurred at time T^* .

A Dummy-Variable Approach That is, we entertain the possibility that

$$y_t = \begin{cases} \beta_1^1 + \beta_2^1 x_t + \varepsilon_t, & t = 1, \dots, T^* \\ \beta_1^2 + \beta_2^2 x_t + \varepsilon_t, & t = T^* + 1, \dots, T \end{cases}$$

Let

$$D_t = \begin{cases} 0, & t = 1, \dots, T^* \\ 1, & t = T^* + 1, \dots, T \end{cases}$$

Then we can write the model as:

$$y_t = (\beta_1^1 + (\beta_2^2 - \beta_1^1)D_t) + (\beta_2^1 + (\beta_2^2 - \beta_2^1)D_t)x_t + \varepsilon_t$$

We simply run:

$$y_t \rightarrow c, D_t, x_t, D_t \cdot x_t$$

The regression can be used both to test for structural change, and to accommodate it if present. It represents yet another use of dummies. The no-break null corresponds to the joint hypothesis of zero coefficients on D_t and $D_t \cdot x_t$, for which an F test is appropriate.

The Chow Test The dummy-variable setup and associated F test above is actually just a laborious way of calculating the so-called Chow breakpoint test statistic,

$$Chow = \frac{(SSR_{res} - SSR)/K}{SSR/(T - 2K)},$$

where SSR_{res} is from the regression using sample $t = 1, \dots, T$ and $SSR = SSR_1 + SSR_2$, where SSR_1 is from the regression using sample $t = 1, \dots, T^*$ and SSR_2 is from the regression using sample $t = T^* + 1, \dots, T$. Under the IC, $Chow$ is distributed F , with K and $T - 2K$ degrees of freedom.

13.2.2 The Chow test with Endogenous Break Selection

Thus far we have (unrealistically) assumed that the potential break date is known. In practice, potential break dates are often unknown and are identified by “peeking” at the data. We can capture this phenomenon in stylized fashion by imagining splitting the sample sequentially at each possible break date, and picking the split at which the Chow breakpoint test statistic is maximized. Implicitly, that’s what people often do in practice, even if they don’t always realize or admit it.

The distribution of such a test statistic is *not* F , as for the traditional Chow breakpoint test statistic. Rather, the distribution is that of the *maximum* of many draws from an F , which will be pushed far to the right of the distribution of a single F draw.

The test statistic is

$$\text{MaxChow} = \max_{\tau_1 \leq \tau \leq \tau_2} \text{Chow}(\tau),$$

where τ denotes sample fraction (typically we take $\tau_1 = .15$ and $\tau_2 = .85$). The distribution of *MaxChow* has been tabulated.

13.3 Dummy Variables and Omitted Variables, Again and Again

13.3.1 Dummy Variables

Notice that dummy (indicator) variables have arisen repeatedly in our discussions. We used 0-1 dummies to handle group heterogeneity in cross-sections. We used time dummies to indicate the date in time series. We used 0-1 seasonal dummies to indicate the season in time series.

Now, in this chapter, we used both (1) time dummies to allow for gradual parameter evolution, and (2) 0-1 dummies to indicate a sharp break date, in time series.

13.3.2 Omitted Variables

Notice that omitted variables have also arisen repeatedly in our discussions.

1. If there are neglected group effects in cross-section regression, we fix the problem (of omitted group dummies) by including the requisite group dummies.
2. If there is neglected trend or seasonality in time-series regression, we fix the problem (of omitted trend or seasonal dummies) by including the requisite trend or seasonal dummies.

3. If there is neglected non-linearity, we fix the problem (effectively one of omitted Taylor series terms) by including the requisite Taylor series terms.
4. If there is neglected structural change in time-series regression, we fix the problem (effectively one of omitted parameter trend dummies or break dummies) by including the requisite trend dummies or break dummies.

You can think of the basic “uber-strategy” as ”If some systematic feature of the DGP is missing from the model, then include it.” That is, if something is missing, then *model* what’s missing, and then the new uber-model won’t have anything missing, and all will be well (i.e., the IC will be satisfied). This is an important recognition. In a subsequent chapter, for example, we’ll study another violation of the IC known as serial correlation (Chapter ??). The problem amounts to a feature of the DGP neglected by the initially-fitted model, and we address the problem by incorporating the neglected feature into the model.

13.4 Recursive Analysis and CUSUM

13.5 Structural Change in Liquor Sales Trend

13.6 Exercises, Problems and Complements

1. Rolling Regression for Generic Structural Change ***

Chapter 14

Vector Autoregression

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*.¹ 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

¹Trends, seasonals, and other exogenous variables may also be included, as long as they're all included in every equation.

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

$$\varepsilon_{1,t} \sim WN(0, \sigma_1^2)$$

$$\varepsilon_{2,t} \sim WN(0, \sigma_2^2)$$

$$cov(\varepsilon_{1,t}, \varepsilon_{2,t}) = \sigma_{12}.$$

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. 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.

When fitting *VAR*'s to data, we can use the Schwarz criterion, just as in the univariate case. The formula differs, however, because we're now working with a multivariate system of equations rather than a single equation. To get an *SIC* value for a *VAR* system, we could add up the equation-by-equation *SIC*'s, 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 information criteria are required,

which account for cross-equation innovation correlation. We interpret the *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 *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 *VAR*'s makes the derivations more tedious, however, so we bypass them. As always, to construct practical forecasts we replace unknown parameters by estimates.

14.1 Predictive Causality

There's an important statistical notion of causality that's intimately related to forecasting and naturally introduced in the context of *VAR*'s. 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 *VAR*'s 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.² 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 forecasts of y_j ?,” and “Does y_j contribute toward improving forecasts of y_i ? ”

²Note 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 .

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:

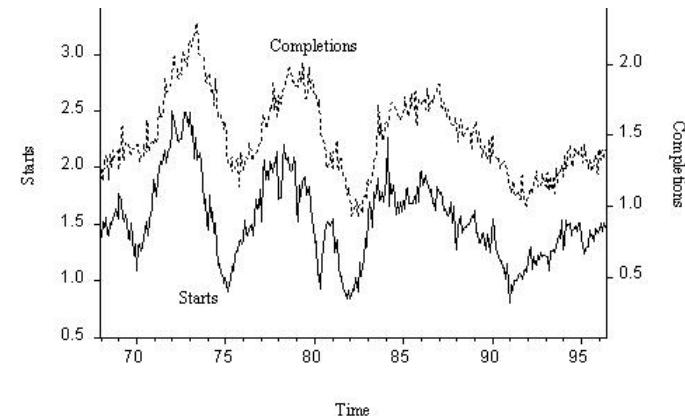
- (a) No lags of variable i aid in one-step-ahead prediction of variable j .

We can broaden the idea, however. Sometimes we test stronger noncausality hypotheses such as:

1. No lags of a set of other variables aid in one-step-ahead prediction of variable j .
2. No lags of any other variables aid in one-step-ahead prediction of variable j .
3. 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 .

14.2 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 14.1. 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.



Notes to figure: The left scale is starts, and the right scale is completions.

Figure 14.1: Housing Starts and Completions, 1968 - 1996

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 14.2 and Figure 14.3. 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 14.4 and Figure 14.5, 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* 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.

	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 14.2: Housing Starts Correlogram

The cross-correlation function (Figure 14.6) 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 SIC , we adopt a $VAR(4)$.

First consider the starts equation (Table 14.7a), residual plot (Figure 14.7b), and residual correlogram (Table 14.8, Figure 14.9). 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,

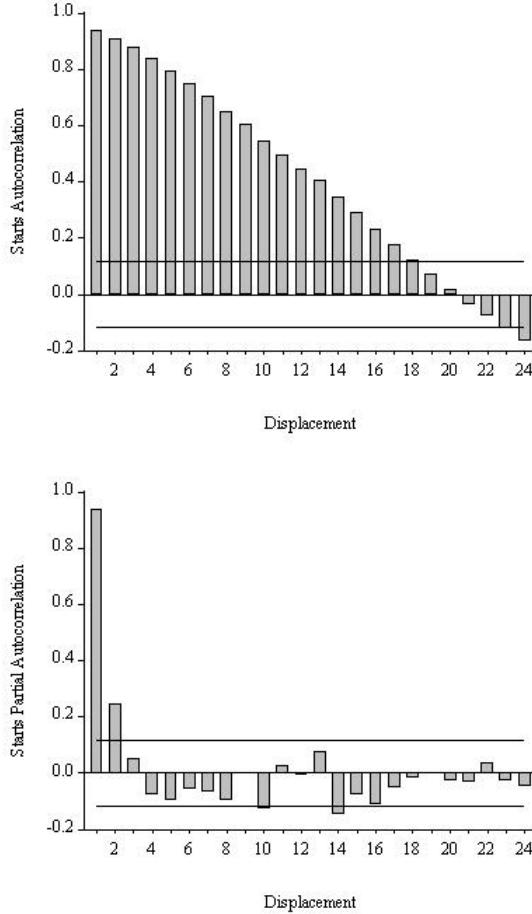


Figure 14.3: Housing Starts Autocorrelations and Partial Autocorrelations

but not conversely. The completions equation (Table 14.10a), residual plot (Figure 14.10b), and residual correlogram (Table 14.11, Figure 14.12) appear similarly good. Lagged starts, moreover, most definitely have a significant effect on completions.

Table 14.13 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-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	0.939	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 14.4: Housing Completions Correlogram

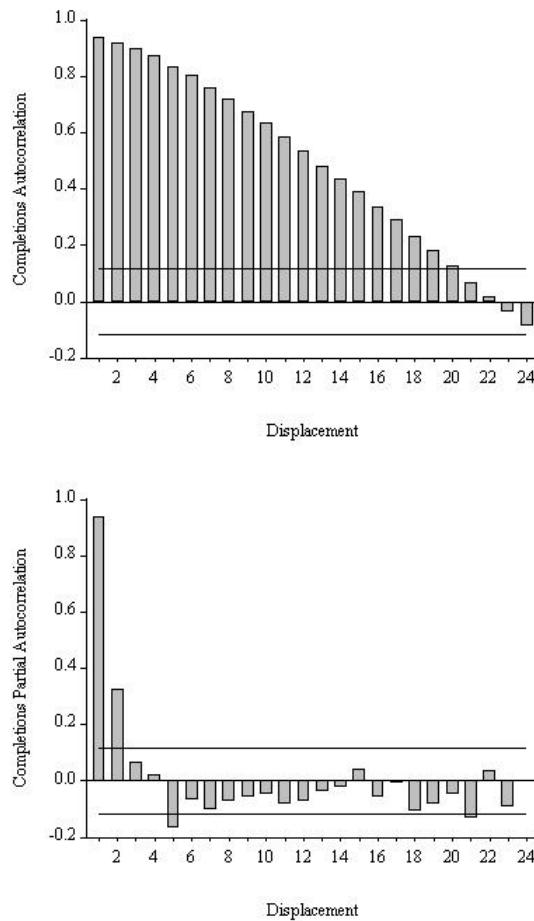
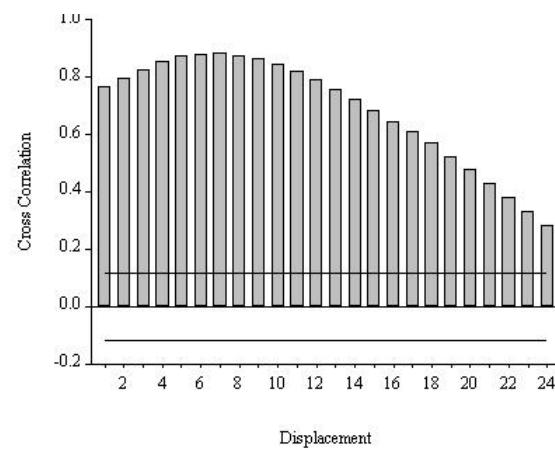


Figure 14.5: Housing Completions Autocorrelations and Partial Autocorrelations

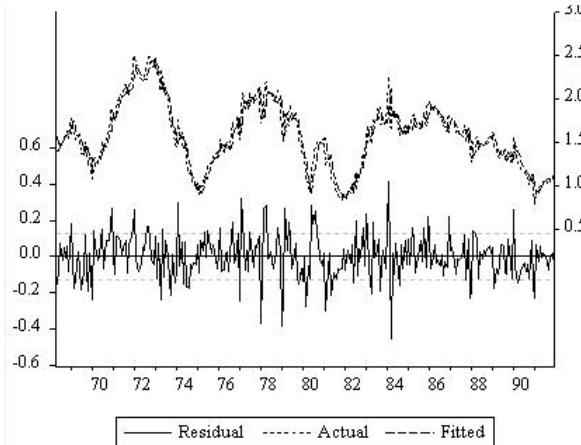


Notes to figure: We graph the sample correlation between completions at time t and starts at time $t-i$, $i = 1, 2, \dots, 24$.

Figure 14.6: 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 14.7: VAR Starts Model

	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 14.8: VAR Starts Residual Correlogram

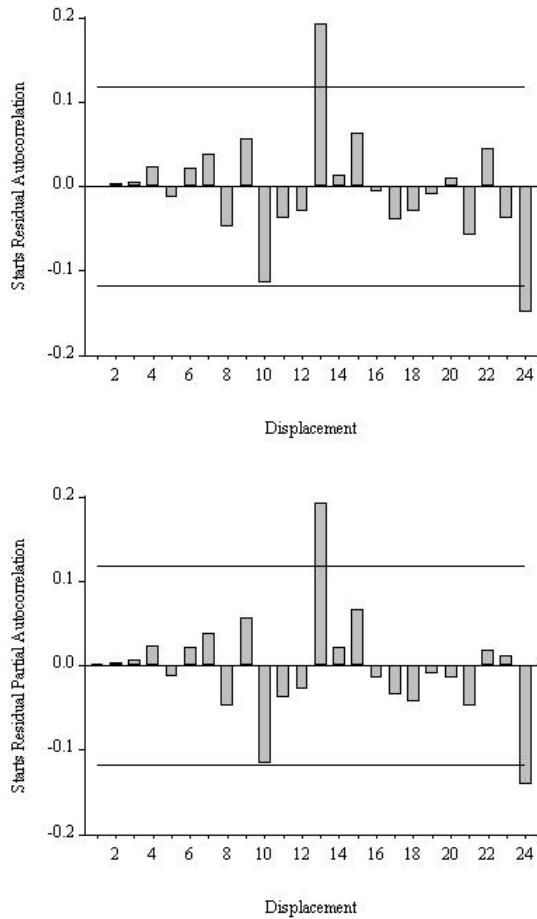
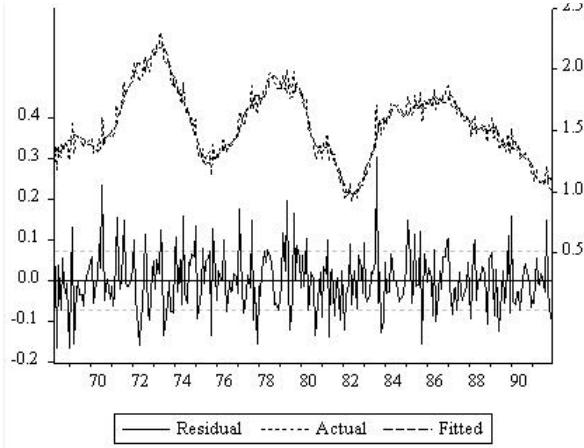


Figure 14.9: VAR Starts Equation - Sample Autocorrelation and Partial Autocorrelation

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 14.10: 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 14.11: VAR Completions Residual Correlogram

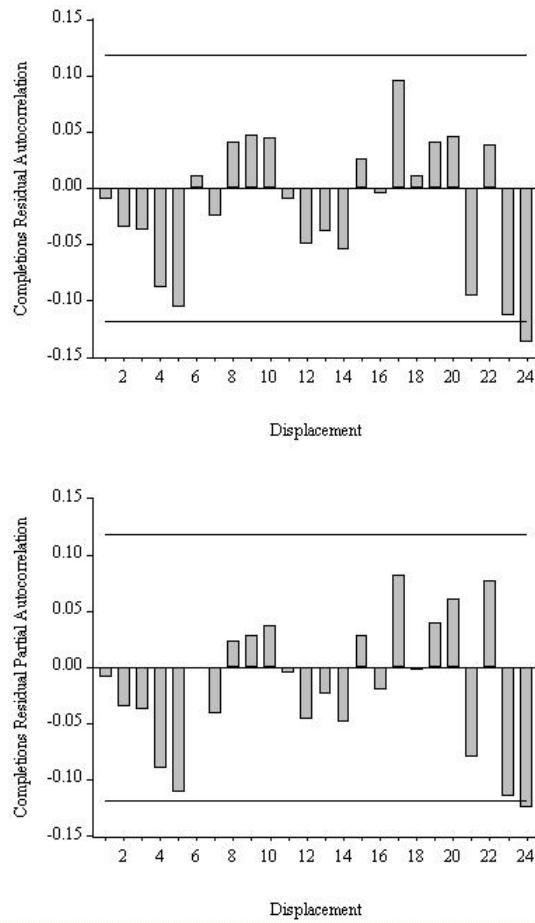


Figure 14.12: VAR Completions Equation - Sample Autocorrelation and Partial Autocorrelation

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 14.13: Housing Starts and Completions - Causality Tests

Finally, we construct forecasts for the out-of-sample period, 1992.01-1996.06. The starts forecast appears in Figure 14.14. Starts begin their recovery before 1992.01, and the *VAR* projects continuation of the recovery. The *VAR* forecasts 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 14.15. 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 14.16 and 14.17. Interestingly, however, completions had not yet turned by 1991.12, but the forecast nevertheless correctly predicts the turning point. (Why?)

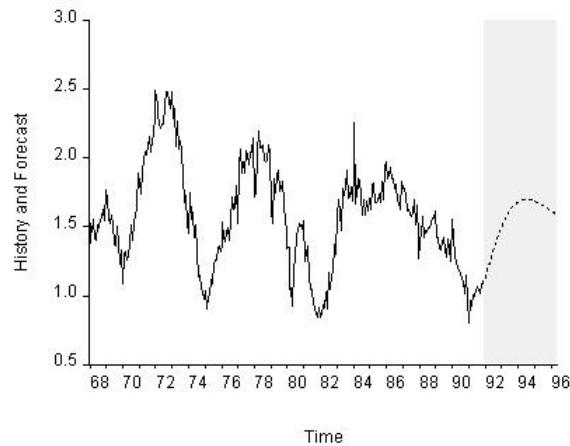


Figure 14.14: Housing Starts Forecast

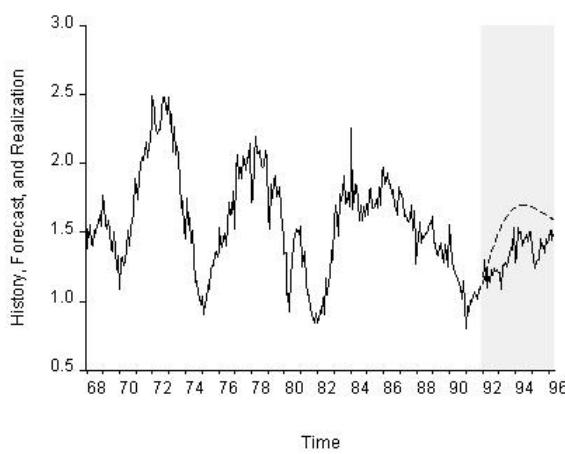


Figure 14.15: Housing Starts Forecast and Realization

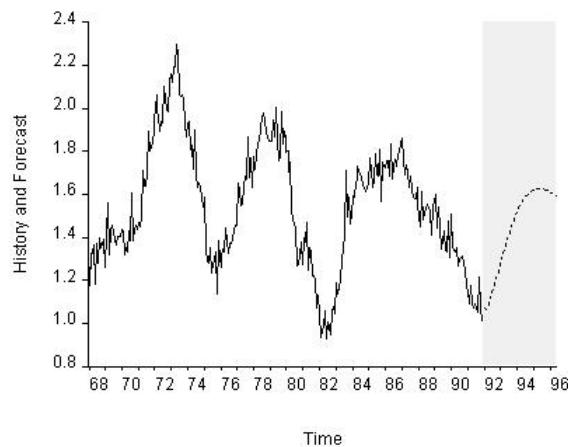


Figure 14.16: Housing Completions Forecast

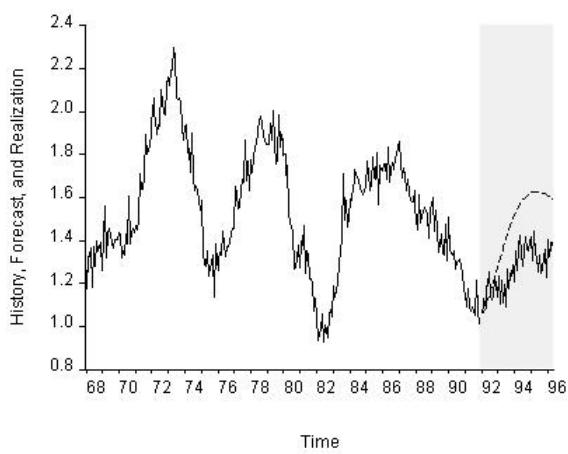


Figure 14.17: Housing Completions Forecast and Realization

14.3 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. 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?

Chapter 15

Dynamic Heteroskedasticity

Recall the full ideal conditions.

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 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 ε_t is white noise. In some situations it is inconsequential whether ε_t is weak or strong white noise, that is, whether ε_t is independent, as opposed to merely serially uncorrelated. Hence, to simplify matters we sometimes assume strong white noise, $\varepsilon_t \sim iid(0, \sigma^2)$. Throughout this book, we have thus far taken that approach, sometimes explicitly and sometimes implicitly.

When ε_t is independent, there is no distinction between the unconditional distribution of ε_t and the distribution of ε_t conditional upon its past, by definition of independence. Hence σ^2 is both the unconditional and conditional variance of ε_t . The Wold decomposition, however, does not require that ε_t be serially independent; rather it requires only that ε_t be serially uncorrelated.

If ε_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 σ_t^2 will in general evolve as Ω_{t-1} evolves, which focuses attention on the possibility of time-varying innovation volatility.¹

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 a 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 models that we have considered thus far, however, do not allow for that possibility. In this chapter we do so.

15.1 The Basic ARCH Process

Consider the general linear process,

$$\begin{aligned} y_t &= B(L)\varepsilon_t \\ B(L) &= \sum_{i=0}^{\infty} b_i L^i \\ \sum_{i=0}^{\infty} b_i^2 &< \infty \end{aligned}$$

¹In 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 ε are due exclusively to fluctuations in σ_t^2 .

$$b_0 = 1$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

We will work with various cases of this process.

Suppose first that ε_t is strong white noise, $\varepsilon_t \sim iid(0, \sigma^2)$. 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

$$E \left((y_{t+h} - E(y_{t+h}|\Omega_t))^2 | \Omega_t \right) = \sigma^2 \sum_{i=0}^{h-1} b_i^2.$$

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 Ω_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 ε_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

$$\begin{aligned}\varepsilon_t | \Omega_{t-1} &\sim N(0, \sigma_t^2) \\ \sigma_t^2 &= \omega + \gamma(L)\varepsilon_t^2 \\ \omega > 0, \gamma(L) = \sum_{i=1}^p \gamma_i L^i \gamma_i &\geq 0 \text{ for all } i \quad \sum \gamma_i < 1.\end{aligned}$$

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. ε_t is serially uncorrelated but not serially independent, because the current conditional variance σ_t^2 depends on the history of ε_t .² 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 ε_t are constant and are given by

$$E(\varepsilon_t) = 0$$

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 ε_t , its conditional variance

²In particular, σ_t^2 depends on the previous p values of ε_t via the distributed lag

$$\gamma(L)\varepsilon_t^2.$$

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 ε_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:

$$E(y_t | \Omega_{t-1}) = \sum_{i=1}^{\infty} b_i \varepsilon_{t-i}$$

$$E((y_t - E(y_t | \Omega_{t-1}))^2 | \Omega_{t-1}) = \omega + \gamma(L)\varepsilon_t^2.$$

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 Ω_{t-1} . In the above development, ε_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 ε_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.³ 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,

$$\sigma_t^2 = \omega + \gamma(L)\varepsilon_t^2,$$

³The variance of the disturbance in a model of household expenditure, for example, may depend on income.

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 *autoregressiveconditional heteroskedasticity*. To understand why, note that the ARCH conditional variance function links today's conditional variance positively to earlier lagged ε_t^2 's, so that large ε_t^2 's in the recent past produce a large conditional variance today, thereby increasing the likelihood of a large ε_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.”

15.2 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 by⁴

$$y_t = \varepsilon_t$$

$$\begin{aligned}\varepsilon_t | \Omega_{t-1} &\sim N(0, \sigma_t^2) \\ \sigma_t^2 &= \omega + \alpha(L)\varepsilon_t^2 + \beta(L)\sigma_t^2 \\ \alpha(L) &= \sum_{i=1}^p \alpha_i L^i, \beta(L) = \sum_{i=1}^q \beta_i L^i \\ \omega > 0, \alpha_i &\geq 0, \beta_i \geq 0, \sum \alpha_i + \sum \beta_i < 1.\end{aligned}$$

The stated conditions ensure that the conditional variance is positive and that y_t is covariance stationary.

Back substitution on σ_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.

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 ω . 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

⁴By “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.

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 uncondi-

tional 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

$$E(\varepsilon_{t+h}|\Omega_t) = 0,$$

and the corresponding prediction error is

$$\varepsilon_{t+h} - E(\varepsilon_{t+h}|\Omega_t) = \varepsilon_{t+h}.$$

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 Ω_t can be exploited to improve interval and density forecasts.

Fourth, consider the relationship between ε_t^2 and σ_t^2 . The relationship is important: GARCH dynamics in σ_t^2 turn out to introduce ARMA dynamics in ε_t^2 .⁵ More precisely, if ε_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

$$\nu_t = \varepsilon_t^2 - \sigma_t^2$$

is the difference between the squared innovation and the conditional variance at time t. To see this, note that if ε_t is GARCH(p,q), then

$$\sigma_t^2 = \omega + \alpha(L)\varepsilon_t^2 + \beta(L)\sigma_t^2.$$

Adding and subtracting

$$\beta(L)\varepsilon_t^2$$

⁵Put differently, the GARCH process approximates conditional variance dynamics in the same way that an ARMA process approximates conditional mean dynamics.

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 ν_t , where

$$\nu_t \in [-\sigma_t^2, \infty).$$

ε_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 σ_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 σ_t^2 , behaving similarly but not identically, with ν_t being the difference, or error. In particular, ε_t^2 is a *noisy* proxy: ε_t^2 is an unbiased estimator of σ_t^2 , but it is more volatile. It seems reasonable, then, that reconciling the noisy proxy ε_t^2 and the true underlying σ_t^2 should involve some sort of smoothing of ε_t^2 . Indeed, in the GARCH(1,1) case σ_t^2 is precisely obtained by exponentially smoothing ε_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.

15.3 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.⁶ Asymmetric response is often present, for example, in stock returns.

15.3.1 Asymmetric Response

The simplest GARCH model allowing for asymmetric response is the **threshold GARCH**, or TGARCH, model.⁷ We replace the standard GARCH conditional variance function, $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$, with $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2 D_{t-1} + \beta\sigma_{t-1}^2$, 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 α . 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 $\gamma < 0$.⁸ Asymmetric response may also be introduced via the **exponential GARCH** (EGARCH) model,

$$\ln(\sigma_t^2) = \omega + \alpha \left| \varepsilon_{\frac{t-1}{\sigma_{t-1}}} \right| + \gamma \varepsilon_{\frac{t-1}{\sigma_{t-1}}} + \beta \ln(\sigma_{t-1}^2).$$

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.⁹ The

⁶In the GARCH model studied thus far, only the *square* of last period's return affects the current conditional variance; hence its sign is irrelevant.

⁷For 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.

⁸Negative 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.

⁹The absolute “size” of news is captured by $|r_{t-1}/\sigma_{t-1}|$, and the sign is captured by r_{t-1}/σ_{t-1} .

log specification also ensures that the conditional variance is automatically positive, because σ_t^2 is obtained by exponentiating $\ln(\sigma_t^2)$; hence the name “exponential GARCH.”

15.3.2 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 γ is a parameter and x is a positive exogenous variable.¹⁰ Allowance for exogenous variables in the conditional variance function is sometimes useful. Financial market volume, for example, often helps to explain market volatility.

15.3.3 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

¹⁰Extension to allow multiple exogenous variables is straightforward.

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.¹¹

15.3.4 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.¹² This is precisely the GARCH(1,1) model introduced earlier, rewritten in a slightly different but equivalent form. In this model, short-run volatility dynamics are governed by the parameters α and β , and there are no long-run volatility dynamics, because $\bar{\omega}$ is constant. Sometimes we might want to allow for both long-run and 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 $q_t = \omega + \rho(q_{t-1} - \omega) + \phi(\varepsilon_{t-1}^2 - \sigma_{t-1}^2)$. 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 ρ , and the transitory dynamics are governed by α and β .¹³

¹¹One may also allow the conditional standard deviation, rather than the conditional variance, to enter the regression.

¹² $\bar{\omega}$ is sometimes called the “long-run” variance, referring to the fact that the unconditional variance is the long-run average of the conditional variance.

¹³It 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.

15.3.5 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: $(\sigma_t^2 - q_t) = \alpha(\varepsilon_{t-1}^2 - q_{t-1}) + \gamma(\varepsilon_{t-1}^2 - q_{t-1})D_{t-1} + \beta($
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.

15.4 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 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.¹⁴ No closed-form expression exists for the GARCH maximum likelihood estimator, so we must maximize the likelihood numerically.¹⁵ 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

$$\sigma_{t+h,t}^2 = E [\varepsilon_{t+h}^2 | \Omega_t] = \omega \left(\sum_{i=1}^{h-1} [\alpha(1) + \beta(1)]^i \right) + [\alpha(1) + \beta(1)]^{h-1} \sigma_{t+1}^2.$$

¹⁴The 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 Notes at the end of the chapter.

¹⁵Routines 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.

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 σ_{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 α and β 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 ε_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 interval forecasts of the form

$$y_{t+h,t} \pm 1.96\sigma_h,$$

where σ_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

$$y_{t+h,t} \pm 1.96\hat{\sigma}_{t+h,t}.$$

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, ε_t^2 . As discussed earlier, ε_t^2 is a proxy for the latent conditional variance; if the conditional variance displays persistence, so too will ε_t^2 .¹⁶ 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 $\varepsilon_t = \sigma_t v_t$, where $v_t \sim iidN(0, 1)$, $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$. Equivalently, the *standardized return*, v , is iid, $\varepsilon_t / \sigma_t = v_t \sim iidN(0, 1)$.

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, $\hat{\sigma}_t$, 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.

15.5 Exercises, Problems and Complements

1. (Graphical regression diagnostic: time series plot of e_t^2 or $|e_t|$)

Plots of e_t^2 or $|e_t|$ reveal patterns (most notably serial correlation) in the squared or *absolute* residuals, which correspond to non-constant volatility, or heteroskedasticity, in the levels of the residuals. As with the standard residual plot, the squared or absolute residual plot is always a simple univariate plot, even when there are many right-hand side variables. Such plots feature prominently, for example, in tracking and forecasting time-varying volatility.

2. (Removing conditional mean dynamics before modeling volatility dy-

¹⁶Note well, however, that the converse is not true. That is, if ε_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 ε_t and hence also in ε_t^2 . Thus, before proceeding to examine and interpret the correlogram of ε_t^2 as a check for volatility dynamics, it is important that any conditional mean effects be appropriately modeled, in which case ε_t should be interpreted as the disturbance in an appropriate conditional mean model.

namics)

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.
 - a. Estimate, diagnose and discuss a threshold GARCH(1,1) model.
 - b. Estimate, diagnose and discuss an EGARCH(1,1) model.
 - c. Estimate, diagnose and discuss a component GARCH(1,1) model.
 - d. 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.

- a. When fitting pure ARCH(p) models, what value of p seems adequate?
 - b. When fitting GARCH(p,q) models, what values of p and q seem adequate?
 - c. 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. (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. It can be shown that the Ljung-Box statistics may be legitimately used on

squared observed returns, in which case it will have the usual χ_m^2 distribution under the null hypothesis of independence. One may also use the Ljung-Box statistic on the squared standardized returns, but a better distributional approximation is obtained in that case by using a χ_{m-k}^2 distribution, where k is the number of estimated GARCH parameters, to account for degrees of freedom used in model fitting.

7. (Allowing for leptokurtic conditional densities)

Thus far we have worked exclusively with conditionally Gaussian GARCH models, which correspond to $\varepsilon_t = \sigma_t v_t$ $v_t \sim iidN(0, 1)$, or equivalently, to normality of the standardized return, ε_t/σ_t .

- a. The conditional normality assumption may sometimes be violated. However, 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 the 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

$$v_t \sim iid \frac{t_d}{std(t_d)}.$$

$$\varepsilon_t = \sigma_t v_t$$

What is the reason for dividing the Student's t variable, t_d , by its standard deviation, $std(t_d)$? How might such a model be estimated?

8. (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 -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 -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?

15.6 Notes

Part IV

Appendices

Appendix A

Probability and Statistics Review

Here we review a few aspects of probability and statistics that we will rely upon at various times.

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 $y = 2$.¹

Discrete random variables, that is, random variables with **discrete probability distributions**, can assume only a countable number of values 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

¹Note 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.

$y = 0$, $f(y) = 0.50$ 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 .²

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.³

The **mean**, or **expected value**, of a discrete random variable is a probability-weighted average of the values it can assume,⁴

$$E(y) = \sum_i p_i y_i.$$

Often we use the Greek letter μ to denote the mean, which measures the **location**, or **central tendency**, of y .

The **variance** of y is its expected squared deviation from its mean,

$$\text{var}(y) = E(y - \mu)^2.$$

We use σ^2 to denote the variance, which measures the **dispersion, or scale**, of y around its mean.

²In addition, the total area under $f(y)$ must be 1.

³In principle, we could of course consider moments beyond the fourth, but in practice only the first four are typically examined.

⁴A similar formula holds in the continuous case, $E(y) = \int y f(y) dy$.

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 σ^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 σ^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 .⁵ 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,

$$\text{cov}(y, x) = E[(y - \mu_y)(x - \mu_x)].$$

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.

Frequently we convert the covariance to a **correlation** by standardizing by the product of σ_y and σ_x ,

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

⁵We could of course consider more than two variables, but for pedagogical reasons we presently limit ourselves to two.

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**.⁶ We can obtain estima-

⁶An estimator is an example of a **statistic**, or **sample statistic**, which is simply a function of the sample observations.

tors 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 σ^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 .

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 χ^2 **distribution** arises from squared normal random variables, the t **distribution** arises from ratios of normal and χ^2 variables, and the F **distribution** arises from ratios of χ^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.

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.

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 μ . In this case, as you learned earlier, \bar{y} is unbiased, consistent, normally distributed with variance σ^2/N , and efficient (minimum variance unbiased, MVUE). 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 estimate σ^2 consistently using s^2 .

We construct exact finite-sample confidence intervals for μ as

$$\mu \in \left[\bar{y} \pm t_{1-\frac{\alpha}{2}}(N-1) \frac{s}{\sqrt{N}} \right] \text{ w.p. } 1 - \alpha,$$

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 consistent, asymptotically normally distributed with variance σ^2/N , and asymptotically efficient. We write,

$$\bar{y} \xrightarrow{a} N\left(\mu, \frac{\sigma^2}{N}\right).$$

More precisely, as $T \rightarrow \infty$,

$$\sqrt{N}(\bar{y} - \mu) \rightarrow_d N(0, \sigma^2).$$

We estimate σ^2 consistently using s^2 .

This result forms the basis for asymptotic inference. We construct asymptotically-

valid confidence intervals for μ as

$$\mu \in \left[\bar{y} \pm z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{N}} \right] \text{ w.p. } 1 - \alpha,$$

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

$$\frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{N}}} \sim N(0, 1).$$

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 of various types, such as $iidN(\mu, \sigma^2)$ (Gaussian simple random sampling).
- 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 CPS wage data)

Use the 1995 CPS 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.

Appendix B

Construction of the Wage Datasets

We construct our datasets from randomly sampling the much-larger Current Population Survey (CPS) datasets.¹

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). We use both personal and family records. Here we focus our discussion on 1995.

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 (_ERNEL=0), leaving 14363 observations. From those, we draw a random unweighted subsample with ten percent selection probability. This results in 1348 observations.

We use seven variables. From the CPS we obtain AGE (age), FEMALE (1 if female, 0 otherwise), NONWHITE (1 if nonwhite, 0 otherwise), and UNION (1 if union member, 0 otherwise). We also create EDUC (years of schooling) based on CPS variable PEEDUCA (educational attainment). Because the CPS does not ask about years of experience, we create EXPER

¹See <http://aspe.hhs.gov/hsp/06/catalog-ai-an-na/cps.htm> for a brief and clear introduction to the CPS datasets.

(potential working experience) as AGE minus EDUC minus 6.

We construct the variable WAGE as follows. 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, NON-WHITE, UNION, EDUC, EXPER and WAGE.

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)

CPS Personal Data Selection Criteria

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.

Variable List

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.)

Definition of EDUC

Appendix C

Some Popular Books Worth Encountering

I have cited many of these books elsewhere, typically in various end-of-chapter complements. Here I list them collectively.

Lewis (2003) [Michael Lewis, *Moneyball*]. “Appearances may lie, but the numbers don’t, so pay attention to the numbers.”

Gladwell (2000) [Malcolm Gladwell, *The Tipping Point*]. “Nonlinear phenomena are everywhere.”

Gladwell pieces together an answer to the puzzling question of why certain things “take off” whereas others languish (products, fashions, epidemics, etc.) More generally, he provides deep insights into nonlinear environments, in which small changes in inputs can lead to small changes in outputs under some conditions, and to *huge* changes in outputs under other conditions.

Taleb (2007) [Nassim Nicholas Taleb, *The Black Swan*] “Warnings, and more warnings, and still more warnings, about non-normality and much else.” See Chapter 5 EPC 1.

Angrist and Pischke (2009) [Joshua Angrist and Jörn-Steffen Pischke, *Mostly Harmless Econometrics*]. “Natural and quasi-natural experiments suggesting instruments.”

This is a fun and insightful treatment of instrumental-variables and related

methods. Just don't be fooled by the book's attempted landgrab, as discussed in a [2015 *No Hesitations*](#) post.

[Silver \(2012\)](#) [Nate Silver, *The Signal and the Noise*]. "Pitfalls and opportunities in predictive modeling."

Bibliography

- Angrist, J.D. and J.-S. Pischke (2009), *Mostly Harmless Econometrics*, Princeton University Press.
- Gladwell, M. (2000), *The Tipping Point*, Little, Brown and Company.
- Harvey, A.C. (1991), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press.
- Lewis, M. (2003), *Moneyball*, Norton.
- Nerlove, M., D.M. Grether, and J.L. Carvalho (1979), *Analysis of Economic Time Series: A Synthesis*. New York: Academic Press. Second Edition.
- Silver, N.. (2012), *The Signal and the Noise*, Penguin Press.
- Taleb, N.N. (2007), *The Black Swan*, Random House.
- Tufte, E.R. (1983), *The Visual Display of Quantitative Information*, Cheshire: Graphics Press.