

Econometrics

Streamlined, Applied and e-Aware



Francis X. Diebold
University of Pennsylvania

Edition 2015
Version Tuesday 17th February, 2015

Econometrics

Econometrics

Streamlined, Applied and e-Aware

Francis X. Diebold

Copyright © 2013, 2014, 2015
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. (20XX), *Book Title Here in Italics*, 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	xxi
About the Cover	xxiii
Guide to e-Features	xxv
Preface	xxxiii
1 Introduction to Econometrics	1
2 Graphics and Graphical Style	11
3 Regression Analysis	27
4 Indicator Variables in Cross Sections	63
5 Indicator Variables in Time Series	71
6 Non-Linearity in Cross Sections	85
7 Non-Linearity in Time Series	99
8 Binary Regression and Classification	111
9 Measurement Error	121
10 Omitted Variables	123
11 Multicollinearity	127
12 Non-normality and Outliers	131

13 Heteroskedasticity in Cross-Sections	143
14 Serial Correlation in Time Series	151
15 Structural Change	255
16 Heteroskedasticity in Time Series	263
17 Endogeneity	305
18 Nonstationarity	315
19 Big Data: Selection, Shrinkage and Distillation	361
I Appendices	371
A Elements of Probability and Statistics	373
B Construction of the Wage Datasets	385
C Some Popular Books Worth Encountering	391

Detailed Table of Contents

About the Author	xxi
About the Cover	xxiii
Guide to e-Features	xxv
Preface	xxxiii
1 Introduction to Econometrics	1
1.1 Welcome	1
1.1.1 Who Uses Econometrics?	1
1.1.2 What Distinguishes Econometrics?	3
1.2 Types of Recorded Economic Data	3
1.3 Online Information and Data	4
1.4 Software	5
1.5 Tips on How to use this book	7
1.6 Exercises, Problems and Complements	8
1.7 Notes	9
2 Graphics and Graphical Style	11
2.1 Simple Techniques of Graphical Analysis	11
2.1.1 Univariate Graphics	12
2.1.2 Multivariate Graphics	12
2.1.3 Summary and Extension	15
2.2 Elements of Graphical Style	17
2.3 U.S. Hourly Wages	19
2.4 Concluding Remarks	19
2.5 Exercises, Problems and Complements	20
2.6 Notes	24

2.7	Graphics Legend: Edward Tufte	25
3	Regression Analysis	27
3.1	Preliminary Graphics	27
3.2	Regression as Curve Fitting	29
3.2.1	Bivariate, or Simple, Linear Regression	29
3.2.2	Multiple Linear Regression	32
3.2.3	Onward	33
3.3	Regression as a Probability Model	33
3.3.1	A Population Model and a Sample Estimator	33
3.3.2	Notation, Assumptions and Results	34
A Bit of Matrix Notation	35	
Assumptions: The Full Ideal Conditions (FIC)	36	
Results	37	
3.4	A Wage Equation	38
3.4.1	Mean dependent var 2.342	41
3.4.2	S.D. dependent var .561	41
3.4.3	Sum squared resid 319.938	42
3.4.4	Log likelihood -938.236	42
3.4.5	F statistic 199.626	43
3.4.6	Prob(F statistic) 0.000000	43
3.4.7	S.E. of regression .492	43
3.4.8	R-squared .232	45
3.4.9	Adjusted R-squared .231	45
3.4.10	Akaike info criterion 1.423	46
3.4.11	Schwarz criterion 1.435	46
3.4.12	Hannan-Quinn criter. 1.427	47
3.4.13	Durbin-Watson stat. 1.926	47
3.4.14	The Residual Scatter	48
3.4.15	The Residual Plot	49
3.5	Quantile Regression	50
3.6	Exercises, Problems and Complements	53
3.7	Notes	59
3.8	Regression's Inventor: Carl Friedrich Gauss	61

4 Indicator Variables in Cross Sections	63
4.1 0-1 Dummy Variables	63
4.2 Group Dummies in the Wage Regression	65
4.3 Exercises, Problems and Complements	67
4.4 Notes	69
4.5 Dummy Variables, ANOVA, and Sir Ronald Fischer	70
5 Indicator Variables in Time Series	71
5.1 Linear Trend	71
5.2 Seasonality	73
5.2.1 Seasonal Dummies	74
5.2.2 More General Calendar Effects	76
5.3 Trend and Seasonality in Liquor Sales	76
5.4 Exercises, Problems and Complements	78
5.5 Notes	82
6 Non-Linearity in Cross Sections	85
6.1 Models Linear in Transformed Variables	85
6.1.1 Logarithms	85
6.1.2 Box-Cox and GLM	87
6.2 Intrinsically Non-Linear Models	88
6.2.1 Nonlinear Least Squares	89
6.3 Interactions	90
6.4 A Final Word on Nonlinearity and the FIC	90
6.5 Testing for Non-Linearity	91
6.5.1 <i>t</i> and <i>F</i> Tests	91
6.5.2 The RESET Test	91
6.6 Non-Linearity in Wage Determination	92
6.6.1 Non-Linearity in Continuous and Discrete Variables Simultaneously	94
6.7 Exercises, Problems and Complements	96
6.8 Notes	98
7 Non-Linearity in Time Series	99
7.1 Exponential Trend	99
7.2 Quadratic Trend	101
7.3 More on Non-Linear Trend	102

7.3.1	Moving-Average Trend and De-Trending	102
7.3.2	Hodrick-Prescott Trend and De-Trending	104
7.4	Non-Linearity in Liquor Sales Trend	104
7.5	Exercises, Problems and Complements	105
7.6	Notes	109
8	Binary Regression and Classification	111
8.1	Binary Regression	111
8.1.1	Binary Response	111
8.1.2	The Logit Model	113
Logit	113
Ordered Logit	114
Dynamic Logit	115
Complications	115
8.1.3	Classification and “0-1 Forecasting”	116
8.2	Exercises, Problems and Complements	117
8.3	Notes	120
9	Measurement Error	121
9.1	Exercises, Problems and Complements	121
9.2	Notes	121
10	Omitted Variables	123
10.1	Omitted Relevant Variables	123
10.2	Included Irrelevant Variables	124
10.3	Exercises, Problems and Complements	124
10.4	Notes	125
11	Multicollinearity	127
11.1	Perfect Multicollinearity	127
11.2	Imperfect Multicollinearity	128
11.3	A Bit More	128
11.4	Exercises, Problems and Complements	129
11.5	Notes	129
12	Non-normality and Outliers	131
12.1	Non-Normality	131

12.1.1 OLS Without Normality	132
12.1.2 Assessing Normality	134
Nonparametric Density Estimation	134
The Residual QQ Plot	134
The Jarque-Bera Test	135
12.2 Outliers and Leverage	135
12.2.1 Outliers	135
12.2.2 Outlier Detection	136
12.2.3 Leverage	137
12.2.4 Robust Estimation	137
Robustness Iteration	138
Least Absolute Deviations	139
12.2.5 Wages and Liquor Sales	139
Wages	139
Liquor Sales	140
12.3 Exercises, Problems and Complements	140
12.4 Notes	141
13 Heteroskedasticity in Cross-Sections	143
13.1 Exercises, Problems and Complements	150
13.2 Notes	150
14 Serial Correlation in Time Series	151
14.1 Characterizing Time-Series Dynamics	151
14.1.1 Covariance Stationary Time Series	152
14.2 White Noise	157
14.3 Estimation and Inference for the Mean, Autocorrelation and Partial Autocorrelation Functions	162
14.3.1 Sample Mean	162
14.3.2 Sample Autocorrelations	163
14.3.3 Sample Partial Autocorrelations	166
14.4 Autoregressive Models for Serially-Correlated Time Series	167
14.4.1 Some Preliminary Notation: The Lag Operator	167
14.4.2 Autoregressions	169
The $AR(1)$ Process	169
14.4.3 The AR(p) Process	175
14.4.4 Alternative Approaches to Estimating Autoregressions	177

14.5 Serial Correlation in Time-Series Regression	179
14.5.1 Serial Correlation in Time-Series Regression	179
14.5.2 Testing for Serial Correlation	181
The Durbin-Watson Test	181
The Breusch-Godfrey Test	184
The Residual Correlogram	186
14.5.3 Estimation with Serial Correlation	187
14.5.4 Regression with Serially-Correlated Disturbances . . .	187
14.5.5 Serially-Correlated Disturbances vs. Lagged Dependent Variables	189
14.5.6 A Full Model of Liquor Sales	193
14.6 Vector Autoregression	196
14.6.1 Distributed Lag Models	207
14.6.2 Regressions with Lagged Dependent Variables, and Regressions with <i>AR</i> Disturbances	218
14.6.3 Vector Autoregressions	221
14.6.4 Predictive Causality	223
14.6.5 Impulse-Response Functions	225
14.6.6 Housing Starts and Completions	230
14.7 Exercises, Problems and Complements	232
14.8 Notes	237
14.9 Christopher A. Sims	239
15 Structural Change	255
15.1 Gradual Parameter Evolution	256
15.2 Sharp Parameter Breaks	256
15.2.1 Exogenously-Specified Breaks	256
15.2.2 Endogenously-Selected Breaks	257
15.3 Recursive Estimation	258
15.3.1 Recursive Residuals	258
15.3.2 Standardized Recursive Residuals and CUSUM . . .	259
15.4 Liquor Sales	259
15.5 Exercises, Problems and Complements	259
15.6 Notes	261

16 Heteroskedasticity in Time Series	263
16.1 The Basic ARCH Process	264
16.2 The GARCH Process	269
16.3 Extensions of ARCH and GARCH Models	276
16.3.1 Asymmetric Response	277
16.3.2 Exogenous Variables in the Volatility Function	278
16.3.3 Regression with GARCH disturbances and GARCH-M	278
16.3.4 Component GARCH	279
16.3.5 Mixing and Matching	280
16.4 Estimating, Forecasting and Diagnosing GARCH Models . . .	280
16.5 Stock Market Volatility	282
16.6 Exercises, Problems and Complements	297
16.7 Notes	303
17 Endogeneity	305
17.1 A Key Subtlety: Causal vs. Non-Causal Regression	305
17.1.1 Causal Predictive Modeling and T-Consistency	306
17.1.2 Non-Causal Predictive Modeling and P-Consistency . .	307
17.1.3 Correlation vs. Causality, and P-Consistency vs. T- Consistency	307
17.1.4 An Example of Correlation Without Causality	308
17.2 Causes of $E(X'\varepsilon) \neq 0$	309
17.2.1 Omitted Variables	309
Over-Controlling I: Controls Can Induce Multicollinearity	309
Over-Controlling II: “Controls” Can Introduce More Endogeneity!	309
17.2.2 Simultaneity	310
17.2.3 Measurement Error	310
17.2.4 Sample Selection	310
17.3 Instrumental Variables	310
17.3.1 The Basic Idea and the IV Estimator	310
17.3.2 Instrument Strength and Exogeneity	310
Weak Instruments	311
“Slightly-Endogenous” Instruments	311
17.3.3 Sources of Instruments	311
Randomized Experiments	311

Natural Experiments	311
Thought Experiments (Structural Econometric Models)	311
Time	311
17.4 Additional Useful Strategies	312
17.4.1 Regression Discontinuity Designs	312
17.4.2 Differences of Differences	312
17.4.3 Matching	312
17.5 “Graphical Models”	312
17.6 Internal and External Validity	312
17.7 Exercises, Problems and Complements	312
17.8 Notes	313
18 Nonstationarity	315
18.1 Nonstationary Series	315
18.1.1 Stochastic Trends and Forecasting	315
18.1.2 Unit-Roots: Estimation and Testing	322
Least-Squares Regression with Unit Roots	322
Effects of Unit Roots on the Sample Autocorrelation and Partial Autocorrelation Functions	324
Unit Root Tests	325
18.1.3 Smoothing	331
Moving Average Smoothing, Revisited	333
Exponential Smoothing	334
Holt-Winters Smoothing	336
Holt-Winters Smoothing with Seasonality	337
Forecasting with Smoothing Techniques	338
18.2 Exercises, Problems and Complements	353
18.3 Notes	360
19 Big Data: Selection, Shrinkage and Distillation	361
19.1 Good and Bad Approaches to Model Selection	361
19.1.1 MSE and R^2	361
19.1.2 s^2 and \bar{R}^2	362
19.2 All Subsets Selection	362
19.2.1 Information Criteria	362
19.2.2 Cross Validation	363
19.3 Stepwise Selection	364

19.3.1 Backward Stepwise Selection	364
19.3.2 Forward Stepwise Selection	364
19.4 Shrinkage	365
19.4.1 Bayesian Shrinkage	365
19.4.2 Ridge Shrinkage	365
19.5 Shrinkage and Selection	365
19.5.1 The Lasso	366
19.5.2 Elastic Net	367
19.5.3 Adaptive Lasso	367
19.5.4 Adaptive Elastic Net	368
19.6 Distillation: Principal Components	368
19.6.1 Distilling “ X Variables” into Principal Components . .	368
19.6.2 Principal Components Regression	369
19.7 Exercises, Problems and Complements	369
19.8 Notes	369
I Appendices	371
A Elements of Probability and Statistics	373
A.1 Populations: Random Variables, Distributions and Moments .	373
A.1.1 Univariate	373
A.1.2 Multivariate	376
A.2 Samples: Sample Moments	377
A.2.1 Univariate	377
A.2.2 Multivariate	379
A.3 Finite-Sample and Asymptotic Sampling Distributions of the Sample Mean	380
A.3.1 Exact Finite-Sample Results	380
A.3.2 Approximate Asymptotic Results (Under Weaker Assumptions)	381
A.4 Exercises, Problems and Complements	382
A.5 Notes	384
B Construction of the Wage Datasets	385
C Some Popular Books Worth Encountering	391

About the Author



Francis X. Diebold is P.F. and W.S. Miller Professor of Economics, and Professor of Finance and Statistics, at the University of Pennsylvania. He 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. Diebold has received several awards for outstanding teaching, and his academic "family" includes nearly 75 Ph.D. students.

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.

Guide to e-Features

- 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.
- 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.
- 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.

List of Figures

1.1	Resources for Economists Web Page	5
1.2	Eviews Homepage	6
1.3	Stata Homepage	7
1.4	R Homepage	8
2.1	1-Year Goverment Bond Yield, Levels and Changes	13
2.2	Histogram of 1-Year Government Bond Yield	14
2.3	Bivariate Scatterplot, 1-Year and 10-Year Government Bond Yields	15
2.4	Scatterplot Matrix, 1-, 10-, 20- and 30-Year Government Bond Yields	16
2.5	Distributions of Wages and Log Wages	19
2.6	Tufte Teaching, with a First Edition Book by Galileo	25
3.1	Distributions of Log Wage, Education and Experience	28
3.2	(Log Wage, Education) Scatterplot	30
3.3	(Log Wage, Education) Scatterplot with Superimposed Regression Line	31
3.4	Regression Output	38
3.5	Wage Regression Residual Scatter	48
3.6	Wage Regression Residual Plot	50
3.7	Carl Friedrich Gauss	61
4.1	Histograms for Wage Covariates	64
4.2	Wage Regression on Education and Experience	66
4.3	Wage Regression on Education, Experience and Group Dummies	66
4.4	Residual Scatter from Wage Regression on Education, Experience and Group Dummies	67
4.5	Sir Ronald Fischer	70

5.1	Various Linear Trends	72
5.2	Liquor Sales	77
5.3	Log Liquor Sales	78
5.4	Linear Trend Estimation	79
5.5	Residual Plot, Linear Trend Estimation	80
5.6	Estimation Results, Linear Trend with Seasonal Dummies	81
5.7	Residual Plot, Linear Trend with Seasonal Dummies	82
5.8	Seasonal Pattern	83
6.1	Basic Linear Wage Regression	92
6.2	Quadratic Wage Regression	93
6.3	Wage Regression on Education, Experience, Group Dummies, and Interactions	94
6.4	Wage Regression with Continuous Non-Linearities and Inter- actions, and Discrete Interactions	95
7.1	Various Exponential Trends	100
7.2	Various Quadratic Trends	102
7.3	Log-Quadratic Trend Estimation	105
7.4	Residual Plot, Log-Quadratic Trend Estimation	106
7.5	Liquor Sales Log-Quadratic Trend Estimation with Seasonal Dummies	107
7.6	Residual Plot, Liquor Sales Log-Quadratic Trend Estimation With Seasonal Dummies	108
14.1	***. ***.	190
14.2	***. ***.	190
14.3	***. ***.	191
14.4	***. ***.	193
14.5	Christopher Sims	239
15.1	Recursive Analysis, Constant Parameter	260
15.2	Recursive Analysis, Breaking Parameter	261
16.1	Time Series of Daily NYSE Returns.	285
16.2	Correlogram of Daily Stock Market Returns.	285
16.3	Histogram and Statistics for Daily NYSE Returns.	286
16.4	Time Series of Daily Squared NYSE Returns	287

16.5 Correlogram of Daily Squared NYSE Returns.	287
16.6 GARCH(1,1) Estimation, Daily NYSE Returns.	295
16.7 Estimated Conditional Standard Deviation, Daily NYSE Re- turns.	296
16.8 Conditional Standard Deviation, History and Forecast, Daily NYSE Returns.	296
16.9 Correlogram of Squared Standardized GARCH(1,1) Residuals, Daily NYSE Returns.	297
16.10 AR(1) Returns with Threshold t-GARCH(1,1)-in Mean. . . .	299
19.1 Degrees-of-Freedom Penalties	363

List of Tables

2.1 Yield Statistics	23
--------------------------------	----

Preface

Most good texts arise from the desire to leave one’s stamp on a discipline by training future generations of students, driven by the recognition that existing texts are deficient in various respects. My motivation is no different, but it is more intense: In recent years I have come to see most existing texts as highly deficient, in four ways.

First, many existing texts attempt exhaustive coverage, resulting in large tomes impossible to cover in a single course (or even two, or three). *Econometrics*, in contrast, does not attempt exhaustive coverage. Indeed the coverage is intentionally selective and streamlined, focusing on the core methods with the widest applicability. Put differently, *Econometrics* is not designed to impress people with the breadth of my knowledge; rather, it’s designed to teach real students, and it can be realistically covered in a one-semester course. Core material appears in the main text, and additional material appears in the end-of-chapter “Exercises, Problems and Complements.”

Second, many existing texts emphasize theory at the expense of serious applications. *Econometrics*, in contrast, is applications-oriented throughout, using detailed real-world applications not simply to illustrate theory, *but to teach it* (in truly realistic situations in which not everything works perfectly!). *Econometrics* uses modern software throughout, but the discussion is not wed to any particular software – students and instructors can use whatever computing environment they like best.

Third, although *Econometrics* is designed for a first course in econometrics and uses only elementary mathematics, it nevertheless conveys a strong feel for the important advances made in recent years. It touches – sometimes extensively – upon topics such as:

- statistical graphics and exploratory data analysis
- nonlinear and non-Gaussian environments
- simulation methods
- unit roots and stochastic trends
- volatility modeling
- Big Data: selection, shrinkage, etc.

Much such material appears in the already-mentioned Exercises, Problems and Complements, which form an integral part of the book. They are organized so that instructors and students can pick and choose, according to their backgrounds and interests. Coverage of such topics helps prepare students for more advanced or specialized books, such as my *Forecasting*.

Finally, almost all existing texts remain shackled by Middle-Ages paper technology. *Econometrics*, in contrast, is e-aware. It's colorful, hyperlinked internally and externally, and tied to a variety of media – effectively a blend of a traditional “book”, a DVD, a web page, a Facebook group, a blog, and whatever else I can find that's useful. It's continually evolving and improving on the web, and it's freely available to all, as opposed to the obscene but now-standard \$250 for a pile of paper. It won't make me any new friends among the traditional publishers, but that's not my goal.

Econometrics should be useful to students in a variety of fields – in economics, of course, but also business, finance, public policy, statistics, and even engineering. It is directly accessible at the undergraduate and master's levels,

as the only prerequisite is an introductory probability and statistics course. 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.

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; James H. Stock, Harvard University; Mark W. 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.

Graduate students include Ross Askanazi, Lorenzo Braccini, Laura Liu, Minchul Shin, and Molin Zhong.

Undergraduate students include Cathy Chen, Tingyan Jia, Mai Li, M.D. Mangini, Ian Masters, Joonyup Park, John Ro, Carlos Rodriguez, Ken Teoh, Han Tian, and Zach Winston.

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

Tuesday 17th February, 2015

Econometrics

Chapter 1

Introduction to Econometrics

1.1 Welcome

1.1.1 Who Uses Econometrics?

Econometric modeling 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 econometrics applications, let's sketch 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, new market entry, and so on.

More generally, 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, have been interested in empirical modeling and understanding of asset returns such as stock returns, interest rates, exchange rates, and 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 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 must confront the fact that economic data is not generated from well-designed experiments. On the contrary, econometricians must generally take whatever so-called “observational data” they’re given.
- Econometrics must confront the special issues and features that arise routinely in economic data, such as trends, seasonality and cycles.
- Econometricians are sometimes interested in non-causal predictive modeling, which requires understanding only correlations (or, more precisely, conditional expectations), and sometimes interested in evaluating treatment effects, which involve deeper issues of causation.

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 data for, say, 1960.I 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. Finally, check out [Quandl](#), which provides access to millions of data series on the web.

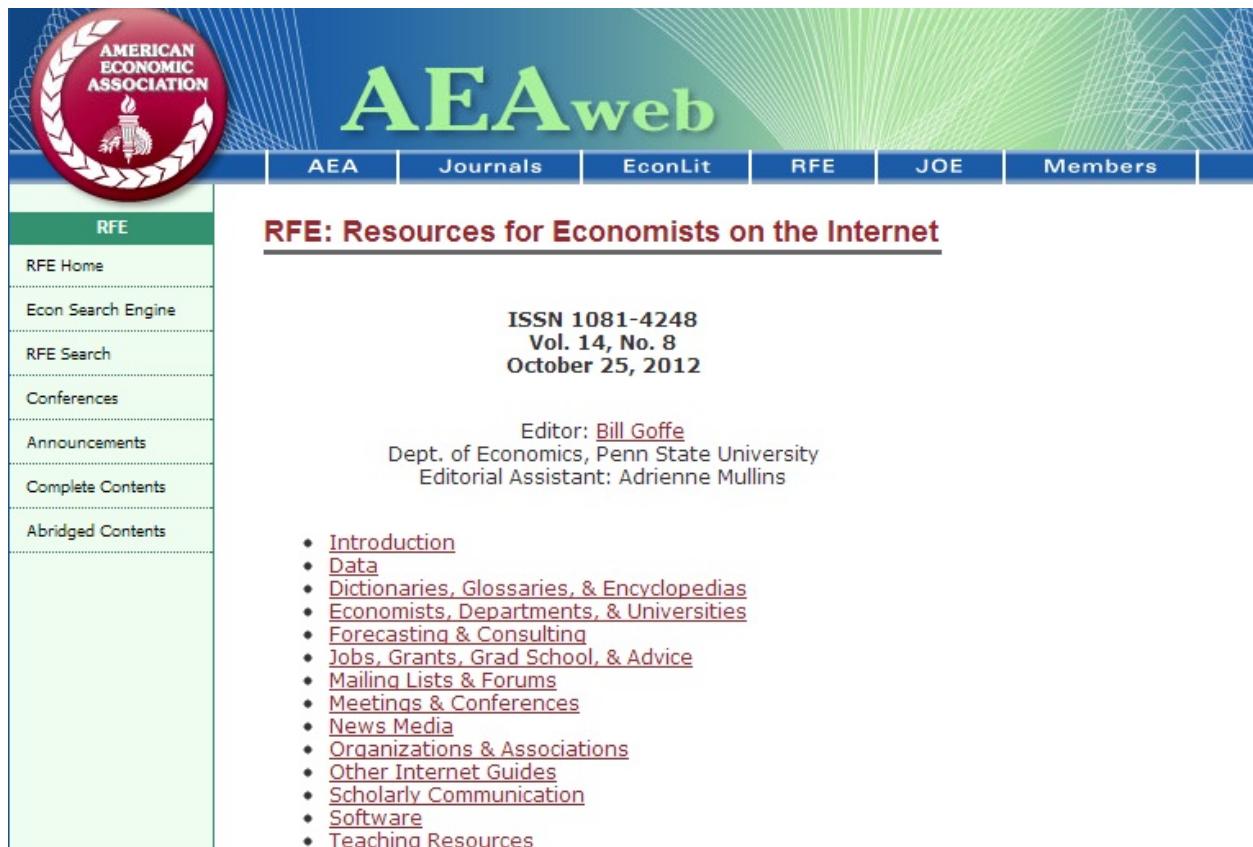


Figure 1.1: Resources for Economists Web Page

1.4 Software

Econometric software tools are widely available. One of the best high-level environments is **Eviews**, a modern object-oriented environment with extensive time series, modeling and forecasting capabilities. (See Figure 1.2.) It implements almost all of the methods described in this book, and many more. Eviews reflects a balance of generality and specialization that makes it ideal for the sorts of tasks that will concern us, and most of the examples in this book are done using it. If you feel more comfortable with another package, however, that's fine – none of our discussion is wed to Eviews in any way, and most of our techniques can be implemented in a variety of packages.

Eviews has particular strength in **time series** environments. **Stata** is



Figure 1.2: Eviews Homepage

a similarly good packaged with particular strength in **cross sections** and **panels**. (See Figure 1.3.)

Eviews and Stata are examples of very high-level modeling environments. If you go on to more advanced econometrics, you'll probably want also to have available slightly lower-level ("mid-level") environments in which you can quickly program, evaluate and apply new tools and techniques. **R** is one very powerful and popular such environment, with special strengths in modern statistical methods and graphical data analysis. (See Figure 1.4.) In this author's humble opinion, R is the key mid-level econometric environment for the foreseeable future. Other notable such environments include **Python** and the rapidly emerging **Julia**.



Figure 1.3: Stata 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 references appear in green.
- Hyperlinks to external items (web pages, video, etc.) appear in cyan.
- Key concepts appear in bold, and they also appear in the (hyperlinked) index.
- Many figures are clickable to reach related material, as are, for example, all figures in this chapter.
- Most chapters contain at least one extensive empirical example in the “Econometrics in Action” section.

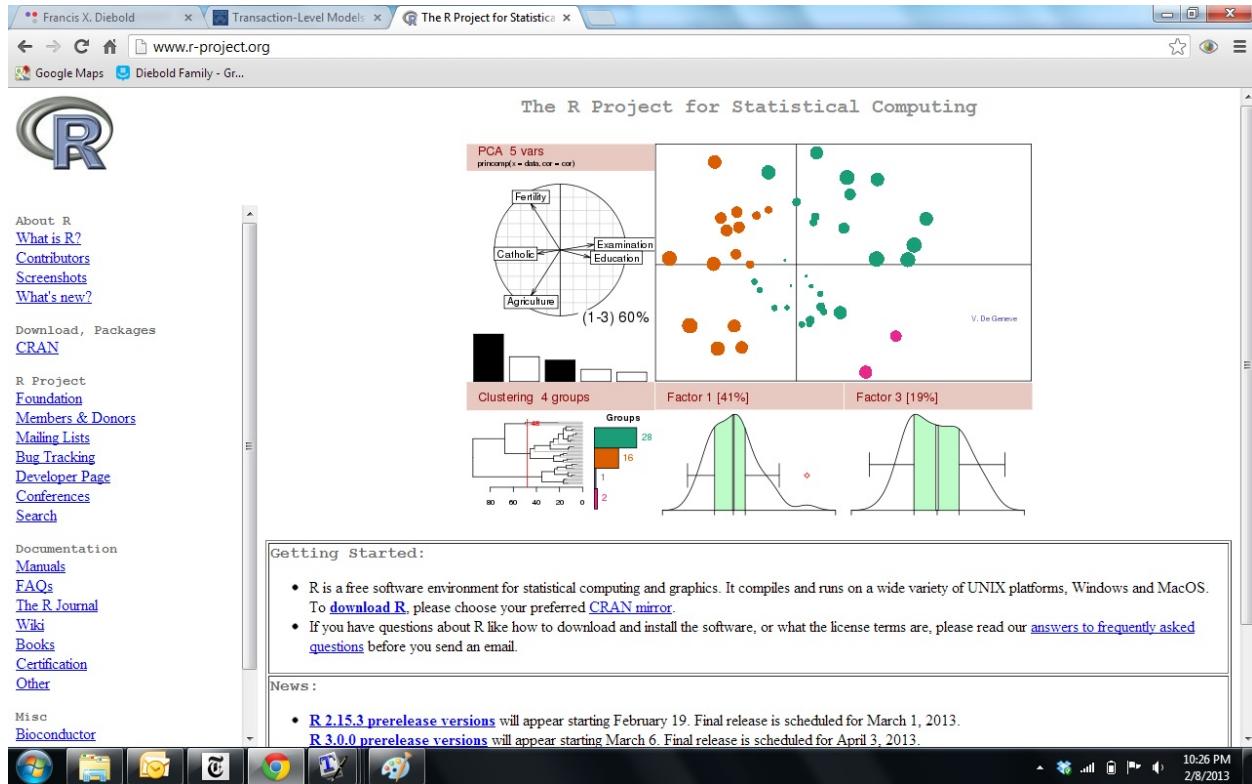


Figure 1.4: R Homepage

- 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 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 “Critique the modeling in this chapter’s Econometrics in Action section, obtain the relevant data, 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

For a compendium of econometric and statistical software, see the [software links](#) site, maintained by Marius Ooms at the *Econometrics Journal*.

R is available for free as part of a massive 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 blog with all sorts of information about all things R. Finally, Quandl has a nice [R interface](#).

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.

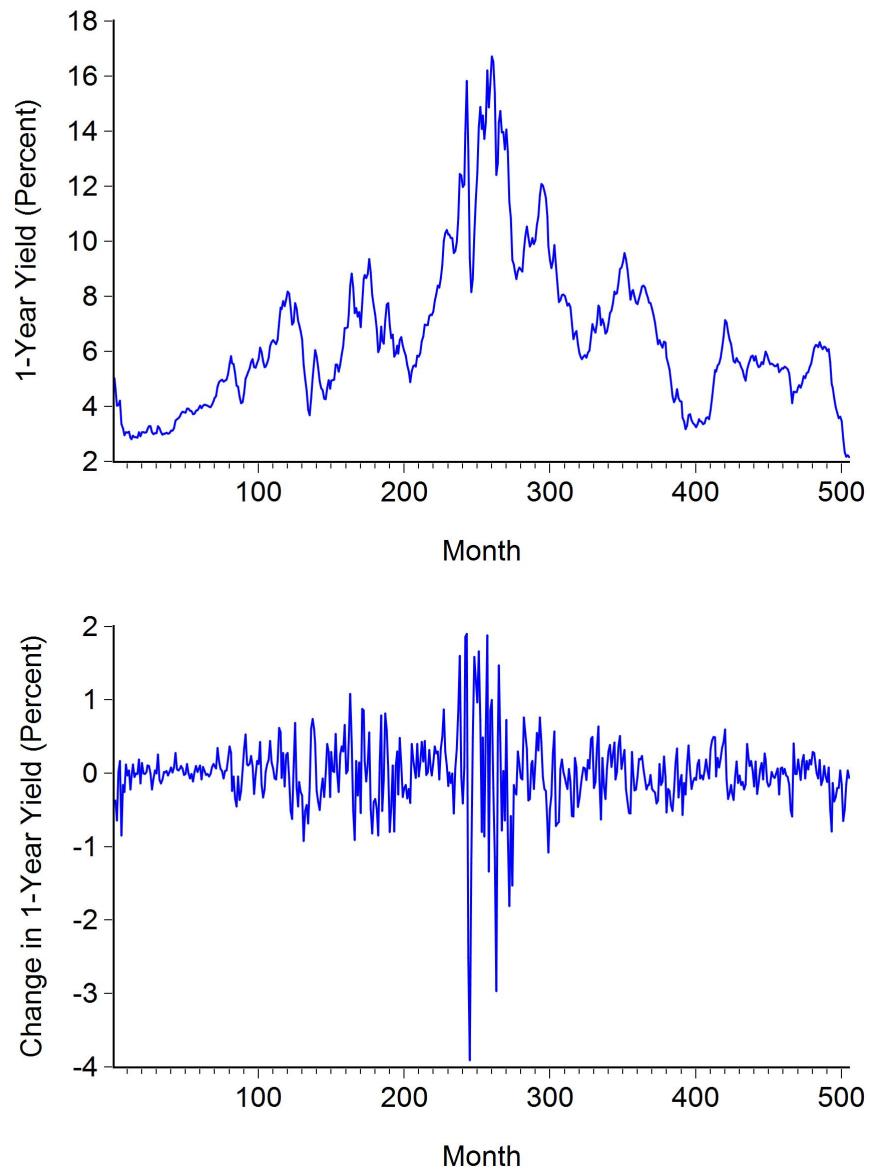


Figure 2.1: 1-Year Goverment Bond Yield, Levels and Changes

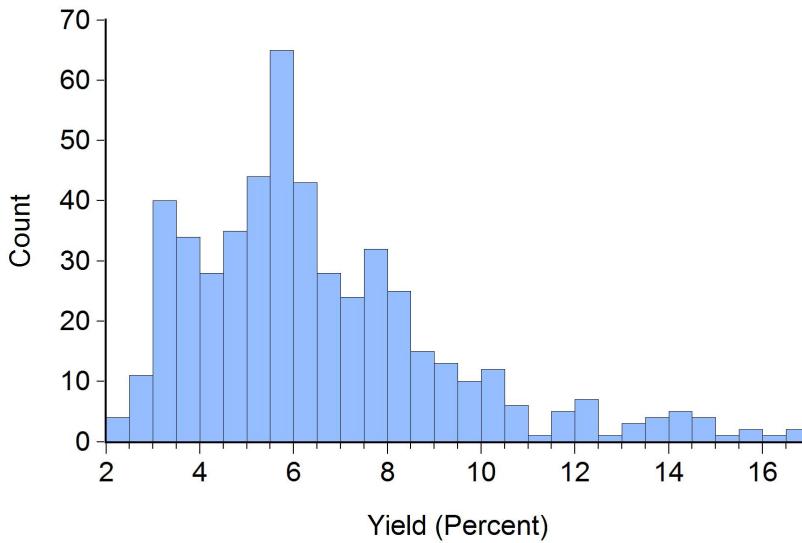


Figure 2.2: Histogram of 1-Year Government Bond Yield

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.

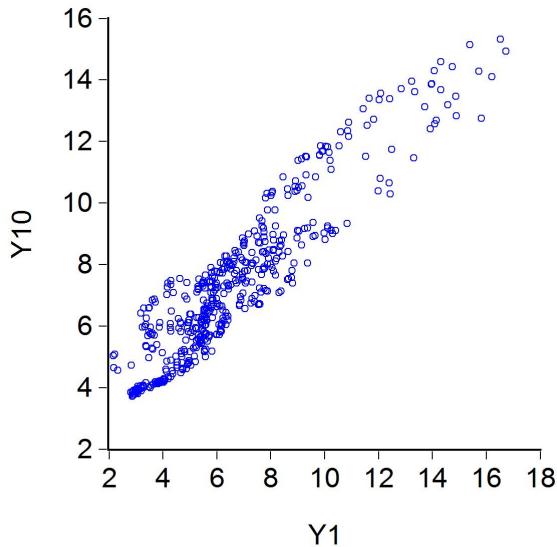


Figure 2.3: Bivariate Scatterplot, 1-Year and 10-Year Government Bond Yields

$\frac{N^2-N}{2}$ 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.

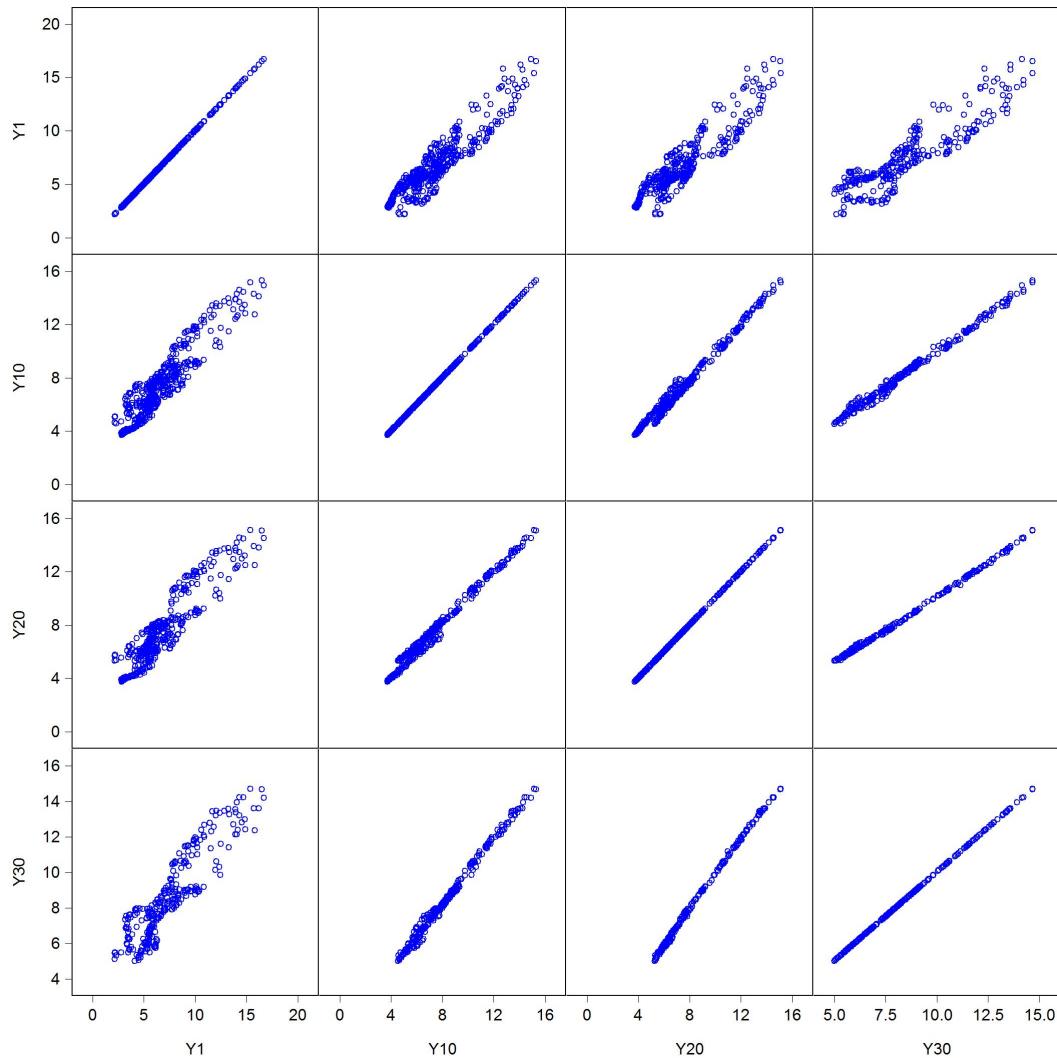


Figure 2.4: Scatterplot Matrix, 1-, 10-, 20- and 30-Year Government Bond Yields

- 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. We will study aspects of big-data econometrics in Chapter ??.

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 shrunk 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.

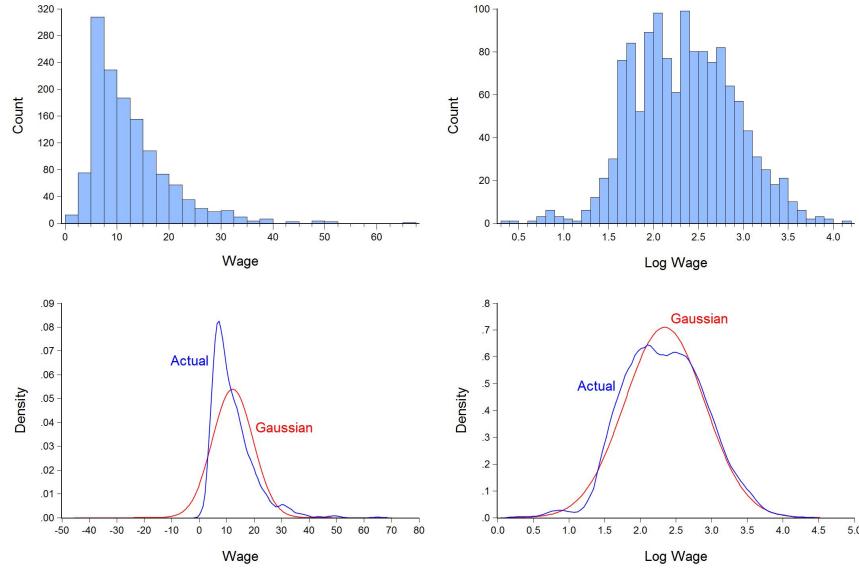


Figure 2.5: Distributions of Wages and Log Wages

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

2.3 U.S. Hourly Wages

We use CPS hourly wage data; for a detailed description see Appendix B.

wage histogram – skewed

wage kernel density estimate with normal superimposed – skewed

log wage histogram – more symmetric

log wage kernel density estimate with normal superimposed – not too bad a fit

2.4 Concluding Remarks

Ultimately good graphics proceeds just like good writing, and if good writing is good thinking, then so too is good graphics good thinking. And good writing *is* just good thinking. So the next time you hear someone pronounce

ignorantly along the lines of “I don’t like to write; I like to think,” rest assured, both his writing and his thinking are likely poor. Indeed many of the classic prose style references contain many insights that can be adapted to improve graphics (even if Strunk and White would view as worthless filler my use of “indeed” earlier in this sentence (“non-thought ink?”)).

So when doing graphics, just as when writing, *think*. Then revise and edit, revise and edit, ...

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. 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)

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.

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

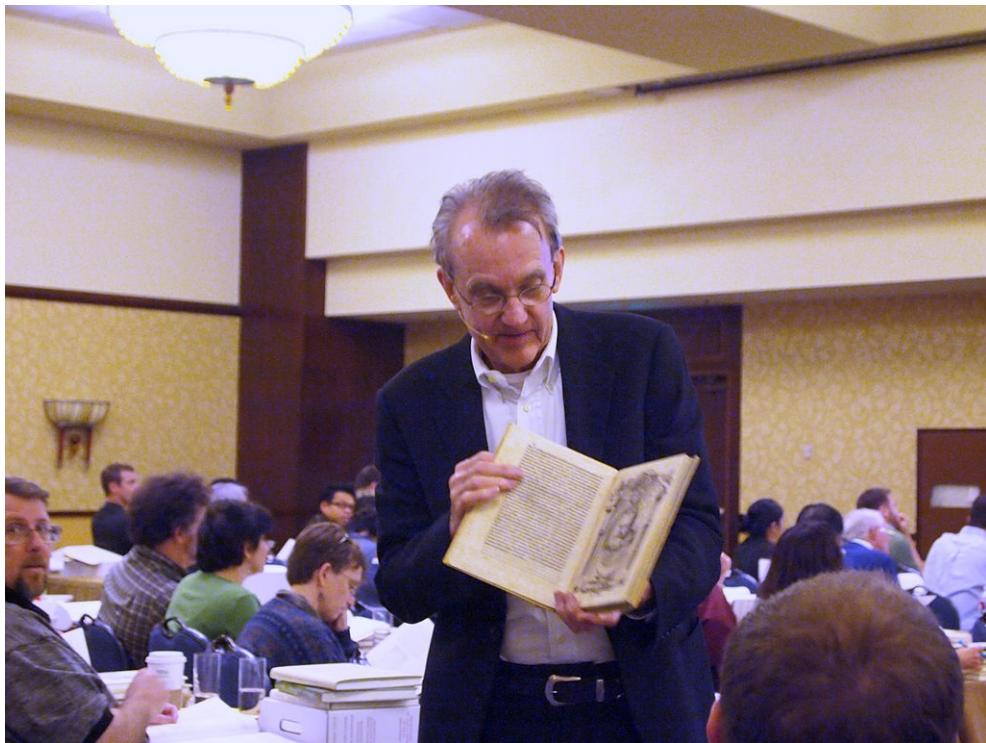


Figure 2.6: Tufte Teaching, with a First Edition Book by Galileo

This chapter has been heavily influenced by [Tufte \(1983\)](#), as are all modern discussions of statistical graphics.⁵ 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.

⁵Photo 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.

Chapter 3

Regression Analysis

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.

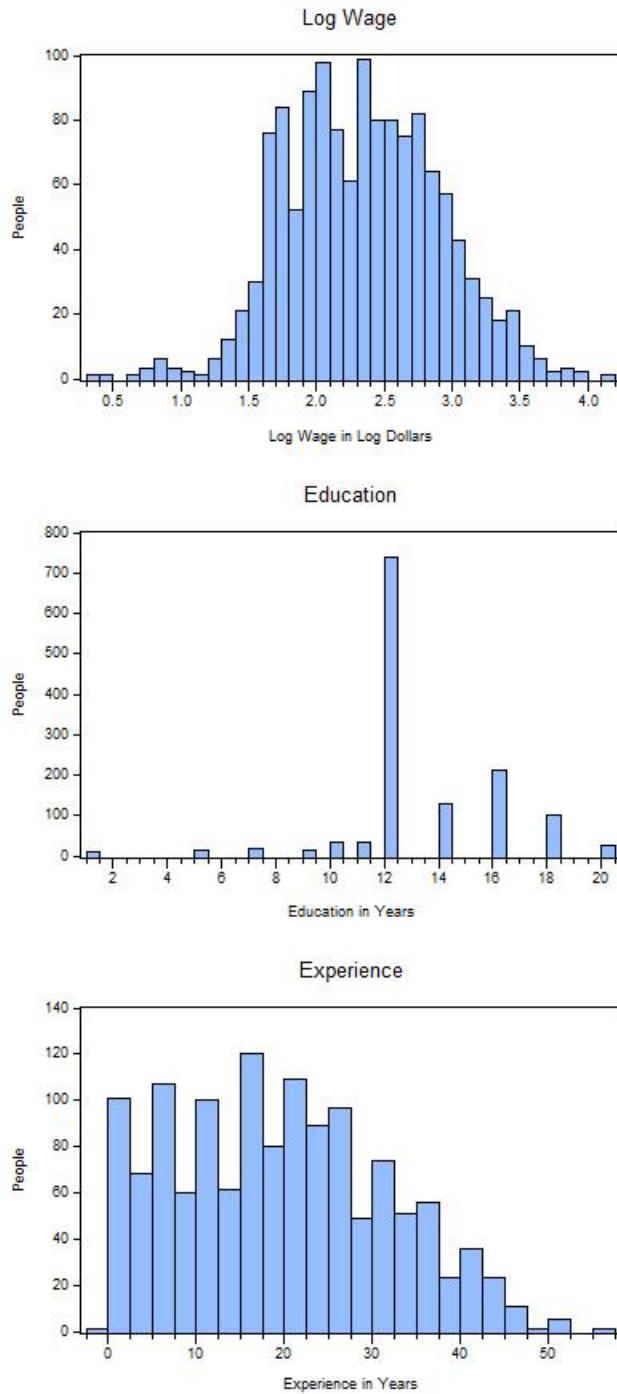


Figure 3.1: Distributions of Log Wage, Education and Experience

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

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

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

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

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

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

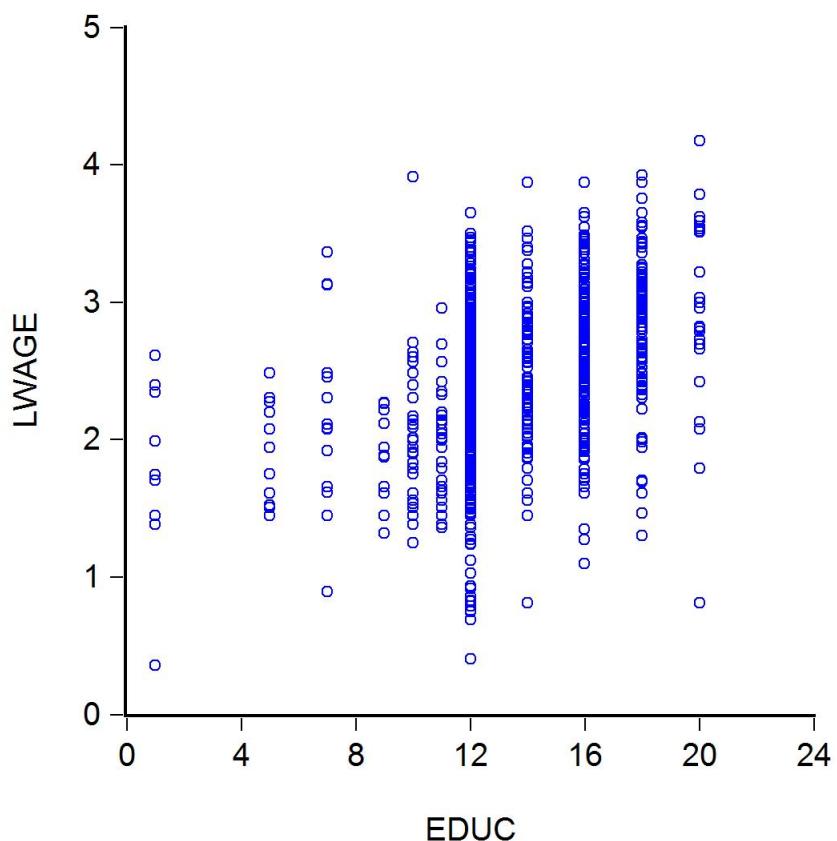


Figure 3.2: (Log Wage, Education) Scatterplot

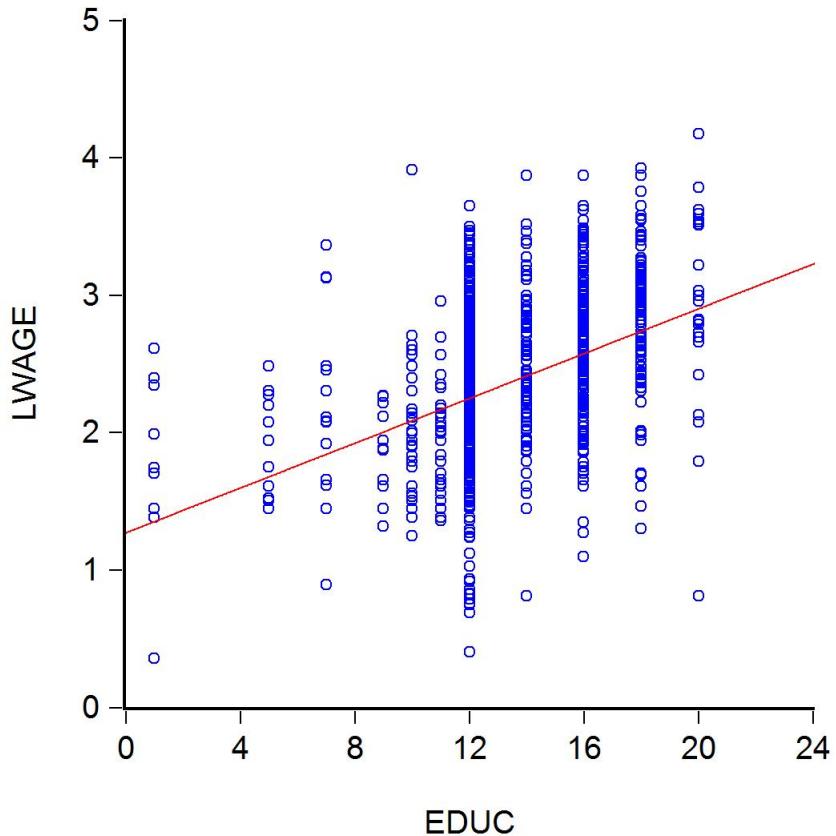


Figure 3.3: (Log Wage, Education) Scatterplot with Superimposed Regression Line

$$t = 1, \dots, T.$$

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

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

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

$$\min_{\beta} \sum_{t=1}^T (y_t - \beta_1 - \beta_2 x_{2t} - \beta_3 x_{3t})^2,$$

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

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

and the residuals are

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

$$t = 1, \dots, T.$$

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

$$\hat{y}_t = \hat{\beta}_1 + \hat{\beta}_2 x_{2t} + \hat{\beta}_3 x_{3t} + \dots + \hat{\beta}_K x_{Kt},$$

which we sometimes write more compactly as

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

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

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 $x'_t = (1, x_{1t}, x_{2t}, \dots, x_{Kt})$.

3.3.1 A Population Model and a Sample Estimator

Thus far we have *not* postulated a probabilistic model that relates y_t and x_t ; instead, we simply ran a mechanical regression of y_t on x_t to find the best fit to y_t formed as a linear function of x_t . It's easy, however, to construct

a probabilistic framework that lets us make statistical assessments about the properties of the fitted line. We assume that y_t is linearly related to an exogenously-determined x_t , and we add an independent and identically distributed zero-mean (iid) Gaussian **disturbance**:

$$y_t = \beta_1 + \beta_2 x_{2t} + \dots + \beta_K x_{Kt} + \varepsilon_t$$

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

$t = 1, \dots, T$. The intercept of the line is β_1 , the slope parameters are the β_i 's, and the variance of the disturbance is σ^2 .² Collectively, we call the β 's the model's **parameters**. The index t keeps track of time; the data sample begins at some time we've called "1" and ends at some time we've called " T ", so we write $t = 1, \dots, T$. (Or, in cross sections, we index cross-section units by i and write $i = 1, \dots, N$.)

Note that in the linear regression model the expected value of y_t conditional upon x_t taking a particular value, say x_t^* , is

$$E(y_t | x_t = x_t^*) = \beta_1 + \beta_2 x_{2t}^* + \dots + \beta_K x_{Kt}^*.$$

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

We assume that the the linear model sketched is true. It is the **population model** in population. 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 $\min_{\beta} \sum_{t=1}^T \varepsilon_t^2$.

3.3.2 Notation, Assumptions and Results

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

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

results 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, $i = 1, \dots, N$, or over time in the time-series case, $t = 1, \dots, T$. With no loss of generality, suppose that we're in a time-series situation; then notationally X is a $T \times K$ matrix.

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

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

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

Alternatively, stack $y_t, t = 1, \dots, T$ into the vector y , where $y' = (y_1, y_2, \dots, y_T)$, and stack $\beta_j, j = 1, \dots, K$ into the vector β , where $\beta' = (\beta_1, \beta_2, \dots, \beta_K)$, and stack $\varepsilon_t, t = 1, \dots, T$, into the vector ε , where $\varepsilon' = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T)$.

Then we can write the complete model over all observations as

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

Our requirement that

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

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 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 Full Ideal Conditions (FIC)

1. The **data-generating process (DGP)** is (3.1)-(3.2), and the fitted model matches the DGP exactly.
2. There is no redundancy among the variables contained in X , so that $X'X$ is non-singular.
3. X is a non-stochastic matrix, fixed in repeated samples.

Note that the first condition above has many important conditions embedded. First, as regards the DGP:

1. Linear relationship, $X\beta$
2. Fixed coefficients, β
3. $\varepsilon \sim N$
4. ε has constant variance σ^2
5. The ε 's are uncorrelated.

Second, as regards the fitted model:

1. No omitted variables
2. No measurement error in observed data.

It is crucial to appreciate that these assumptions are surely heroic – indeed preposterous! – in many econometric contexts, and we shall relax them all in turn. But first we need to understand what happens under the ideal conditions.

For completeness, let us combine everything and write:

The Full Ideal Conditions (FIC):

1. The DGP is:

$$y = X\beta + \varepsilon$$

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

- (a) Linear relationship, $X\beta$.
- (b) Fixed coefficients, β .
- (c) $\varepsilon \sim N$.
- (d) ε has constant variance σ^2 .
- (e) The ε 's are uncorrelated.
- (f) There is no redundancy among the variables contained in X , so that $X'X$ is non-singular.
- (g) X is a non-stochastic matrix, fixed in repeated samples.

2. The fitted model matches the DGP exactly:

- (a) No omitted variables
- (b) No measurement error

Results

The least squares estimator is

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

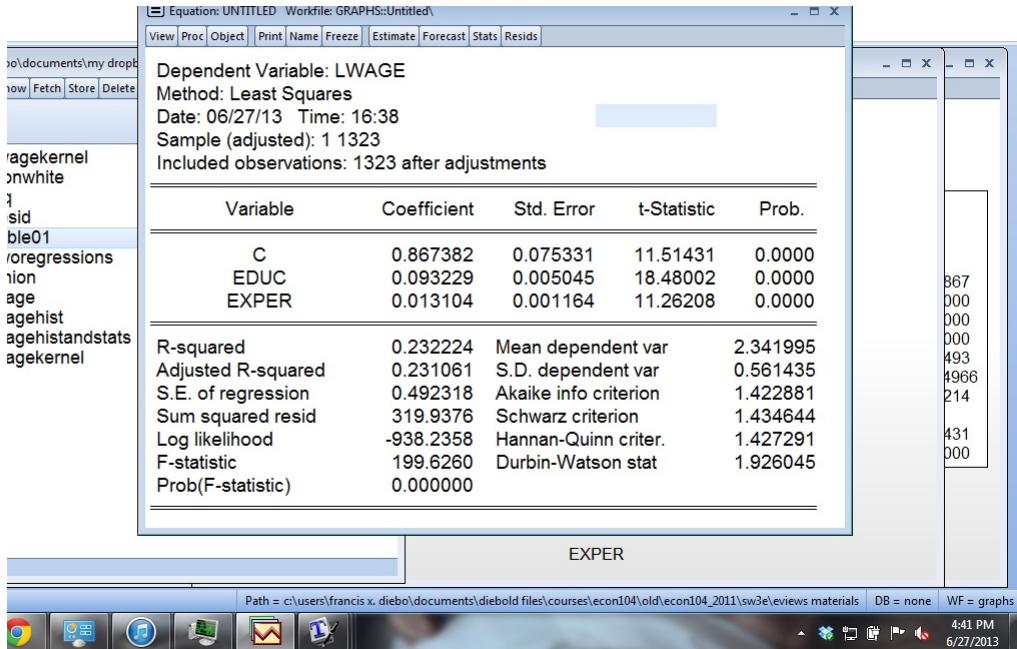


Figure 3.4: Regression Output

and under the full ideal conditions it is consistent, normally distributed with covariance matrix $\sigma^2(X'X)^{-1}$, and indeed MVUE. We write

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

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 4.2 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 full ideal conditions are surely *not* 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 printout begins by reminding us that we’re running a least-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.³

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 forecasting 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%

³Sometimes the population coefficient on C is called the **constant term**, and the regression estimate is called the estimated constant term.

confidence interval for the parameter. If so, we reject irrelevance. The t statistic is just the ratio of the estimated coefficient to its standard error, so if zero is outside the 95% confidence interval, then the t statistic must be bigger than two in absolute value. Thus we can quickly test irrelevance at the 5% level by checking whether the t statistic is greater than two in absolute value.⁴

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 is 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.

⁴If 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.

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-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

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

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

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

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

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

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.⁵ Like the sum of squared residuals, it's not of direct use, but it's useful for comparing models and testing hypotheses. We will rarely use the log likelihood function directly; instead, we'll focus for the most part on the sum of squared residuals.

⁵Throughout this book, “log” refers to a natural (base e) logarithm.

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.⁶ That is, we test whether, taken jointly as a set, the variables included in the forecasting model have any predictive value. This contrasts with the t statistics, which we use to examine the predictive worth of the variables one at a time.⁷ If no variable has predictive value, the F statistic follows an F distribution with $k - 1$ and $T - k$ degrees of freedom. The formula is

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

where SSR_{res} is the sum of squared residuals from a *restricted* regression that contains only an intercept. Thus the test proceeds by examining how much 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 predictive content.

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 forecasted y_t using $x'_t\beta$, then our forecast errors would be the ε_t 's, with variance σ^2 . We'd like an estimate of σ^2 ,

⁶We 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.

⁷In 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 $F \neq t^2$.

because it tells us whether our forecast errors are likely to be large or small. The observed residuals, the e_t 's, are effectively estimates of the unobserved population disturbances, the ε_t 's. Thus the sample variance of the e 's, which we denote s^2 (read “**s-squared**”), is a natural estimator of σ^2 :

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

s^2 is an estimate of the dispersion of the regression disturbance and hence is used to assess goodness of fit of the model, as well as the magnitude of forecast errors that we're likely to make. The larger is s^2 , the worse the model's fit, and the larger the forecast errors we're likely to make. s^2 involves a degrees-of-freedom correction (division by $T - K$ rather than by $T - 1$, reflecting the fact that K regression coefficients have been estimated), which is an attempt to get a good estimate of the out-of-sample forecast 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

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

The standard error of the regression is easier to interpret than s^2 , because its units are the same as those of the e 's, whereas the units of s^2 are not. If the e 's are in dollars, then the squared e 's are in dollars squared, so s^2 is in dollars squared. By taking the square root at the end of it all, *SER* converts the units back to dollars.

Sometimes it's informative to compare the standard error of the regression (or a close relative) to the standard deviation of the dependent variable (or a close relative). The standard error of the regression is an estimate of the standard deviation of forecast errors from the regression model, and the standard deviation of the dependent variable is an estimate of the standard

deviation of the forecast errors from a simpler forecasting model, in which the forecast each period is simply \bar{y} . If the ratio is small, the variables in the model appear very helpful in forecasting y . R -squared measures, to which we now turn, are based on precisely that idea.

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 forecasting y ; hence it is widely used as a quick check of **goodness of fit**, or forecastability of y based on the variables included in the regression. Here the R^2 is about 55% – good but not great. The formula is

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

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

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

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 quite close (53% vs. 55%). Adjusted R^2 is often denoted \bar{R}^2 ; the formula is

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

where K is the number of 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 forecasting models. The formula is:

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

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:

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

The AIC and SIC are tremendously important for guiding model selection in ways that avoid **data mining** and **in-sample overfitting**. In Chapter ?? we discuss in detail the sum of squared residuals, the standard error of the regression, R^2 , adjusted R^2 , the AIC , and the SIC , the relationships among them, and their role in selecting forecasting models.

3.4.12 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.

3.4.13 Durbin-Watson stat. 1.926

The Durbin-Watson statistic is useful in time series environments for assessing whether the ε_t 's are correlated over time; that is, whether the *iid* assumption (part of the full ideal conditions) is violated. It is irrelevant in the present application to wages, which uses cross-section data. We nevertheless introduce it briefly here.

The **Durbin-Watson statistic** tests for correlation over time, called **serial correlation**, in regression disturbances. It works within the context of a regression model with disturbances

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

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

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, but when $\phi \neq 0$, the disturbance is serially correlated. More specifically, when $\phi \neq 0$, we say that ε_t follows an autoregressive process of order one, or *AR(1)* for short.⁸ If $\phi > 0$ the disturbance is positively serially correlated, and if $\phi < 0$ the disturbance is negatively serially correlated. **Positive serial correlation** is typically the relevant alternative in the applications that will concern us. The formula for the Durbin-Watson (*DW*) statistic is

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

⁸Although the Durbin-Watson test is designed to be very good at detecting serial correlation of the *AR(1)* type. Many other types of serial correlation are possible; we'll discuss them extensively in Chapter 14.1.

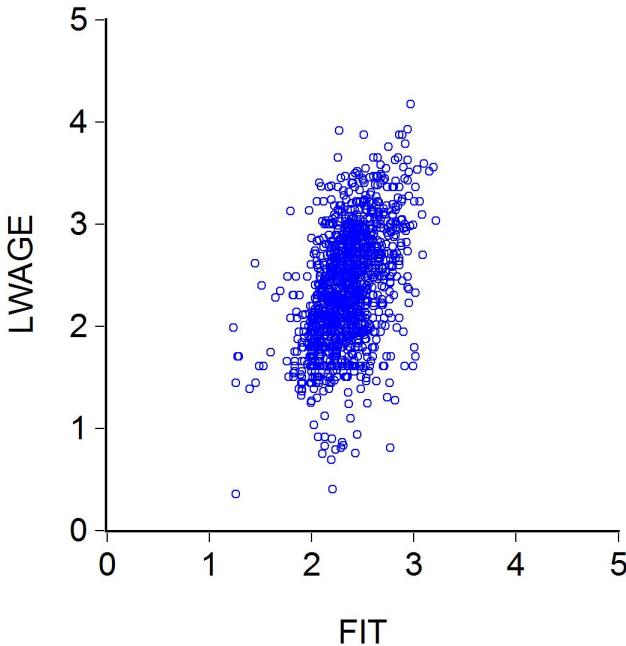


Figure 3.5: Wage Regression Residual Scatter

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.

3.4.14 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.

3.4.15 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_t 's), the fitted values (\hat{y}_t 's), and the residuals (e_t 's). Often we'll refer to such plots, shown together in a single graph, as a **residual plot**.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_t '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_t 's are in fact substantially smaller and less variable than either the y_t 's or the \hat{y}_t 's. We draw the zero line through the residuals for visual comparison. 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.

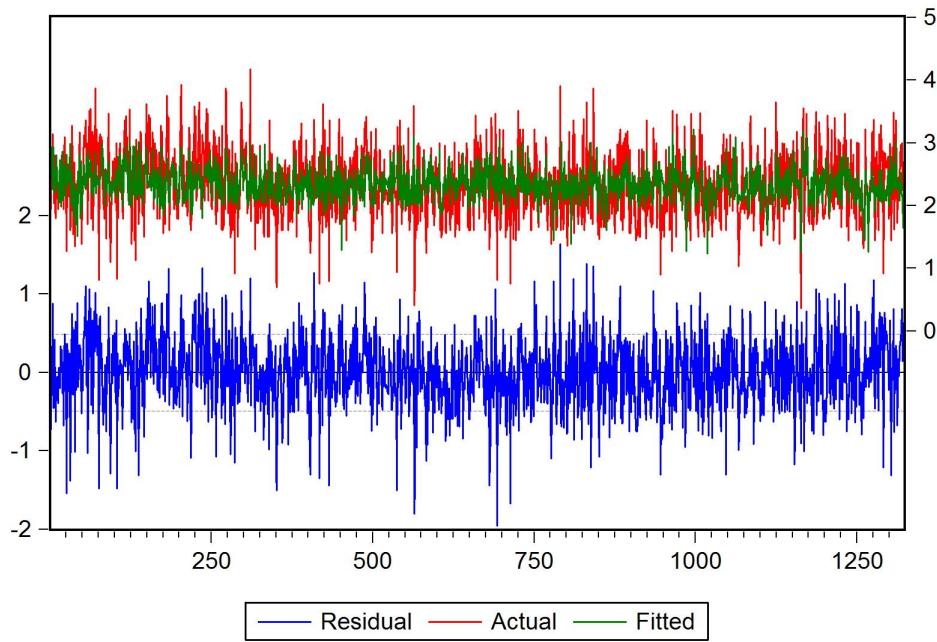


Figure 3.6: Wage Regression Residual Plot

3.5 Quantile Regression

Beyond OLS: Non-Quadratic Objectives

Ordinary Least Squares (OLS)

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

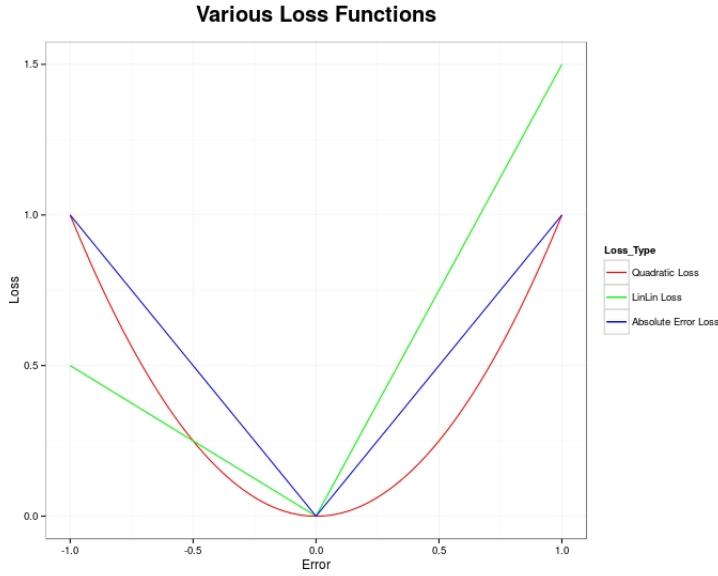
$$\min_{\beta} \sum_{t=1}^T (y_t - \beta_1 - \beta_2 x_{2t} - \dots - \beta_K x_{Kt})^2 = \min_{\beta} \sum_{t=1}^T \varepsilon_t^2$$

- Simple
(analytic closed-form expression, $(X'X)^{-1}X'y$)

- Wonderful properties under FIC
(Unbiased, consistent, Gaussian, MVUE)

But other approaches are possible and sometimes useful.

Least Absolute Deviations (LAD)



The LAD estimator, $\hat{\beta}_{LAD}$, solves:

$$\min_{\beta} \sum_{t=1}^T |\varepsilon_t|$$

- Not as simple as OLS, but still simple
(Solves a linear programming problem)
- Useful properties under some violations of FIC
(Robust to outliers; more on that later)
- But there's a much bigger reason to be interested
Conditional Mean and Median Functions
- 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 two are equal under symmetry as with FIC, but not under asymmetry, in which case the median is a better measure of central tendency
-

Quantile Regression (QR)

Objective like LAD but unequal slopes on each side of 0.

QR estimator $\hat{\beta}_{QR}$ minimizes “linlin loss,” or “check function loss”:

$$\min_{\beta} \sum_{t=1}^T \text{linlin}(\varepsilon_t),$$

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.

Not as simple as OLS, but still simple

(solves a linear programming problem)

What Does Quantile Regression Fit?

- QR fits the $d \cdot 100\%$ quantile:

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

where

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

- Median regression (LAD) is special case of $d = .5$

– Important generalization of median 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?)

3.6 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 $y_t = \beta_1 + \beta_2 x_t + \beta_3 z_t + \varepsilon_t$, where y_t is the number of hot dogs sold at an amusement park on a given day, x_t is the number of admission tickets sold that day, z_t is the daily maximum temperature, and ε_t is a random error.
 - a. State whether each of y_t , x_t , z_t , 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. Durbin-Watson statistic
 - k. Mean of the dependent variable
 - l. Standard deviation of the dependent variable
 - m. Akaike information criterion
 - n. Schwarz information criterion
 - o. F statistic
 - p. 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 full ideal conditions?

- c. Does the variance-covariance matrix of the OLS estimator under the full ideal conditions depend on any unknown parameters, and if so, how would you estimate them?
- d. What is your estimate of the standard error of the sample mean?

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 full 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 FIC might fail.) Which of the FIC might fail? One? A few? All? Why? Insofar as possible, discuss the FIC, 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. Quantile Regression.

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.

10. (Understanding model selection criteria)

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?

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

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

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

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

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.

12. The sample mean and the OLS estimator.

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 estimator under the full ideal conditions.

- (a) What *are* the full ideal conditions?
- (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.
- (d) How would you estimate and plot the exact distribution of the OLS estimator?

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

13. 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 , S^2 , and SIC . What role does SSR play in each of the test statistics?
- (d) Under the full ideal conditions, is the maximized log likelihood related to the SSR ? If so, how, and what would happen to the relationship if we dropped normality from the full ideal conditions?

3.7 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.

3.8 Regression's Inventor: Carl Friedrich Gauss



Figure 3.7: Carl Friedrich Gauss

This is a photographic reproduction of public domain work of art, 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). Photo by A. Wittmann.

Chapter 4

Indicator Variables in Cross Sections

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.

4.1 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,

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

In Figure 4.1 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); we have examined them before and there is nothing new to say. The new vari-

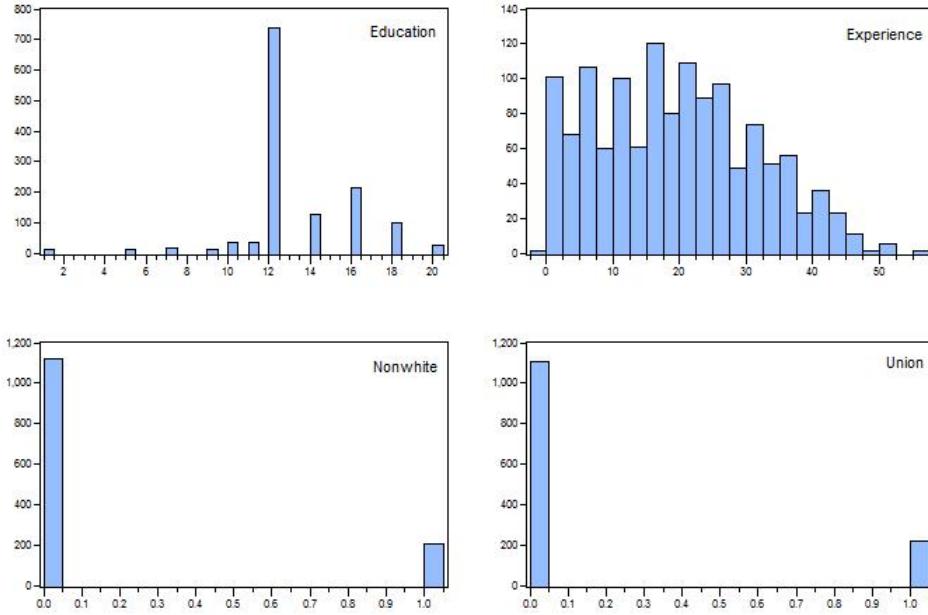


Figure 4.1: Histograms for Wage Covariates

ables are 0-1 dummies, UNION (already defined) and NONWHITE, where

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

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

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

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.¹ If an intercept is included, one of the dummy categories must be dropped.

4.2 Group Dummies in the Wage Regression

Recall our basic wage regression,

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

shown in Figure 4.2. Both explanatory variables highly significant, with expected signs.

Now consider the same regression, but with our three group dummies added, as shown in Figure 4.3. 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 4.4. Of course it’s hardly the forty-five degree line (the regression R^2 is higher but still only .31), but it’s

¹We’ll examine such issues in detail later when we study “multicollinearity” in Chapter ??.

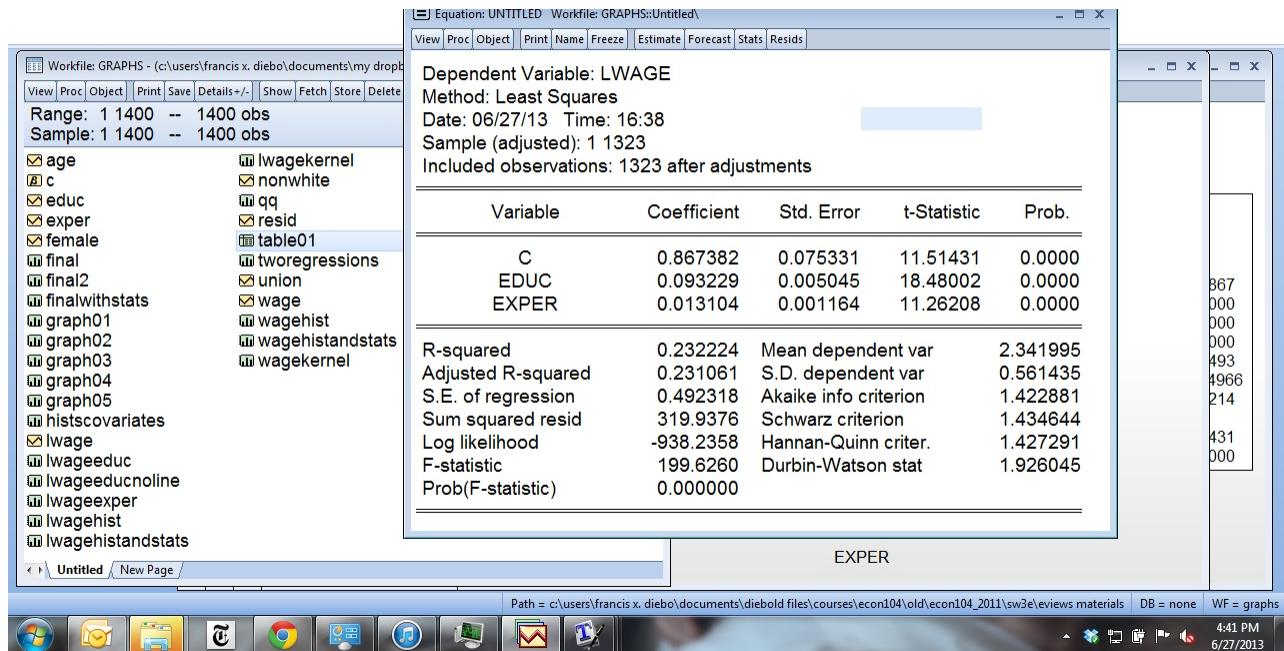


Figure 4.2: Wage Regression on Education and Experience

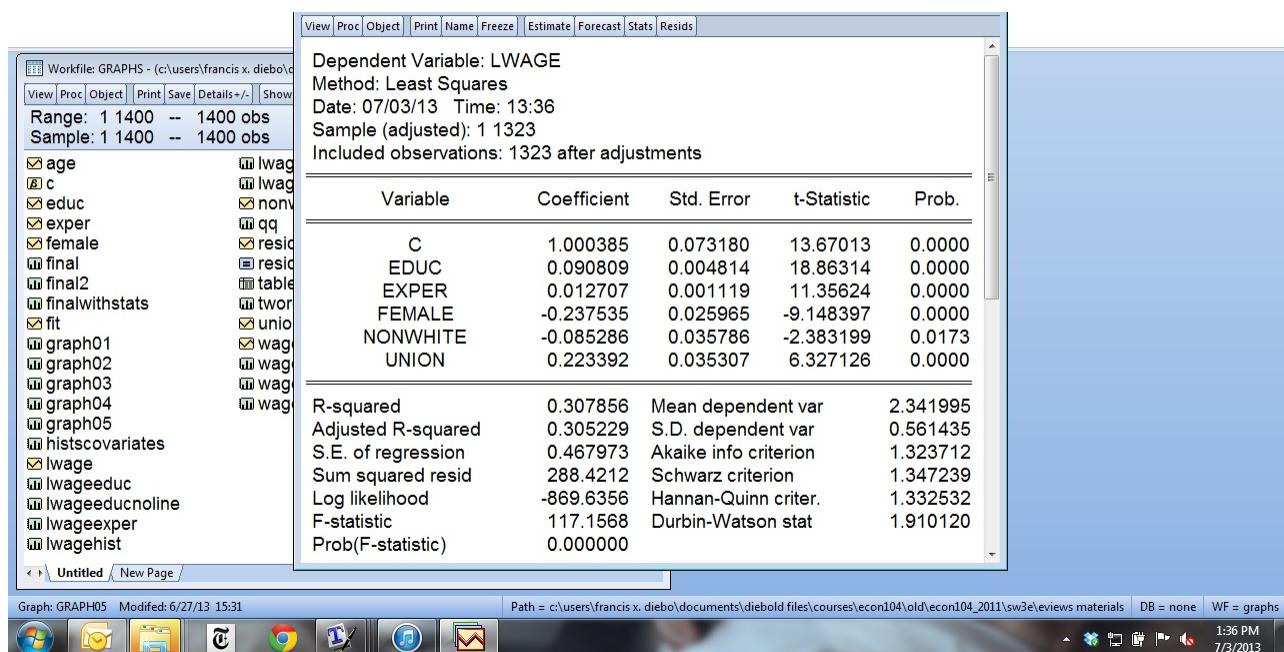


Figure 4.3: Wage Regression on Education, Experience and Group Dummies

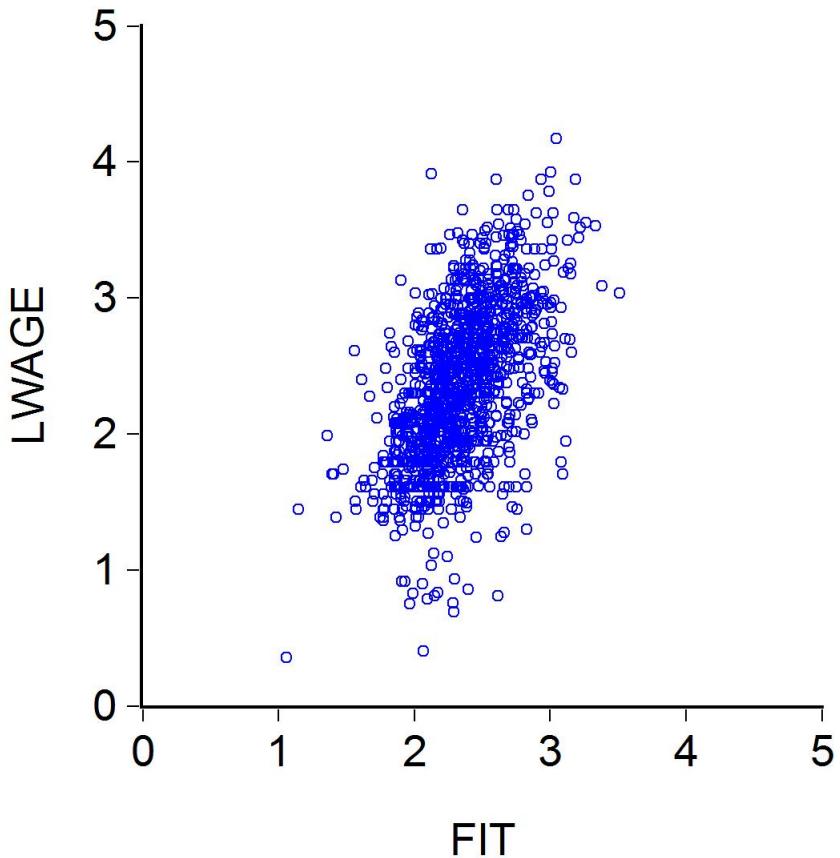


Figure 4.4: Residual Scatter from Wage Regression on Education, Experience and Group Dummies

getting closer.

4.3 Exercises, Problems and Complements

1. (Slope dummies)

Consider the regression

$$y_t = \beta_1 + \beta_2 x_t + \varepsilon_t.$$

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 $\beta_1 + \delta D_t$.

We can also allow slope coefficients to vary with groups. Instead of writing the slope as β_2 , we write it as $\beta_2 + \gamma D_t$. Hence to capture slope variation across groups we regress 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, $y_t \rightarrow c, x_t$.

- (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. (The wage equation with intercept and slope dummies)

Try allowing for slope dummies in addition to intercept dummies in our wage regression. Discuss your results.

4. (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 randomly treat each of 1000 randomly-selected U.S. farms that use fertilizer A, either keeping in place the old fertilizer A used or replacing

it with one of four new experimental fertilizers, B, C, D or E. Using a dummy variable regression setup:

- (a) How would you test the hypothesis that none of the four new fertilizers is better or worse than the old fertilizer?
- (b) Assuming that you reject the null, how would you estimate the improvement (or worsening) due to replacing fertilizer A with fertilizer D?

4.4 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*.

4.5 Dummy Variables, ANOVA, and Sir Ronald Fischer



Figure 4.5: Sir Ronald Fischer

Fischer is in many ways the “father” of much of modern statistics.

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 5

Indicator Variables in Time Series

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**.

5.1 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.¹

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 dummy**.” $TIME$ equals 1 in the first period of the sample, 2 in the second period, and so on. Thus, for a sample of size T , $TIME = (1, 2, 3, \dots, T - 1, T)$. Put differently, $TIME_t = t$, so that the $TIME$ variable

¹Later we'll broaden our discussion to allow for **stochastic trend**.

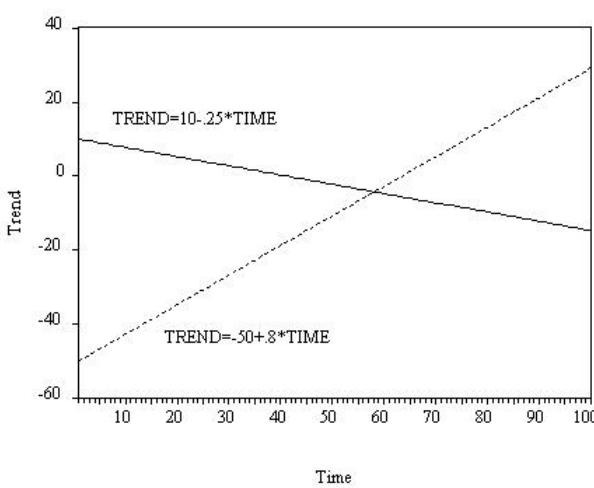


Figure 5.1: Various Linear Trends

simply indicates the time. β_1 is the **intercept**; it's the value of the trend at time $t=0$. β_2 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 5.1, 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 such need not be the case. In recent decades, for example, male labor force participation rates have been falling, as have the times between trades on stock exchanges. In other 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 $y \rightarrow c, TIME$.

5.2 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.² The annual repetition can be exact, in which case we speak of **deterministic seasonality**, or approximate, in which case we speak of **stochastic 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.

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.

²Note therefore that seasonality is impossible, and therefore not an issue, in data recorded once per year, or less often than once per year.

5.2.1 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, $s = 52$ if we have weekly data, and so forth.

The pure seasonal dummy model is

$$\text{Seasonal}_t = \sum_{i=1}^s \gamma_i \text{SEAS}_{it}$$

$$\text{where } \text{SEAS}_{it} = \begin{cases} 1 & \text{if observation } t \text{ falls in season } i \\ 0 & \text{otherwise} \end{cases}$$

The SEAS_{it} 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 variables³:

$$\text{SEAS}_1 = (1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, \dots, 0)'$$

$$\text{SEAS}_2 = (0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, \dots, 0)'$$

$$\text{SEAS}_3 = (0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, \dots, 0)'$$

$$\text{SEAS}_4 = (0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, \dots, 1)'.$$

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

³For illustrative purposes, assume that the data sample begins in Q1 and ends in Q4.

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_i '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.

Trend may be included as well. For example, we can account for seasonality and linear trend by running⁴

$$y \rightarrow TIME, SEAS_1, \dots, SEAS_s.$$

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.

⁴Note well that we drop the intercept! (Why?)

5.2.2 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 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.

5.3 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 5.2, which displays clear trend (sales are increasing) and seasonality (sales skyrocket during the Christ-

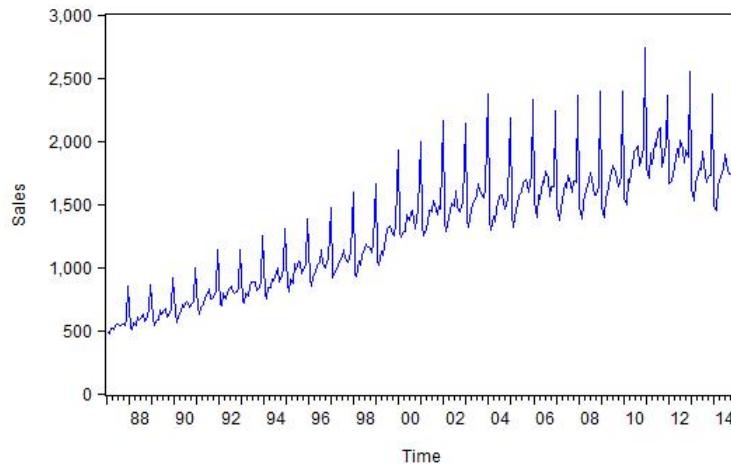


Figure 5.2: Liquor Sales

mas season, among other things).

We show log liquor sales in Figure 5.3 ; we take logs to stabilize the variance, which grows over time.⁵ Log liquor sales has a more stable variance, and it's the series for which we'll build models.⁶

Linear trend estimation results appear in Table 5.4. The trend is increasing and highly significant. The adjusted R^2 is 84%, reflecting the fact that trend is responsible for a large part of the variation in liquor sales.

The residual plot (Figure 5.5) 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. The Durbin-Watson statistic missed it, evidently because it's not designed to have power against seasonal dynamics.⁷

In Figure 5.6 we show estimation results for a model with linear trend

⁵The 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.

⁶From this point onward, for brevity we'll simply refer to “liquor sales,” but remember that we've taken logs.

⁷Recall that the Durbin-Watson test is designed to detect simple $AR(1)$ dynamics. It also has the ability to detect other sorts of dynamics, but evidently not those relevant to the present application, which are very different from a simple $AR(1)$.

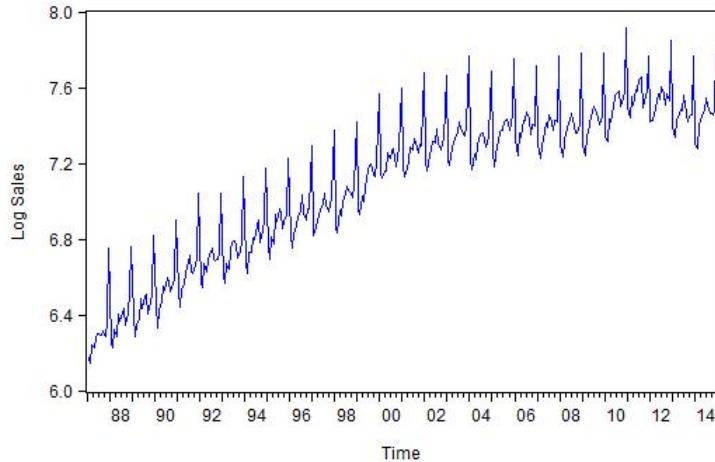


Figure 5.3: Log Liquor Sales

and seasonal dummies. (Note that we dropped the intercept!) The seasonal dummies are highly significant, and in many cases significantly different from each other. R^2 is higher.

In Figure 5.7 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.

In Figure 5.8 we plot the estimated seasonal pattern, 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 (nonlinear) trends.

5.4 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.

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 5.4: Linear Trend Estimation

- 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. (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 with moving average methods, with the dummy variable methods discussed in this chapter, or with sophisticated hybrid methods developed at the U.S. Census Bureau and elsewhere.

- a. Discuss in detail how you'd use a linear trend plus seasonal dummies model to seasonally adjust a series.

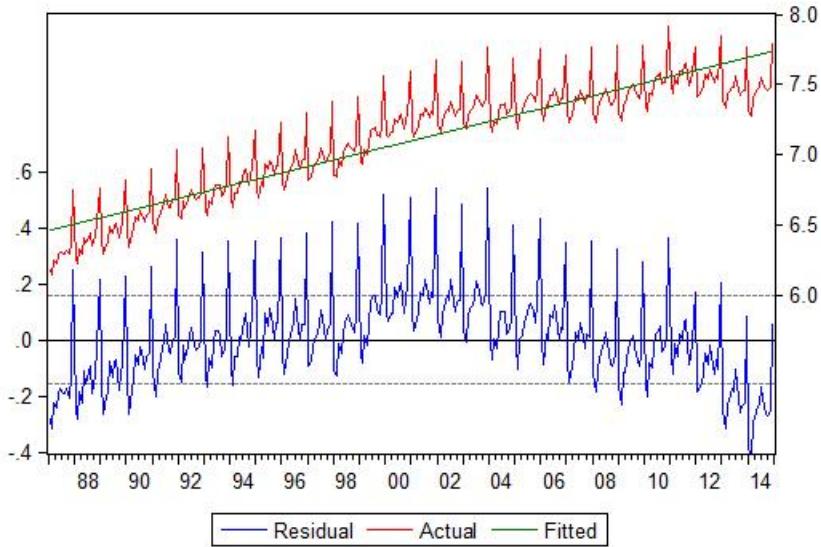


Figure 5.5: Residual Plot, Linear Trend Estimation

- b. 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.
 - c. Search the Web (or the library) for information on the latest U.S. Census Bureau seasonal adjustment procedure, and report what you learned.
3. (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. A sporting goods store finds 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 and March, in the spring months of April, May and June, and so on.
- b. A campus bookstore finds that detrended sales are roughly the same

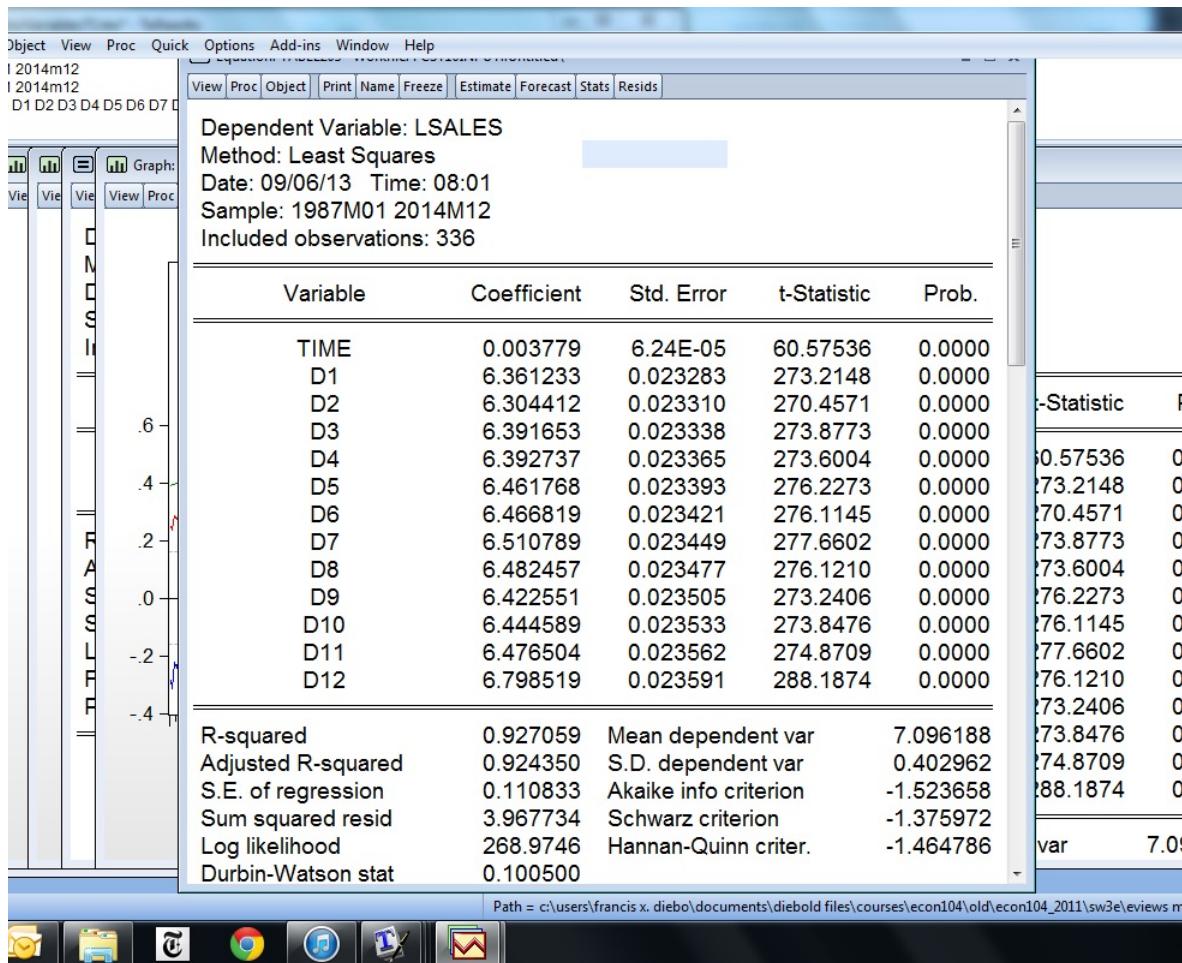


Figure 5.6: Estimation Results, Linear Trend with Seasonal Dummies

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.

- c. A Christmas ornament store is only open in November and December, so sales are zero in all other months.
4. (Testing for seasonality)

Using the log liquor sales data:

- a. As in the chapter, construct and estimate a model with a full set of seasonal dummies.

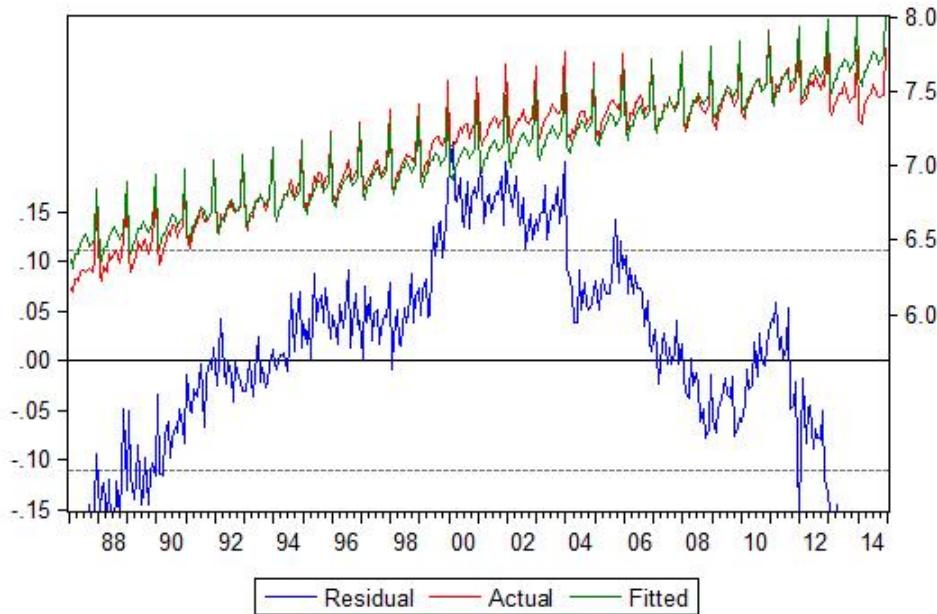


Figure 5.7: Residual Plot, Linear Trend with Seasonal Dummies

- 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.

5.5 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

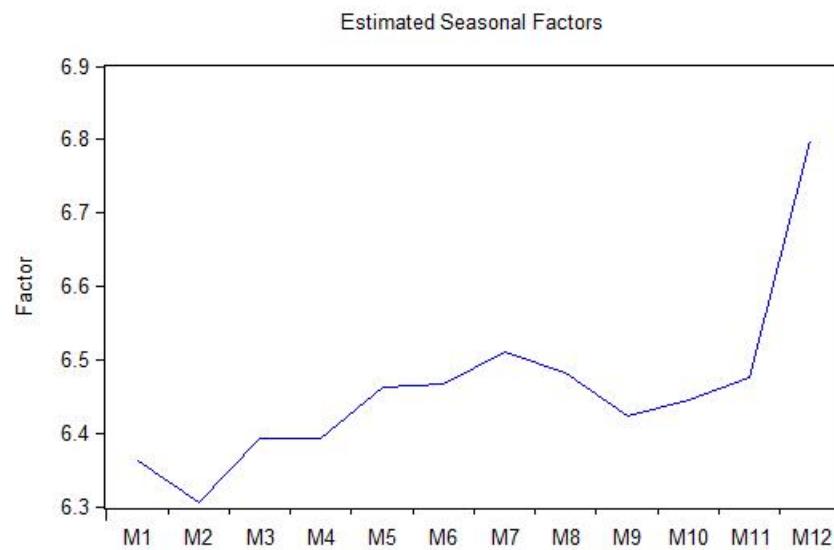


Figure 5.8: Seasonal Pattern

from the Bank of Spain.

Chapter 6

Non-Linearity in Cross Sections

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.

6.1 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.

6.1.1 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 regres-

sion case, but of course we could have more than one regressor. We have

$$\ln y_t = \beta_1 + \beta_2 \ln x_t + \varepsilon_t.$$

y_t is a non-linear function of the x_t , 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,

$$y_t = AL_t^\alpha K_t^\beta \exp(\varepsilon_t).$$

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

$$\ln y_t = \ln A + \alpha \ln L_t + \beta \ln K_t + \varepsilon_t.$$

This transformed model can be immediately estimated by ordinary least squares. We simply regress $\ln y_t$ on an intercept, $\ln L_t$ and $\ln K_t$. 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_t and x_t , $(\ln y_t - \ln x_t)$ is approximately the percent difference between y_t and x_t . Hence the coefficients in log-log regressions give the expected percent change in $E(y_t|x_t)$ for a one-percent change in x_t , the so-called *elasticity of y_t with respect to x_t* .

Log-Lin Regression

Second, consider **log-lin regression**, in which $\ln y_t = \beta x_t + \varepsilon$. We have a log on the left but not on the right. The classic example involves the

workhorse model of exponential growth:

$$y_t = Ae^{rt}$$

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

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

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**:

$$y_t = \beta \ln x_t + \varepsilon.$$

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

6.1.2 Box-Cox and GLM

Box-Cox

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

$$B(y_t) = \beta_1 + \beta_2 x_t + \varepsilon_t,$$

where

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

Hence

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

Because

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

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

$$G(y_t) = \beta_1 + \beta_2 x_t + \varepsilon_t,$$

so that

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

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

6.2 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**,

$$y = \frac{1}{a + br^x},$$

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.

6.2.1 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, however, 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 FIC. But the violation is not a big deal. Under the remaining FIC (that is, dropping only linearity), $\hat{\beta}_{NLS}$ has a sampling distribution similar to that under the FIC.

6.3 Interactions

Suppose that the regression relationship is truly quadratic rather than linear,

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

In the multiple regression case we would also have interaction terms. Consider, for example, $f(x_t, z_t)$:

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

Such **interaction effects** are also relevant in situations involving dummy variables. There we capture interactions by including products of dummies.¹

6.4 A Final Word on Nonlinearity and the FIC

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

Models linear in transformed variables (e.g., log-log regression) actually *don't* violate the FIC, 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 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 FIC) due to the omitted higher-order expansion terms. Hence the failure of the FIC discussed in this chapter can be viewed either as:

1. The linearity assumption ($E(y|X) = X'\beta$) is incorrect, or

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

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.

6.5 Testing for Non-Linearity

6.5.1 t and F Tests

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 .

6.5.2 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 $y_t \rightarrow c, X_t$ and obtain the fitted values \hat{y}_t . Then, to test for non-linearity, we run the regression again with various powers of \hat{y}_t included,

$$y_t \rightarrow c, X_t, \hat{y}_t^2, \dots, \hat{y}_t^m.$$

Note that the powers of \hat{y}_t 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}_t , because that would be redundant with the included X variables. Instead we include powers $\hat{y}_t^2, \hat{y}_t^3, \dots$ Typically a small

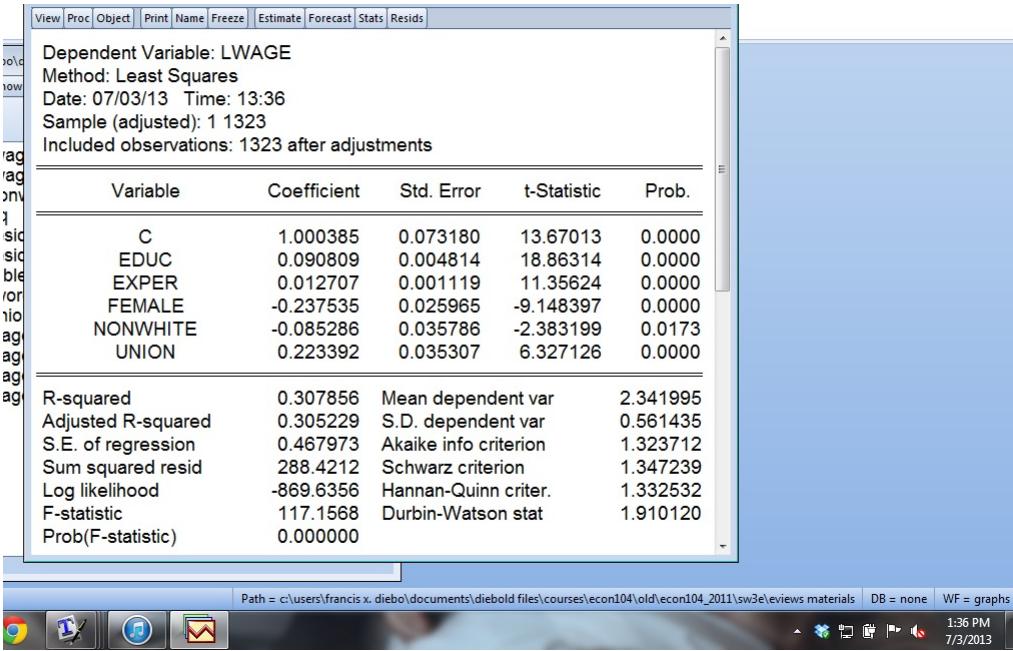


Figure 6.1: Basic Linear Wage Regression

m is adequate. Significance of the included set of powers of \hat{y}_t can be checked using an F test. This procedure is called RESET (Regression Specification Error Test).

6.6 Non-Linearity in Wage Determination

For convenience we reproduce in Figure 6.1 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.

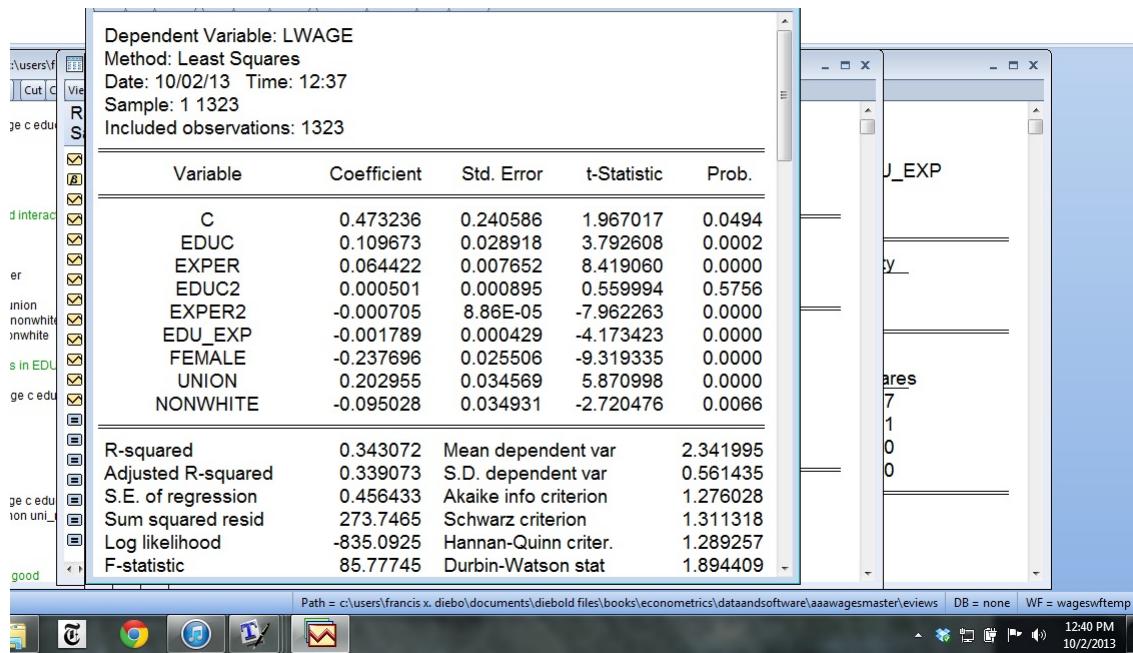


Figure 6.2: Quadratic Wage Regression

In Figure 6.2 we show the results of the quadratic regression

$$LWAGE \rightarrow EDUC, EXPER$$

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

$$FEMALE, UNION, NONWHITE$$

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*.

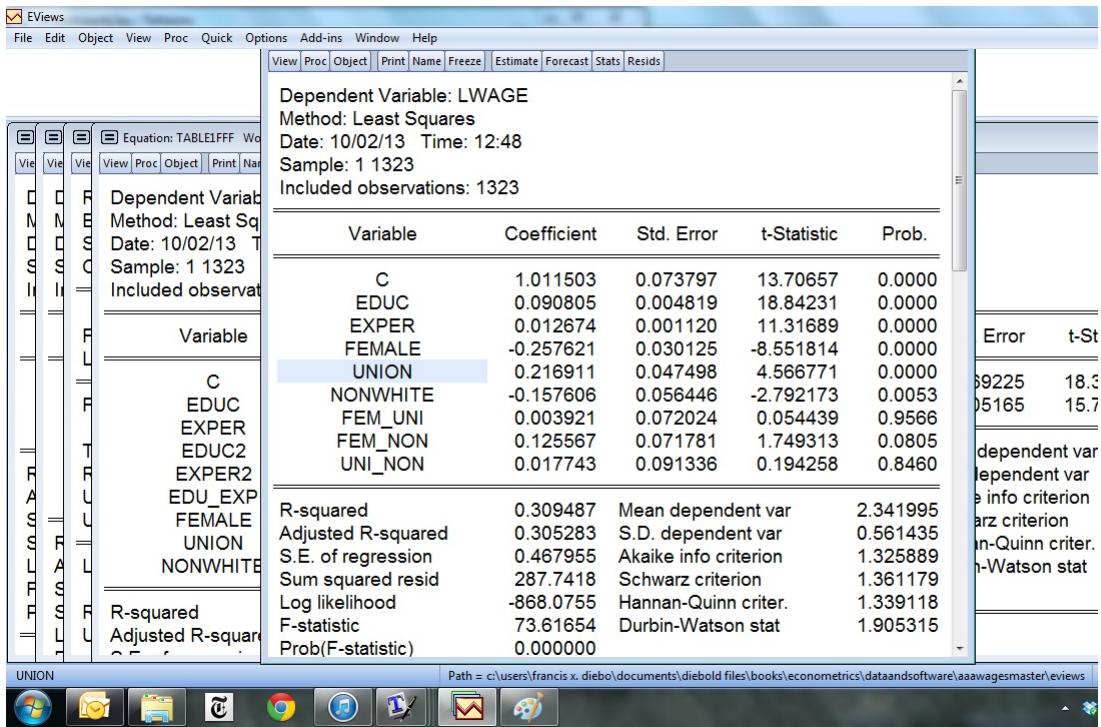


Figure 6.3: Wage Regression on Education, Experience, Group Dummies, and Interactions

In Figure 6.4 we show results for

$$LWAGE \rightarrow EDUC, EXPER,$$

$$FEMALE, UNION, NONWHITE,$$

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

The dummy interactions are insignificant.

6.6.1 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.

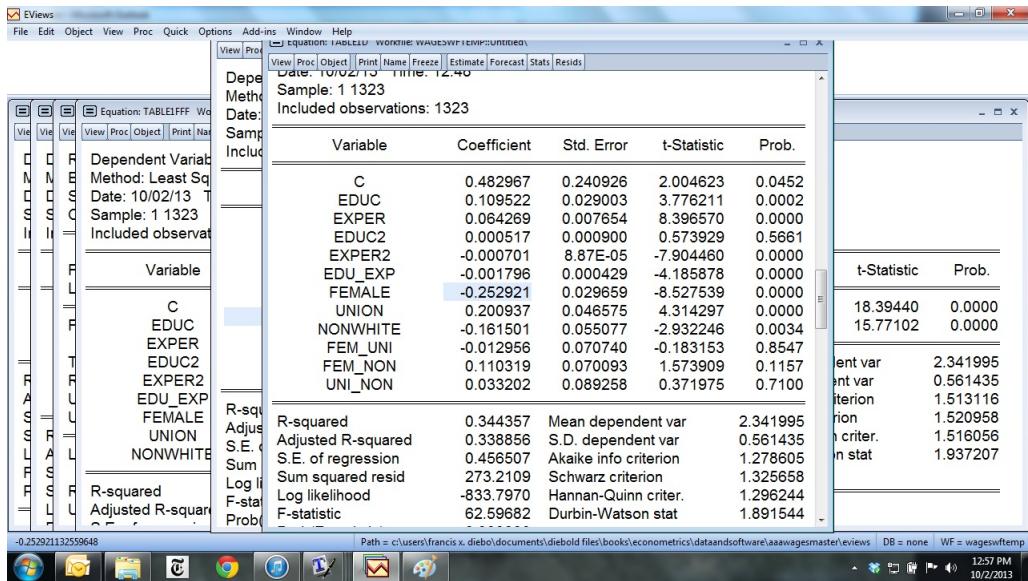


Figure 6.4: Wage Regression with Continuous Non-Linearities and Interactions, and Discrete Interactions

In Figure 6.4 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* 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 6.6,

$$LWAGE \rightarrow EDUC, EXPER,$$

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

$$FEMALE, UNION, NONWHITE.$$

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

6.7 Exercises, Problems and Complements

1. (Non-linear vs. linear relationships)

The U.S. Congressional Budget Office (CBO) is helping the president 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 23 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? Is the relationship likely linear? (Hint: how much revenue would be collected at a tax rate of zero percent? Fifty percent? One hundred percent?)

2. (Graphical regression diagnostic: scatterplot of e_t vs. x_t)

This plot helps us assess whether the relationship between y and the set of x 's 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_t vs. x_t . 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,

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

under the full ideal conditions. Find the mean of y_t conditional upon $x_t = x_t^*$ and $z_t = z_t^*$. Is the conditional mean linear in $(x_t^*? z_t^*)$?

5. (OLS vs. NLS)

Consider the following three regression models:

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

$$y_t = \beta_1 e^{\beta_2 x_t} \varepsilon_t$$

$$y_t = \beta_1 + e^{\beta_2 x_t} + \varepsilon_t.$$

- 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. (Graphical regression diagnostic: scatterplot of e_t vs. x_t)

This plot helps us assess whether the relationship between y and the set of x ’s 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_t vs. x_t . 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.

7. (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.

6.8 Notes

Chapter 7

Non-Linearity in Time Series

In time series a central issue is nonlinear *trend*. Here we focus on it.

7.1 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.¹

Exponential trend is common because economic variables often display roughly constant real growth rates (e.g., two percent per year). If trend is 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

$$\ln(Trend_t) = \ln(\beta_1) + \beta_2 TIME_t. \quad (7.1)$$

Thus, $\ln(Trend_t)$ is a linear function of time.

In Figure 7.1 we show the variety of exponential trend shapes that can be obtained depending on the parameters. Depending on the signs and sizes

¹Throughout this book, logarithms are *natural* (base e) logarithms.

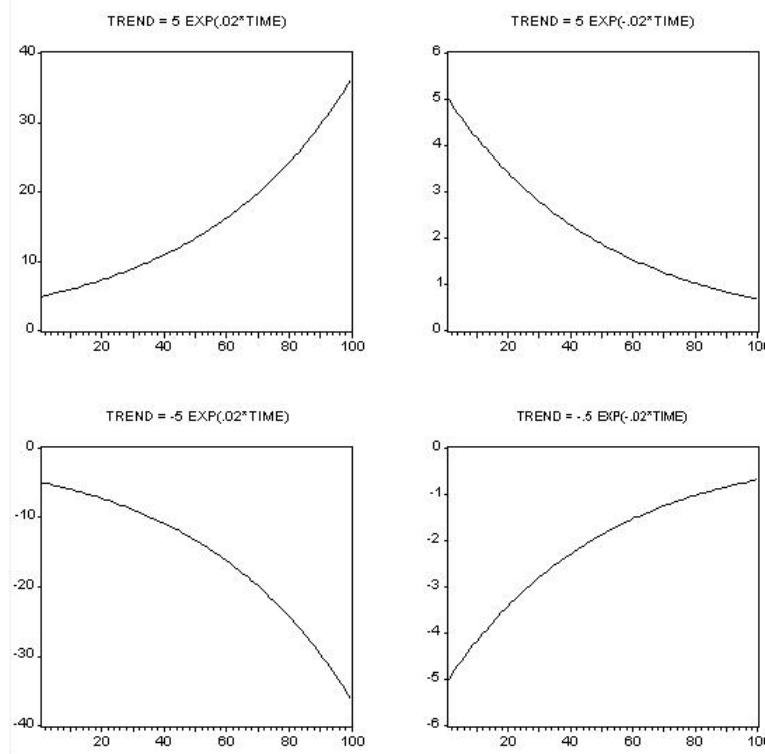


Figure 7.1: Various Exponential Trends

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 (7.1) 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 be achieved with quadratic and exponential trend, there are sub-

tle 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.

7.2 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 $TIME$.² 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 7.2. 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 $TIME2_t = TIME_t^2$. Because $TIME = (1, 2, \dots, T)$, $TIME2 = (1, 4, \dots, T^2)$.

²Higher-order **polynomial trends** are sometimes entertained, but it's important to use low-order polynomials to maintain smoothness.

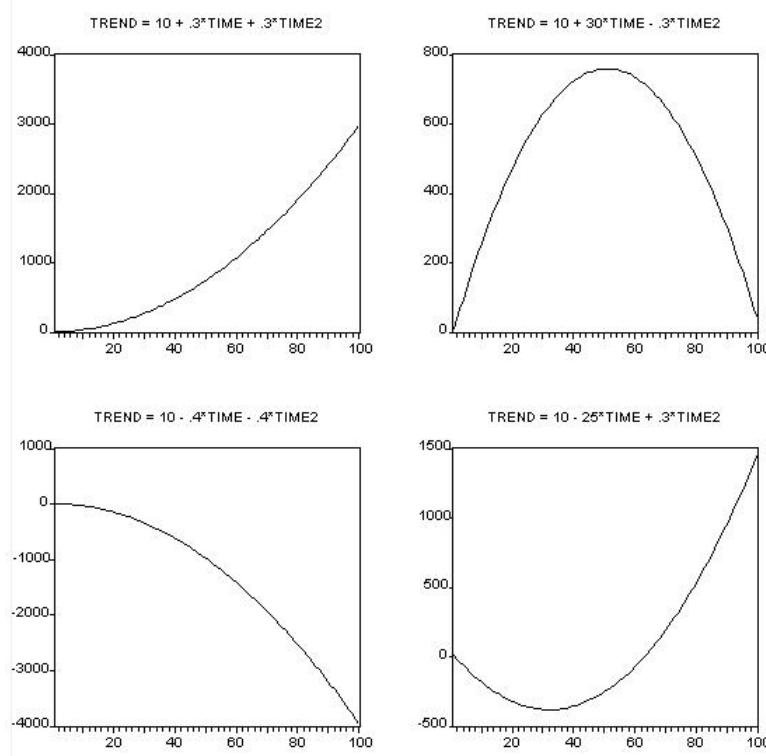


Figure 7.2: Various Quadratic Trends

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 $TIME2$.

7.3 More on Non-Linear Trend

The trend regression technique is one way to estimate trend. Two additional ways involve model-free **smoothing** techniques. They are moving-average smoothers and Hodrick-Prescott smoothers. We briefly introduce them here.

7.3.1 Moving-Average Trend and De-Trending

We'll focus on three: 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 $\{s_t\}_{t=1}^T$. Then the **two-sided moving average** is

$$s_t = (2m + 1)^{-1} \sum_{i=-m}^m y_{t-i},$$

the **one-sided moving average** is

$$s_t = (m + 1)^{-1} \sum_{i=0}^m y_{t-i},$$

and the **one-sided weighted moving average** is

$$s_t = \sum_{i=0}^m w_i y_{t-i},$$

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?

7.3.2 Hodrick-Prescott Trend and De-Trending

A final 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.)

7.4 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 7.3. 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 7.4) 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?)

In Figure 7.5 we show the results of regression on quadratic trend and a full set of seasonal dummies. The trend remains highly significant, and

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 7.3: Log-Quadratic Trend Estimation

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 7.6 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.

7.5 Exercises, Problems and Complements

1. (Properties of polynomial trends)

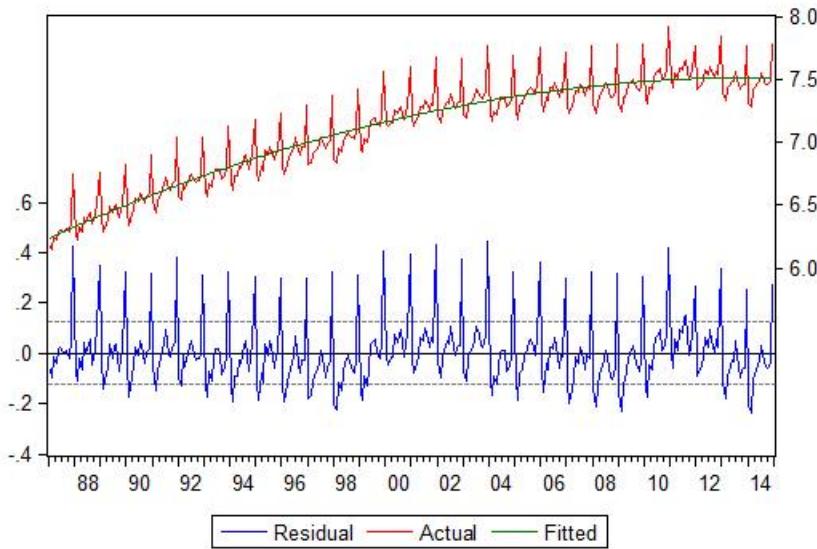


Figure 7.4: Residual Plot, Log-Quadratic Trend Estimation

Consider a sixth-order deterministic polynomial trend:

$$T_t = \beta_1 + \beta_2 \text{TIME}_t + \beta_3 \text{TIME}_t^2 + \dots + \beta_7 \text{TIME}_t^6.$$

- 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.
2. (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?
3. (Difficulties with non-linear optimization)

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 7.5: Liquor Sales Log-Quadratic Trend Estimation with Seasonal Dummies

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. (Direct estimation of exponential trend in levels)

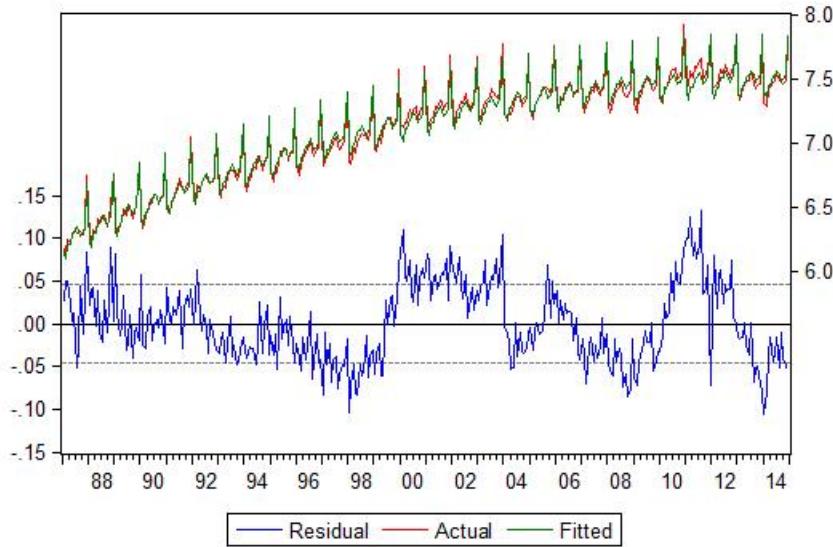


Figure 7.6: Residual Plot, Liquor Sales Log-Quadratic Trend Estimation With Seasonal Dummies

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.)
5. (Logistic trend)

In the main text we introduced the logistic functional form. A key

example is **logistic trend**, which is

$$Trend_t = \frac{1}{a + br^{TIME_t}},$$

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.

6. (Empirical Liquor Sales Analysis)

Consider the liquor sales data. Work throughout with log liquor sales (LSALES). Never include an intercept. Discuss all results in detail.

- (a) Replicate the linear trend + seasonal results from the text.
From this point onward, include three autoregressive lags of LSALES in every model estimated.
- (b) Contrast linear, exponential and quadratic trends for LSALES, and simultaneously consider "tightening" the seasonal specification to include fewer than 12 seasonal dummies. What is your "best" model?
- (c) What do the autoregressive lags do? Do they seem important?
- (d) Assess the robustness of your "best" model by re-estimating it using LAD.
- (e) What is your "final" "best" model? Are you completely happy with it? Why or why not? (And what does that even mean?)

7.6 Notes

Chapter 8

Binary Regression and Classification

8.1 Binary Regression

Another appearance of dummy variables: LHS.

Here we work with “**limited dependent variables**,” meaning that they can take only a limited number of values. The classic case is a 0-1 “dummy variable.” **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.

8.1.1 Binary Response

Note that the basic regression model,

$$y_t = x'_t \beta + \varepsilon,$$

immediately implies that

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

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

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

In that case we have

$$E(I_t(z)|x_t) = x_t' \beta.$$

A key insight, however, is that

$$E(I_t(z)|x_t) = P(I_t(z) = 1|x_t),$$

so the model is effectively

$$P(I_t(z) = 1|x_t) = x_t' \beta. \quad (8.1)$$

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. Leading examples include recessions, bankruptcies, loan or credit card defaults, financial market crises, and consumer choices.

But how should we “fit a line” when the LHS variable is binary? The **linear probability model** does it by brute-force OLS regression $I_t(z) \rightarrow x_t$. 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 $E(I_t(z)|x_t) = P(I_t(z) = 1|x_t)$ to lie in the unit interval, in which probabilities must of course lie. We now consider models that impose that constraint by running $x_t' \beta$ through a monotone “squashing function,” $F(\cdot)$, that keeps $P(I_t(z) = 1|x_t)$ in the unit interval. That is, we move to models with

$$P(I_t(z) = 1|x_t) = F(x_t' \beta),$$

where $F(\cdot)$ is monotone increasing, with $\lim_{w \rightarrow \infty} F(w) = 1$ and $\lim_{w \rightarrow -\infty} F(w) = 0$.

0. Many squashing functions can be entertained, and many *have* been entertained.

8.1.2 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.

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

$$P(I_t(z) = 1|x_t) = \frac{e^{x'_t \beta}}{1 + e^{x'_t \beta}}.$$

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 deep and useful insights. In particular, consider a latent variable, y_t^* , where

$$y_t^* = x'_t \beta + \varepsilon_t$$

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

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

$$\begin{aligned} E(I_t(y_t^* > 0)|x_t) &= P(y_t^* > 0|x_t) = P(\varepsilon_t > -x'_t \beta) \\ &= P(\varepsilon_t < x'_t \beta) \text{ (by symmetry of the logistic density of } \varepsilon) \end{aligned}$$

$$= \frac{e^{x_t' \beta}}{1 + e^{x_t' \beta}},$$

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_t^* with an observed indicator that turns “on” when $y_t^* > 0$ – 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, bond ratings as a function of continuous underlying firm “health,” etc.

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

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”; S&P and Moody’s bond ratings in several “buckets”; and surveys of that ask about degree of belief in several categories ranging from “strongly disagree” through “strongly agree.”

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

$$y_t^* = x_t' \beta + \varepsilon_t$$

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

But now we have an indicator with a finer gradation:

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

We can estimate this **ordered logit** model by maximum likelihood, just as with the standard logit model. All interpretation remains the same, so long as relationships between all pairs of outcome groups are the same, an assumption known as *proportional odds*.

Dynamic Logit

Note that one or more of the x variables could be lagged dependent variables, $I_{t-i}(z)$, $i = 1, 2, \dots$

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,

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

because f must be positive. In addition, ratios of β 's do give ratios of effects, because the f 's cancel.

$$R^2$$

Recall that traditional R^2 for continuous LHS variables is

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

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

$$R^2 = 1 - \frac{\sum(y_t - \hat{P}(I_t(z) = 1|x_t))^2}{\sum(y_t - \bar{y}_t)^2}.$$

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.

8.1.3 Classification and “0-1 Forecasting”

- Examples: Make loan or not, grant credit card or not, hire a worker or not, will consumer buy or not
 - Classification maps probabilities into 0-1 forecasts. Bayes classifier uses a cutoff of .5.
 - Decision boundary. Suppose we use a Bayes classifier.

We predict 1 when $\text{logit}(x'\beta) > 1/2$. But that's the same as predicting 1 when $x'\beta > 0$. If there are 2 x 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.

** Figures illustrating linear and non-linear decision boundaries

8.2 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 (8.1). 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** $O(I_t(z) = 1|x_t)$ of an event z are just a simple transformation of its probability

$$O(I_t(z) = 1|x_t) = \frac{P(I_t(z) = 1|x_t)}{1 - P(I_t(z) = 1|x_t)}.$$

Consider a linear model for log odds

$$\ln \left(\frac{P(I_t(z) = 1|x_t)}{1 - P(I_t(z) = 1|x_t)} \right) = x_t' \beta.$$

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

$$P(I_t(z) = 1|x_t) = \frac{1}{1 + e^{-x_t' \beta}} = \frac{e^{x_t' \beta}}{1 + e^{x_t' \beta}}.$$

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

A full statement of the model is

$$y_t \sim Bern(p_t)$$

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

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

$$P(I_t(z) = 1|x_t) = \Phi(x'_t \beta),$$

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,

$$E(y_t|x_t) = G^{-1}(x'_t \beta),$$

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_t , where $y_t = y_t^*$ if $y_t^* \geq 0$, and 0 otherwise, and where

$$y_t^* = \beta_0 + \beta_1 x_t + \varepsilon_t.$$

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...

8.3 Notes

Chapter 9

Measurement Error

Suppose the DGP is

$$y_t = \beta_1 + \beta_2 x_t + \varepsilon_t,$$

but that we can't measure x_t accurately. Instead we measure

$$x_t^m = x_t + v_t.$$

Think of v_t as an *iid* measurement error with variance σ_v^2 . (Assume that it is also independent of ε_t) Clearly, as σ_v^2 gets larger relative to σ_x^2 , the fitted regression

$$y \rightarrow c, x^m$$

is progressively less able to identify the true relationship. In the limit as $\sigma_v^2 / \sigma_x^2 \rightarrow \infty$, it is impossible. In any event, $\hat{\beta}_{LS}$ is biased toward zero, in small as well as large samples. We speak of the “errors in variables” problem.

9.1 Exercises, Problems and Complements

1. ***

9.2 Notes

Chapter 10

Omitted Variables

10.1 Omitted Relevant Variables

Many of the diagnostic statistics that we introduced in this chapter are designed to detect violations of the full ideal conditions (e.g., the Durbin-Watson test for serial correlation in disturbances). Omission of relevant variables is another violation of the full ideal conditions.

Suppose that the DGP is

$$y_t = \beta_1 + \beta_2 z_t + \varepsilon_t,$$

but that we incorrectly run

$$y \rightarrow c, x$$

where $\text{corr}(x_t, z_t) > 0$.

Clearly we'll estimate a positive effect of x on y , in large as well as small samples, even though it's completely spurious and would vanish if z had been included in the regression. The positive bias arises because in our example we assumed that $\text{corr}(z_t, z_t) > 0$; in general the sign of the bias could go either way. We speak of "**omitted variable bias**."

Note that 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’s not useful for determining the effect on y of an exogenous shift in x (the true effect is 0!) (“consistency for a treatment effect”).

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

10.2 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 a variable included in a regression is irrelevant? A set of variables?

10.3 Exercises, Problems and Complements

1. (Omitted Variables, Again and Again)

Notice that omitted variables have also arisen repeatedly in our discussions.

- (a) If there are neglected group effects in cross-section regression, we fix the problem (of omitted group dummies) by including the requisite group dummies.
- (b) 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.
- (c) If there is neglected non-linearity, we fix the problem (effectively one of omitted Taylor series terms) by including the requisite Taylor series terms.

- (d) 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 FIC will be satisfied). This is an important recognition. In subsequent chapters, for example, we’ll study violations of the FIC known as heteroskedasticity (Chapters 13 and 16) and serial correlation (Chapter 14.5). In each case 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.

10.4 Notes

Chapter 11

Multicollinearity

Collinearity refers to two x variables that are highly correlated. But all pairwise correlations could be small and yet an x variable is 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. Hence collinearity is a special case of multicollinearity, and from now on we will simply speak of multicollinearity.

There are two types of multicollinearity, perfect and imperfect. As we'll see, *perfect* multicollinearity is disastrous, but unlikely to occur unless you do something really dumb. Imperfect multicollinearity, in contrast occurs routinely but is not necessarily problematic, although in extreme cases it may require some attention.

11.1 Perfect Multicollinearity

Perfect multicollinearity refers to perfect correlation among some regressors, or linear combinations of regressors. The classic example is the dummy variable trap, in which we include a full set of dummies *and* an intercept. Perfect multicollinearity is indeed a problem; $X'X$ matrix is singular, so $(X'X)^{-1}$ does not exist, and the OLS estimator cannot even be computed!

But perfect multicollinearity will only arise if you do something really

dumb (like including both a full set of dummies and an intercept). In any event the solution is trivial: simply drop one of the redundant variables.

11.2 Imperfect Multicollinearity

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 FIC. 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.

11.3 A Bit More

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 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,”

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

which is just a transformation of R_k^2 .

11.4 Exercises, Problems and Complements

1. ***

11.5 Notes

Chapter 12

Non-normality and Outliers

Much of econometrics is about *failure* of one or more of the FIC. Here we consider non-normality and the related issue of outliers.

12.1 Non-Normality

Recall the full ideal conditions.

Here we consider a violation of the full ideal conditions, namely non-normal disturbances.

Non-normality and **non-linearity**, which we already studied in chapter 12, are very closely related. 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.

Non-normality and **outliers**, which we introduce in this chapter, are also 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).

12.1.1 OLS Without Normality

To understand the properties of OLS under non-normality, it is helpful first to consider the properties of the sample mean (which, as you know from Problem 6 of Chapter 3, is just OLS regression on an intercept.)

As reviewed in Chapter A, for a Gaussian simple random sample,

$$y_t \sim iidN(\mu, \sigma^2), i = 1, \dots, T,$$

the sample mean \bar{y} is unbiased, consistent, normally distributed with variance σ^2/T , and indeed the minimum variance unbiased (MVUE) estimator. We write

$$\bar{y} \sim N\left(\mu, \frac{\sigma^2}{T}\right),$$

or equivalently

$$\sqrt{T}(\bar{y} - \mu) \sim N(0, \sigma^2).$$

Moving now to a non-Gaussian simple random sample,

$$y_t \sim iid(\mu, \sigma^2), i = 1, \dots, T,$$

as also reviewed in Chapter A, we still have that \bar{y} is unbiased, consistent, *asymptotically* normally distributed with variance σ^2/T , and best linear unbiased (BLUE). We write,

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

or more precisely, as $T \rightarrow \infty$,

$$\sqrt{T}(\bar{y} - \mu) \rightarrow_d N(0, \sigma^2).$$

This result forms the basis for asymptotic inference. It is a Gaussian central limit theorem, and it also has a law of large numbers ($\bar{y} \rightarrow_p \mu$) imbedded within it.

Now consider the full linear regression model under normality. The least squares estimator is

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

Recall from Chapter 3 that under the full ideal conditions (which include normality) $\hat{\beta}_{LS}$ is consistent, normally distributed with covariance matrix $\sigma^2(X'X)^{-1}$, and indeed MVUE. We write

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

Moving now to non-Gaussian regression, we still have that $\hat{\beta}_{LS}$ is consis-

tent, *asymptotically* normally distributed, and BLUE. We write

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

or more precisely,

$$\sqrt{T}(\hat{\beta}_{LS} - \beta) \rightarrow_d N\left(0, \sigma^2 \left(\frac{X'X}{T}\right)^{-1}\right).$$

Clearly the linear regression results for Gaussian vs. non-Gaussian situations precisely parallel those for the sample mean in Gaussian vs. non-Gaussian situations. Indeed they must, as the sample mean corresponds to regression on an intercept.

12.1.2 Assessing Normality

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

Nonparametric Density Estimation

The Residual QQ Plot

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.

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} .

The Jarque-Bera Test

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

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

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

12.2 Outliers and Leverage

12.2.1 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

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

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 perfectly “clean” datasets have 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.

12.2.2 Outlier Detection

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, residual $\hat{y}y$ scatterplots remain invaluable, as does the good-old residual plot of $y - \hat{y}$.

Other quantities also inform us about aspects of fatness of tails. One such quantity is the outlier probability,

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

(there is of course nothing magical about our choice of 5).

Another possibility is the “tail index” γ , such that

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

In practice, as always, we use sample versions of the above population objects to inform us about different aspects of non-normality.

12.2.3 Leverage

For observation i in simple regression, the leverage statistic is

$$l_i = \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^N (x_{i'} - \bar{x})^2}.$$

This is a special case of the statistic for the general multiple regression,

$$l_i = (X(X'X)^{-1}X')_{ii},$$

the i^{th} diagonal element of the $N \times N$ “hat matrix, $X(X'X)^{-1}X'$. Note that $l_i \in [0, 1]$. Average leverage across all observations is K/N , so we examine the leverage of any individual observation relative to K/N .

It is sometimes useful to graph standardized residuals vs. leverage. In that way, outliers and leverage can be assessed simultaneously.

12.2.4 Robust Estimation

Robust estimation provides a useful middle ground between completely discarding allegedly-outlying observations (“dummying 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.

Robustness Iteration

Fit at robustness iteration 0:

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

where

$$\hat{\beta}^{(0)} = \operatorname{argmin} \left[\sum_{t=1}^T (y_t - x'_t \beta)^2 \right].$$

Robustness weight at iteration 1:

$$\rho_t^{(1)} = S \left(\frac{\hat{\varepsilon}_t^{(0)}}{6h} \right)$$

where

$$\hat{\varepsilon}_t^{(0)} = y_t - \hat{y}_t^{(0)}$$

$$h = \operatorname{med} |\hat{\varepsilon}_t^{(0)}|.$$

Fit at iteration 1:

$$\hat{y}_t^{(1)} = x' \hat{\beta}^{(1)}$$

where

$$\hat{\beta}^{(1)} = \operatorname{argmin} \left[\sum_{t=1}^T \rho_t^{(1)} (y_t - x'_t \beta)^2 \right].$$

Continue as desired.

Least Absolute Deviations

Recall that the OLS estimator solves

$$\min_{\beta} \sum_{t=1}^T (y_t - \beta_1 - \beta_2 x_{2t} - \dots - \beta_K x_{Kt})^2.$$

Now we simply change the objective to

$$\min_{\beta} \sum_{t=1}^T |y_t - \beta_1 - \beta_2 x_{2t} - \dots - \beta_K x_{Kt}|.$$

That is, we change from squared-error loss to absolute-error loss. We call the new estimator “**least absolute deviations**” (LAD). By construction, the solution under absolute-error loss is not influenced by outliers as much as the solution under quadratic loss. 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.²

12.2.5 Wages and Liquor Sales

Wages

Use best *LWAGE* model, including non-linear effects.

Residual plot.

Residual scatterplot.

²Indeed computation of the *LAD* estimator turns out to be a linear programming problem, which is well-studied and simple.

Residual histogram and stats

Residual Gaussian QQ plot

LAD estimation

Liquor Sales

Hist and stats for $LSALES$ regression residuals

QQ plot for $LSALES$ regression residuals

$LSALES$ regression residual plots

$LSALES$ regression (include three AR terms) residual plots

$LSALES$ regression (include three AR terms) coefficient influence plots

$LSALES$ regression (include three AR terms) LAD estimation

12.3 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. (Bootstrapping Standard Errors)

We can go farther than trying to robustify ourselves to failures of the normality assumption. Rather, using simulation methods we can embrace the precise nature of the non-normality (and small samples) with which we are confronted.

3. (“Leave-One-Out” Plots)

We seek to determine which observations have the most impact on which estimated parameters. In a “leave-one-out” plot, for example, we use the computer to sweep through the sample, leaving out successive observations, examining differences in parameter estimates with observation t in vs. out. That is, in an obvious notation, we examine and plot

$$\hat{\beta}_k - \hat{\beta}_k(-t),$$

or some suitably-scaled version thereof, $k = 1, \dots, K, t = 1, \dots, T$.

This procedure is more appropriate for cross-section data than for time-series data. Why?

12.4 Notes

The Jarque-Bera test is developed in [Jarque and Bera \(1987\)](#). [Koenker \(2005\)](#) provides extensive discussion of *LAD* and its extensions. *LAD* will later lead us to the idea of quantile regression, which is a way of fitting lines to describe not only the conditional median of $y|X$ (which is the conditional mean under symmetry), but also other conditional percentiles.

Chapter 13

Heteroskedasticity in Cross-Sections

Generalized Least Squares (GLS)

Consider the FIC except that we now let:

$$\varepsilon \sim N(\underline{0}, \sigma^2 \Omega)$$

The old case is $\Omega = I$, but things are very different when $\Omega \neq I$:

- OLS parameter estimates consistent but inefficient
(no longer MVUE or BLUE)
- OLS standard errors are biased and inconsistent. Hence t ratios do not have the t distribution in finite samples and do not have the $N(0, 1)$ distribution asymptotically

The GLS estimator is:

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

Under the remaining full ideal conditions it is consistent, normally distributed with covariance matrix $\sigma^2 (X' \Omega^{-1} X)^{-1}$, and MVUE:

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

Heteroskedasticity in Cross-Section Regression

Homoskedasticity: variance of ε_i is constant across i

Heteroskedasticity: variance of ε_i is not constant across i

Relevant cross-sectional heteroskedasticity situation
(on which we focus for now):

ε_i independent across i but not identically distributed across i

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

- Can arise for many reasons
- Engel curve (e.g., food expenditure vs. income) is classic example

Consequences

OLS inefficient (no longer MVUE or BLUE),
in finite samples and asymptotically

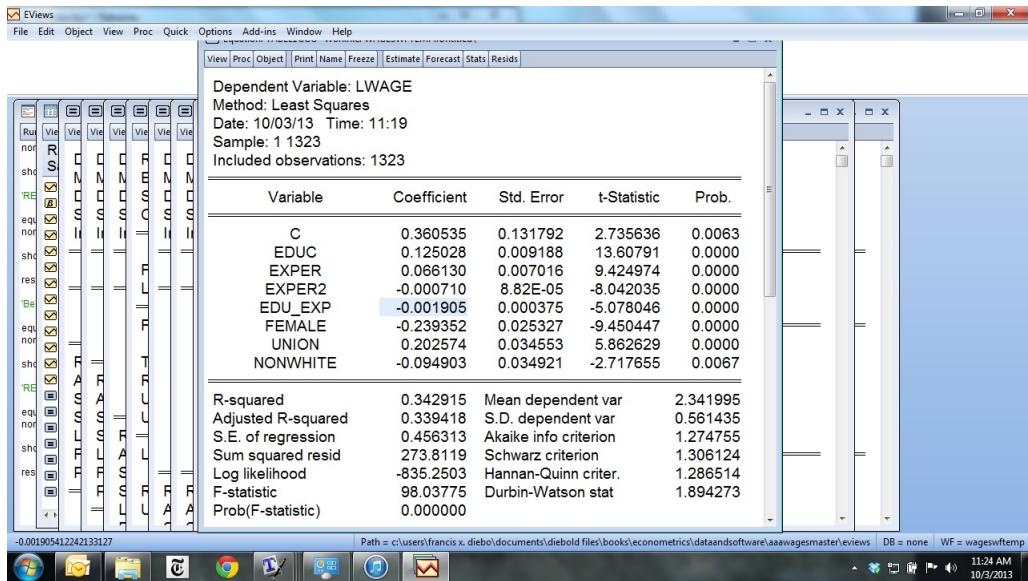
Standard errors biased and inconsistent.

Hence t ratios do not have the t distribution in finite samples
and do not have the $N(0, 1)$ distribution asymptotically

Detection

- Graphical heteroskedasticity diagnostics
- Formal heteroskedasticity tests

Graphical Diagnostics



Graph e_i^2 against x_i , for various regressors

Problem: Purely pairwise

Recall Our “Final” Wage Regression

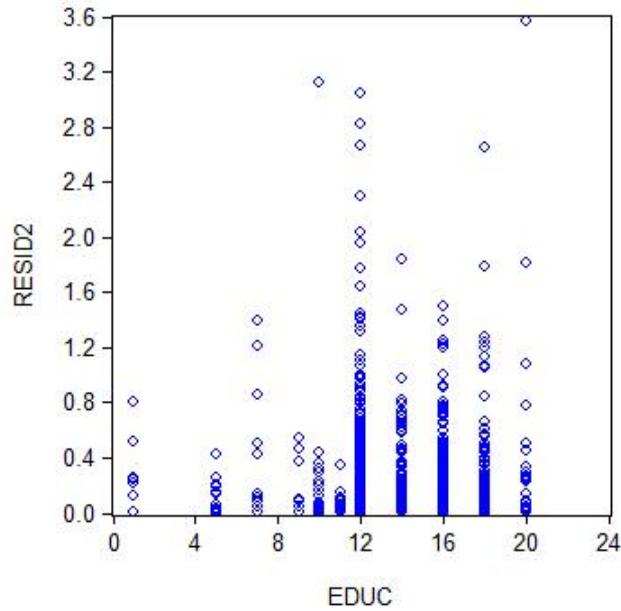
Squared Residual vs. EDUC

The Breusch-Godfrey-Pagan Test (BGP)

- Estimate the OLS regression, and obtain the squared residuals
- Regress the squared residuals on all regressors
- To test the null hypothesis of no relationship, examine NR^2 from this regression. In large samples $NR^2 \sim \chi^2$ under the null.

BGP Test

White's Test



- Estimate the OLS regression, and obtain the squared residuals
- Regress the squared residuals on all regressors, squared regressors, and pairwise regressor cross products
- To test the null hypothesis of no relationship, examine NR^2 from this regression. In large samples $NR^2 \sim \chi^2$ under the null.
 (White's test is a natural and flexible generalization of the Breusch-Godfrey-Pagan test)

White Test

GLS for Heteroskedasticity

- “Weighted least squares” (WLS)
 - Take a stand on the DGP. Get consistent standard errors and efficient parameter estimates.

EViews

File Edit Object View Proc Quick Options Add-ins Window Help

Is wage c educ exper exper2 educ

View Proc Object Print Name Freeze Estimate Forecast Stats Resids

Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	5.414870	Prob. F(7,1315)	0.0000
Obs*R-squared	37.06628	Prob. Chi-Square(7)	0.0000
Scaled explained SS	49.66045	Prob. Chi-Square(7)	0.0000

Test Equation:

Dependent Variable: RESID^2

Method: Least Squares

Date: 10/30/13 Time: 10:54

Sample: 1 1323

Included observations: 1323

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.170309	0.097349	-1.749473	0.0804
EDUC	0.024074	0.006787	3.547204	0.0004
EXPER	0.011701	0.005183	2.257616	0.0241
EXPER2	-5.53E-05	6.52E-05	-0.849150	0.3960
EDU_EXP	-0.000478	0.000277	-1.725513	0.0847
FEMALE	-0.009757	0.018708	-0.521530	0.6021
UNION	-0.079648	0.025523	-3.120623	0.0018
NONWHITE	0.000486	0.025794	0.018829	0.9850

Path = c:\users\francis x. diebo\documents\diebold files\books\ehconomics\dataandsoftware\aaawagesmaster\ewviews DB = none WF = wageswtemp

10:55 AM 10/30/2013

EViews

File Edit Object View Proc Quick Options Add-ins Window Help

Is wage c educ exper exper2 educ_exp female union nonwhite

Workfile: WAGESWTEMP - (c:\users\francis x. diebo\documents\diebold files\books\ehconomics\dataandsoftware\aaawagesmaster\ewviews)

View Proc Object Print Save Details Show Fetch Store Delete Genr Sample

Range: 1 1400 -- 1400 obs Filter: *

Sample: 1 1323 -- 1323 obs

age c educ educ2 exper exper2 fem_non fem_uni female i wage nonwhite resid table1 table1a table1b table1c table1d table1dd table1ddd

table1 table1f

Equation: UNTITLED Workfile: WAGESWTEMP-Untitled

Heteroskedasticity Test: White

F-statistic	2.431488	Prob. F(29,1293)	0.0000
Obs*R-squared	68.41804	Prob. Chi-Square(29)	0.0000
Scaled explained SS	91.66473	Prob. Chi-Square(29)	0.0000

Path = c:\users\francis x. diebo\documents\diebold files\books\ehconomics\dataandsoftware\aaawagesmaster\ewviews DB = none WF = wageswtemp

11:00 AM 10/30/2013

(Infeasible) Weighted Least Squares

DGP:

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

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

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 DGP is now:

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

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

- OLS is MVUE!
- Problem: We don't know σ_i^2

Remark on Weighted Least Squares

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

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

Feasible Weighted Least Squares

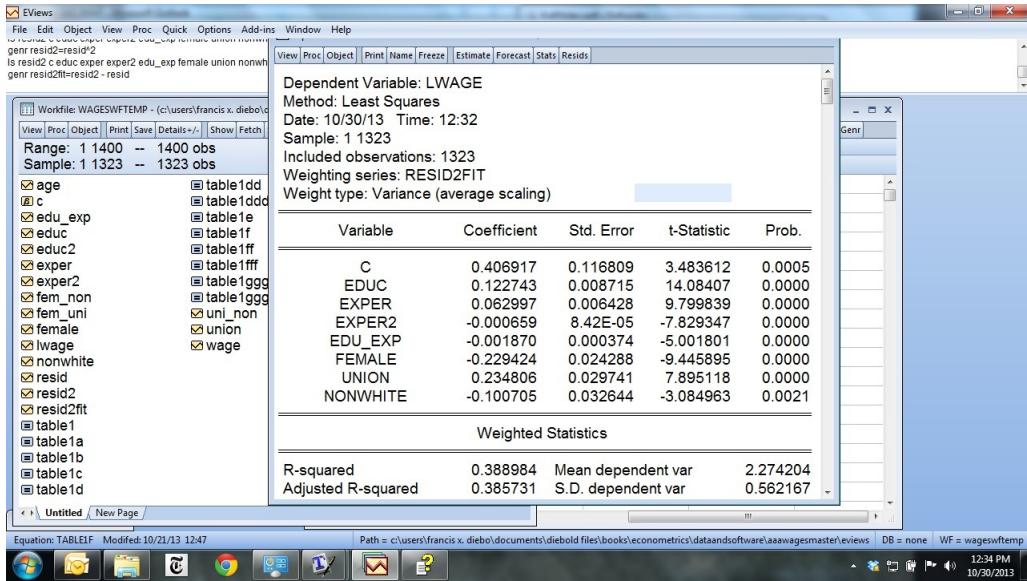
Intuition: Replace the unknown σ_i^2 values with estimates

Some good ideas:

- Use $w_i = 1/\hat{e}_i^2$, where \hat{e}_i^2 are from the BGP test regression
- Use $w_i = 1/\hat{e}_i^2$, where \hat{e}_i^2 are from the White test regression

What about WLS directly using $w_i = 1/e_i^2$?

- Not such 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 $e^2 \rightarrow X$

Regression Weighted by Fit From White Test Regression

A Different Approach

(Advanced but Very Important)

White's Heteroskedasticity-Consistent Standard Errors

Perhaps surprisingly, we make direct use of e_i^2

Don't take a stand on the DGP

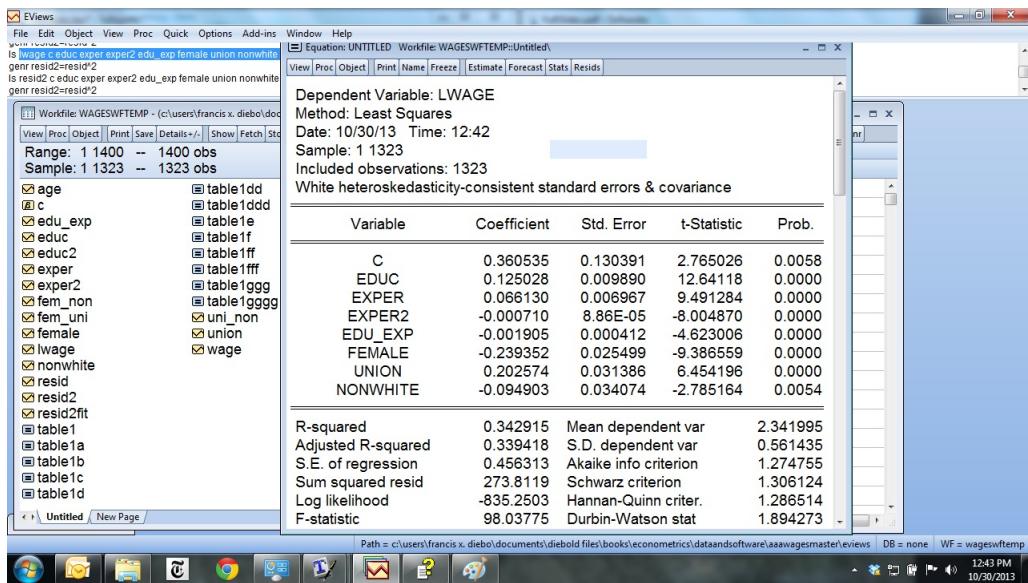
Give up on efficient parameter estimates, but get consistent s.e.'s.

Using advanced methods, one *can* obtain consistent s.e.'s (if not an efficient $\hat{\beta}$) using only e_i^2

- Standard errors are rendered consistent.
- $\hat{\beta}$ remains unchanged at its OLS value. (Is that a problem?)

"Robustness to heteroskedasticity of unknown form"

Regression with White's Heteroskedasticity-Consistent Standard Errors



13.1 Exercises, Problems and Complements

1. (Robustness iteration)

Do a second-stage WLS with weights $1/|e_t|$, or something similar. This is not a heteroskedasticity correction, but a purely mechanical strategy to downweight outliers.

2. (Vocabulary)

All these have the same meaning: White standard errors, “White-washed” standard errors, heteroskedasticity-robust standard errors, heteroskedasticity-consistent standard errors, and robust standard errors.

13.2 Notes

Chapter 14

Serial Correlation in Time Series

Observed Time Series.

14.1 Characterizing Time-Series Dynamics

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

Cycles, according to our broad interpretation, may display the sort of back-and-forth movement characterized in Figure ???, but they don't have to. All we require is that there be 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 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. Thus we split the discussion into three parts. First we develop methods for *characterizing* cycles, and then we introduce *models* of cycles. All of the material is crucial, and it's also a bit difficult the first time around because it's unavoidably rather mathematical, so careful, systematic study is required.

14.1.1 Covariance Stationary Time Series

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 $\gamma(0) = cov(y_t, y_t) = var(y_t)$.

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 is simple: whether we work directly in levels and include special components for the nonstationary elements of our models, or we work on transformed data such as growth rates, the covariance stationarity assumption is not as unrealistic as it may seem.

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

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

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,

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

as claimed. Note that we always have $\rho(0) = \frac{\gamma(0)}{\gamma(0)} = 1$, because any series

is perfectly correlated with itself. Thus the autocorrelation at displacement 0 isn't of interest; rather, only the autocorrelations *beyond* displacement 0 inform us about a series' dynamic structure.

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

As with the autocorrelations, we often graph the partial autocorrelations 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 *** we show an autocorrelation function that displays gradual one-sided damping, and in Figure *** we show a constant autocorrelation function; the latter could not be the autocorrelation function of a stationary process, whose autocorrelation function must eventually decay. The precise decay patterns

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.

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.

of autocorrelations and partial autocorrelations of a covariance stationary series, however, depend on the specifics of the series. In Figure ***, for example, we show an autocorrelation function that displays damped oscillation – the autocorrelations are positive at first, then become negative for a while, then positive again, and so on, while continuously getting smaller in absolute value. Finally, in Figure *** we show an autocorrelation function that differs in the way it approaches zero – the autocorrelations drop abruptly to zero beyond a certain displacement.

14.2 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_t 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.⁴ Sometimes for short we write

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

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.⁵ If in addition to being serially uncorrelated, y is serially independent, then we say that y is **independent white noise**.⁶ 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**.⁷ 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 *** 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.

⁴It’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.

⁵Recall that zero correlation implies independence only in the normal case.

⁶Another name for independent white noise is **strong white noise**, in contrast to standard serially uncorrelated **weak white noise**.

⁷Carl Friedrich Gauss, one of the greatest mathematicians of all time, discovered the normal distribution some 200 years ago; hence the adjective “Gaussian.”

Recall that the disturbance in a regression model is typically assumed to be white noise of one sort or another. There's a subtle difference here, however. Regression disturbances are not observable, whereas we're working with an observed series. Later, however, we'll see how all of our models for observed series can be used to model unobserved variables such as regression disturbances. Let's characterize the dynamic stochastic structure of white noise, $y_t \sim WN(0, \sigma^2)$. By construction the unconditional mean of y is $E(y_t) = 0$, and the unconditional variance of y is $\text{var}(y_t) = \sigma^2$.

Note 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.⁸

To understand fully the linear dynamic structure of a covariance stationary time series process, we need to compute and examine its mean and its autocovariance function. For white noise, we've already computed the mean and the variance, which is the autocovariance at displacement 0. We have yet to compute the rest of the autocovariance function; fortunately, however, it's very simple. Because white noise is, by definition, uncorrelated over time, all the autocovariances, and hence all the autocorrelations, are zero beyond displacement 0.⁹ Formally, then, the autocovariance function for a white noise process is

$$\gamma(\tau) = \begin{cases} \sigma^2, & \tau = 0 \\ 0, & \tau \geq 1, \end{cases}$$

⁸Recall that $\sigma^2 = \gamma(0)$.

⁹If the autocovariances are all zero, so are the autocorrelations, because the autocorrelations are proportional to the autocovariances.

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 *** we plot the white noise autocorrelation function.

Finally, consider the partial autocorrelation function for a white noise series. For the same reason that the autocorrelation at displacement 0 is always one, so too is the partial autocorrelation at displacement 0. For a white noise process, all partial autocorrelations beyond displacement 0 are zero, which again follows from the fact that white noise, by construction, is serially uncorrelated. Population regressions of y_t on y_{t-1} , or on y_{t-1} and y_{t-2} , or on any other lags, produce nothing but zero coefficients, because the process is serially uncorrelated. Formally, the partial autocorrelation function of a white noise process is

$$p(\tau) = \begin{cases} 1, & \tau = 0 \\ 0, & \tau \geq 1. \end{cases}$$

We show the partial autocorrelation function of a white noise process in Figure ***. 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.¹⁰ 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 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

¹⁰If you need to refresh your memory on conditional means, consult any good introductory statistics book, such as Wonnacott and Wonnacott (1990).

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.

14.3 Estimation and Inference for the Mean, Autocorrelation and Partial Autocorrelation Functions

Now suppose we have a sample of data on a time series, and we don't know the true model that generated the data, or the mean, autocorrelation function or partial autocorrelation function associated with that true model. Instead, we want to use the data to estimate the mean, autocorrelation function, and partial autocorrelation function, which we might then use to help us learn about the underlying dynamics, and to decide upon a suitable model or set of models to fit to the data.

14.3.1 Sample Mean

The mean of a covariance stationary series is

$$\mu = E y_t.$$

A fundamental principle of estimation, called the **analog principle**, suggests that we develop estimators by replacing expectations with sample averages. Thus our estimator for the population mean, given a sample of size T , is the **sample mean**,

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

Typically we're not directly interested in the estimate of the mean, but it's needed for estimation of the autocorrelation function.

14.3.2 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 yields¹¹

$$T\hat{\rho}^2(\tau) \sim \chi_1^2.$$

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.¹² 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

¹¹Recall 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.

¹² m is a maximum displacement selected by the user. Shortly we'll discuss how to choose it.

a large part of the autocorrelation function. On the other hand, as m grows relative to T , the quality of the distributional approximations we've invoked deteriorates. In practice, focusing on m in the neighborhood of \sqrt{T} is often reasonable.

14.3.3 Sample Partial Autocorrelations

Recall that the partial autocorrelations are obtained from population linear regressions, which correspond to a thought experiment involving linear regression using an infinite sample of data. The sample partial autocorrelations 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.

14.4 Autoregressive Models for Serially-Correlated Time Series

14.4.1 Some Preliminary Notation: The Lag Operator

The **lag operator** and related constructs are the natural language in which time series models are expressed. If you want to understand and manipulate time series models – indeed, even if you simply want to be able to read the software manuals – you have to be comfortable with the lag operator. The lag operator, L , is very simple: it “operates” on a series by lagging it. Hence $Ly_t = y_{t-1}$. Similarly, $L^2y_t = L(L(y_t)) = L(y_{t-1}) = y_{t-2}$, and so on. Typically we'll operate on a series not with the lag operator but with a **polynomial in the lag operator**. A lag operator polynomial of degree m is just a linear function of powers of L , up through the m -th power,

$$B(L) = b_0 + b_1L + b_2L^2 + \dots + b_mL^m.$$

To take a very simple example of a lag operator polynomial operating on a series, consider the m -th order lag operator polynomial L^m , for which

$$L^m y_t = y_{t-m}.$$

A well-known operator, the first-difference operator 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

$$B(L) = b_0 + b_1L + b_2L^2 + \dots = \sum_{i=0}^{\infty} b_i L^i.$$

Thus, for example, to denote an infinite distributed lag of current and past shocks we might write

$$B(L)\varepsilon_t = b_0\varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \dots = \sum_{i=0}^{\infty} b_i \varepsilon_{t-i}.$$

At first sight, infinite distributed lags may seem esoteric and of limited practical interest, because models with infinite distributed lags have infinitely many parameters (b_0, b_1, b_2, \dots) and therefore can't be estimated with a finite sample of data. On the contrary, and surprisingly, it turns out that models involving infinite distributed lags are central to time series modeling. Wold's theorem, to which we now turn, establishes that centrality.

14.4.2 Autoregressions

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.

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

$$y_t = \phi y_{t-1} + \varepsilon_t$$

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**.

$$\varepsilon_t \sim WN(0, \sigma^2).$$

In lag operator form, we write

$$(1 - \phi L)y_t = \varepsilon_t.$$

In Figure *** 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 covariance stationary. If we begin with the $AR(1)$ process,

$$y_t = \phi y_{t-1} + \varepsilon_t,$$

and substitute backward for lagged y's on the right side, we obtain

$$y_t = \varepsilon_t + \phi \varepsilon_{t-1} + \phi^2 \varepsilon_{t-2} + \dots$$

In lag operator form we write

$$y_t = \frac{1}{1 - \phi L} \varepsilon_t.$$

This moving average representation for y is convergent if and only if $|\phi| < 1$

; thus, $|\phi| < 1$ is the condition for covariance stationarity in the $AR(1)$ case. Equivalently, the condition for covariance stationarity is that the inverse of the root of the autoregressive lag operator polynomial be less than one in absolute value.

From the moving average representation of the covariance stationary $AR(1)$ process, we can compute the unconditional mean and variance,

$$\begin{aligned} E(y_t) &= E(\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \dots) \\ &= E(\varepsilon_t) + \phi E(\varepsilon_{t-1}) + \phi^2 E(\varepsilon_{t-2}) + \dots \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} var(y_t) &= var(\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \dots) \\ &= \sigma^2 + \phi^2\sigma^2 + \phi^4\sigma^2 + \dots \\ &= \sigma^2 \sum_{i=0}^{\infty} \phi^{2i} \\ &= \sigma^2 \frac{1}{1-\phi^2}. \end{aligned}$$

The conditional moments, in contrast, are

$$\begin{aligned} E(y_t|y_{t-1}) &= E(\phi y_{t-1} + \varepsilon_t|y_{t-1}) \\ &= \phi E(y_{t-1}|y_{t-1}) + E(\varepsilon_t|y_{t-1}) \\ &= \phi y_{t-1} + 0 \\ &= \phi y_{t-1} \end{aligned}$$

and

$$\begin{aligned} \text{var}(y_t|y_{t-1}) &= \text{var}((\phi y_{t-1} + \varepsilon_t)|y_{t-1}) \\ &= \phi^2 \text{var}(y_{t-1}|y_{t-1}) + \text{var}(\varepsilon_t|y_{t-1}) \\ &= 0 + \sigma^2 \\ &= \sigma^2. \end{aligned}$$

Note in particular that the simple way that the conditional mean adapts to the changing information set as the process evolves.

To find the autocovariances, we proceed as follows. The process is

$$y_t = \phi y_{t-1} + \varepsilon_t,$$

so that multiplying both sides of the equation by $y_{t-\tau}$ we obtain

$$y_t y_{t-\tau} = \phi y_{t-1} y_{t-\tau} + \varepsilon_t y_{t-\tau}.$$

For $\tau \geq 1$, taking expectations of both sides gives

$$\gamma(\tau) = \phi \gamma(\tau - 1).$$

This is called the **Yule-Walker equation**. It is a recursive equation; that is, given $\gamma(\tau)$, for any 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 Figure *** and *** 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,

$$\begin{aligned} p(\tau) &= \phi, \tau = 1 \\ &= 0, \tau > 1. \end{aligned}$$

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 *** and *** 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.

More on the Stability Condition in $AR(1)$

The key stability condition is $|\phi| < 1$

$$\begin{aligned} \text{Recall } y_t &= \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j} \\ \implies \text{var}(y_t) &= \sum_{j=0}^{\infty} \phi^{2j} \sigma^2 \end{aligned}$$

This is the sum of a geometric series. Hence:

$$\text{var}(y_t) = \frac{\sigma^2}{1 - \phi^2} \text{ if } |\phi| < 1$$

$$\text{var}(y_t) = \infty \text{ otherwise}$$

A More Complete Picture of $AR(1)$ Stability (On Your Own)

- Series y_t is persistent but eventually reverts to a fixed mean
- Shocks ε_t have persistent effects but eventually die out

Hint: Consider $y_t = \mu + \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$, $|\phi| < 1$

- Autocorrelations $\rho(\tau)$ nonzero but decay to zero
- Autocorrelations $\rho(\tau)$ depend on τ (of course) but not on time

Hint: Use back substitution to relate y_t and y_{t-2} . How does it compare to the relation between y_t and y_{t-1} when $|\phi| < 1$?

- Series y_t varies but not too extremely

Hint: Consider $\text{var}(y_t) = \frac{\sigma^2}{1 - \phi^2}$, $|\phi| < 1$

All of this makes for a nice, stable environment.

“Covariance stationarity”

14.4.3 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 ***. Because the roots are complex, the autocorrelation function oscillates, and because the roots are close to the unit circle, the oscillation damps slowly.

Finally, let's step back once again to consider in greater detail the precise way that finite-order autoregressive processes approximate the Wold repre-

¹⁵Note that complex roots can't occur in the $AR(1)$ case.

sentation. As always, the Wold representation is

$$y_t = B(L)\varepsilon_t,$$

where $B(L)$ is of infinite order. The moving average representation associated with the $AR(1)$ process is

$$y_t = \frac{1}{1 - \phi L}\varepsilon_t.$$

Thus, when we fit an $AR(1)$ model, we're using $\frac{1}{1 - \phi L}$, a rational polynomial with degenerate numerator polynomial (degree zero) and denominator polynomial of degree one, to approximate $B(L)$. The moving average representation associated with the $AR(1)$ process is of infinite order, as is the Wold representation, but it does not have infinitely many free coefficients. In fact, only one parameter, ϕ , underlies it.

The $AR(p)$ is an obvious generalization of the $AR(1)$ strategy for approximating the Wold representation. The moving average representation associated with the $AR(p)$ process is

$$y_t = \frac{1}{\Phi(L)}\varepsilon_t.$$

When we fit an $AR(p)$ model to approximate the Wold representation we're still using a rational polynomial with degenerate numerator polynomial (degree zero), but the denominator polynomial is of higher degree.

14.4.4 Alternative Approaches to Estimating Autoregressions

We can estimate autoregressions directly by OLS.

Alternatively, we can write the AR model as a regression on an intercept, with a serially correlated disturbance. We have

$$y_t = \mu + \varepsilon_t$$

$$\Phi(L)\varepsilon_t = v_t$$

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

We can estimate each model in identical fashion using nonlinear least squares. Eviews and other packages proceed in precisely that way.¹⁶

This framework – regression on a constant with serially correlated disturbances – has a number of attractive features. First, the mean of the process is the regression constant term.¹⁷ Second, it leads us naturally toward regression on more than just a constant, as other right-hand side variables can be added as desired.

Non-Zero Mean I (*AR*(1) Example): Regression on an Intercept and y_{t-1} , With White Noise Disturbances

$$(y_t - \mu) = \phi(y_{t-1} - \mu) + \varepsilon_t$$

$$\varepsilon_t \sim iidN(0, \sigma^2), |\phi| < 1$$

$$\implies y_t = c + \phi y_{t-1} + \varepsilon_t, \text{ where } c = \mu(1 - \phi)$$

Back-substitution reveals that:

$$y_t = \mu + \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$$

$$\implies E(y_t) = \mu$$

Non-Zero Mean II (*AR*(1) Example, Cont'd): Regression on an Intercept Alone, with AR(1) Disturbances

¹⁶That's why, for example, information on the number of iterations required for convergence is presented even for estimation of the autoregressive model.

¹⁷Hence the notation “ μ ” for the intercept.

$$y_t = \mu + \varepsilon_t$$

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

$$v_t \sim iidN(0, \sigma^2), |\phi| < 1$$

14.5 Serial Correlation in Time-Series Regression

Recall the full ideal conditions.

Here we deal with violation of the assumption that ***

14.5.1 Serial Correlation in Time-Series Regression

Consider:

$$\varepsilon \sim N(\underline{0}, \sigma^2 \Omega)$$

The FIC case is $\Omega = I$. When is $\Omega \neq I$?

We've already seen heteroskedasticity.

Now we consider “serial correlation” or “autocorrelation.”

$\rightarrow \varepsilon_t$ is correlated with $\varepsilon_{t-\tau} \leftarrow$

Can arise for many reasons, but they all boil down to:

The included X variables fail to capture all the dynamics in y .

– No additional explanation needed!

On Ω with Heteroskedasticity vs. Serial Correlation

With heteroskedasticity, ε_i is independent across i but not identically distributed across i (variance of ε_i varies with i):

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

With serial correlation, ε_t is correlated across t but unconditionally identically distributed across t :

$$\sigma^2 \Omega = \begin{pmatrix} \sigma^2 & \gamma(1) & \dots & \gamma(T-1) \\ \gamma(1) & \sigma^2 & \dots & \gamma(T-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(T-1) & \gamma(T-2) & \dots & \sigma^2 \end{pmatrix}$$

Consequences of Serial Correlation

OLS inefficient (no longer BLUE),
in finite samples and asymptotically

Standard errors biased and inconsistent. Hence t ratios do not have the t distribution in finite samples and do not have the $N(0, 1)$ distribution asymptotically

Does this sound familiar?

Detection

- Graphical autocorrelation diagnostics
 - Residual plot

- Scatterplot of e_t against $e_{t-\tau}$

14.5.2 Testing for 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.

- Formal autocorrelation tests and analyses
 - Durbin-Watson
 - Breusch-Godfrey
 - Residual correlogram

Liquor Sales Regression on Trend and Seasonals

Graphical Diagnostics - Residual Plot

Graphical Diagnostics - Scatterplot of e_t against e_{t-1}

The Durbin-Watson Test

Formal Tests and Analyses: Durbin-Watson (0.59!)

The Durbin-Watson test (discussed in Chapter 3) is the most popular.

Simple paradigm ($AR(1)$):

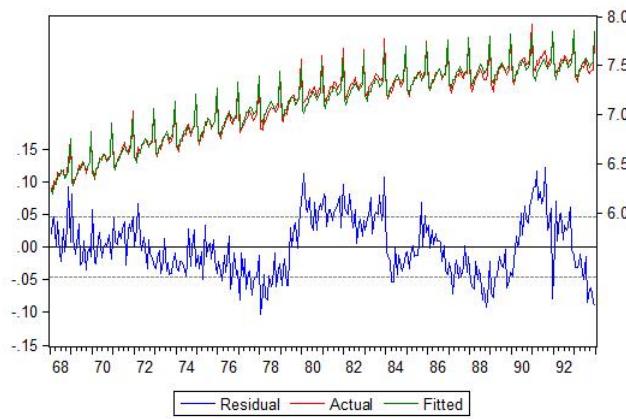
$$y_t = x'_t \beta + \varepsilon_t$$

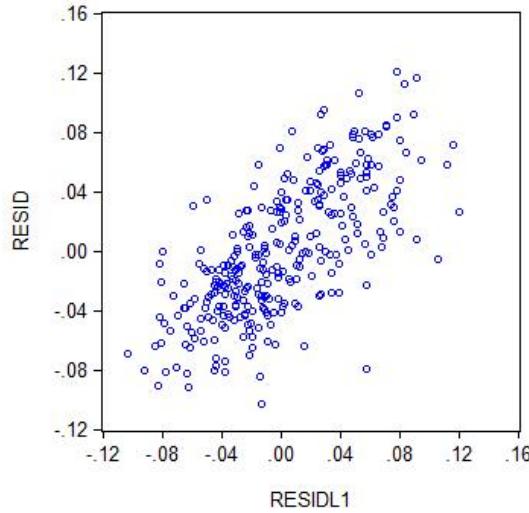
$$\varepsilon_t = \phi \varepsilon_{t-1} + v_t$$

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

Dependent Variable: LSALES
 Method: Least Squares
 Date: 10/13/12 Time: 12:32
 Sample: 1968M01 1993M12
 Included observations: 312

Variable	Coef.	Std. Err.	t-Statistic	Prob.
TIME	0.007656	0.000123	62.35882	0.0000
TIME2	-1.14E-05	3.56E-07	-32.06823	0.0000
D1	6.147456	0.012340	498.1699	0.0000
D2	6.088653	0.012353	492.8890	0.0000
D3	6.174127	0.012366	499.3008	0.0000
D4	6.175220	0.012378	498.8970	0.0000
D5	6.246086	0.012390	504.1398	0.0000
D6	6.250387	0.012401	504.0194	0.0000
D7	6.295979	0.012412	507.2402	0.0000
D8	6.268043	0.012423	504.5509	0.0000
D9	6.203832	0.012433	498.9630	0.0000
D10	6.229197	0.012444	500.5968	0.0000
D11	6.259770	0.012453	502.6602	0.0000
D12	6.580068	0.012463	527.9819	0.0000
R-squared	0.986111	Mean dependent var	7.112383	
Adjusted R-squared	0.985505	S.D. dependent var	0.379308	
S.E. of regression	0.045666	Akaike info criterion	-3.291086	
Sum squared resid	0.621448	Schwarz criterion	-3.123131	
Log likelihood	527.4094	Hannan-Quinn criter.	-3.223959	
Durbin-Watson stat	0.586187			





We want to test $H_0 : \phi = 0$ against $H_1 : \phi \neq 0$

Regress $y \rightarrow X$ and obtain the residuals e_t

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

Understanding the Durbin-Watson Statistic

$$\begin{aligned} DW &= \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T e_t^2} = \frac{\frac{1}{T} \sum_{t=2}^T (e_t - e_{t-1})^2}{\frac{1}{T} \sum_{t=1}^T e_t^2} \\ &= \frac{\frac{1}{T} \sum_{t=2}^T e_t^2 + \frac{1}{T} \sum_{t=2}^T e_{t-1}^2 - 2 \frac{1}{T} \sum_{t=2}^T e_t e_{t-1}}{\frac{1}{T} \sum_{t=1}^T e_t^2} \end{aligned}$$

Hence as $T \rightarrow \infty$:

$$DW \approx \frac{\sigma^2 + \sigma^2 - 2\text{cov}(e_t, e_{t-1})}{\sigma^2} = 2(1 - \underbrace{\text{corr}(e_t, e_{t-1})}_{\rho_e(1)})$$

$\implies DW \in [0, 4]$, $DW \rightarrow 2$ as $\phi \rightarrow 0$, and $DW \rightarrow 0$ as $\phi \rightarrow 1$

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.¹⁸ (See EPC ***) On both counts, we'd like more general and flexible approaches for diagnosing 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.

General $AR(p)$ environment:

$$y_t = x'_t \beta + \varepsilon_t$$

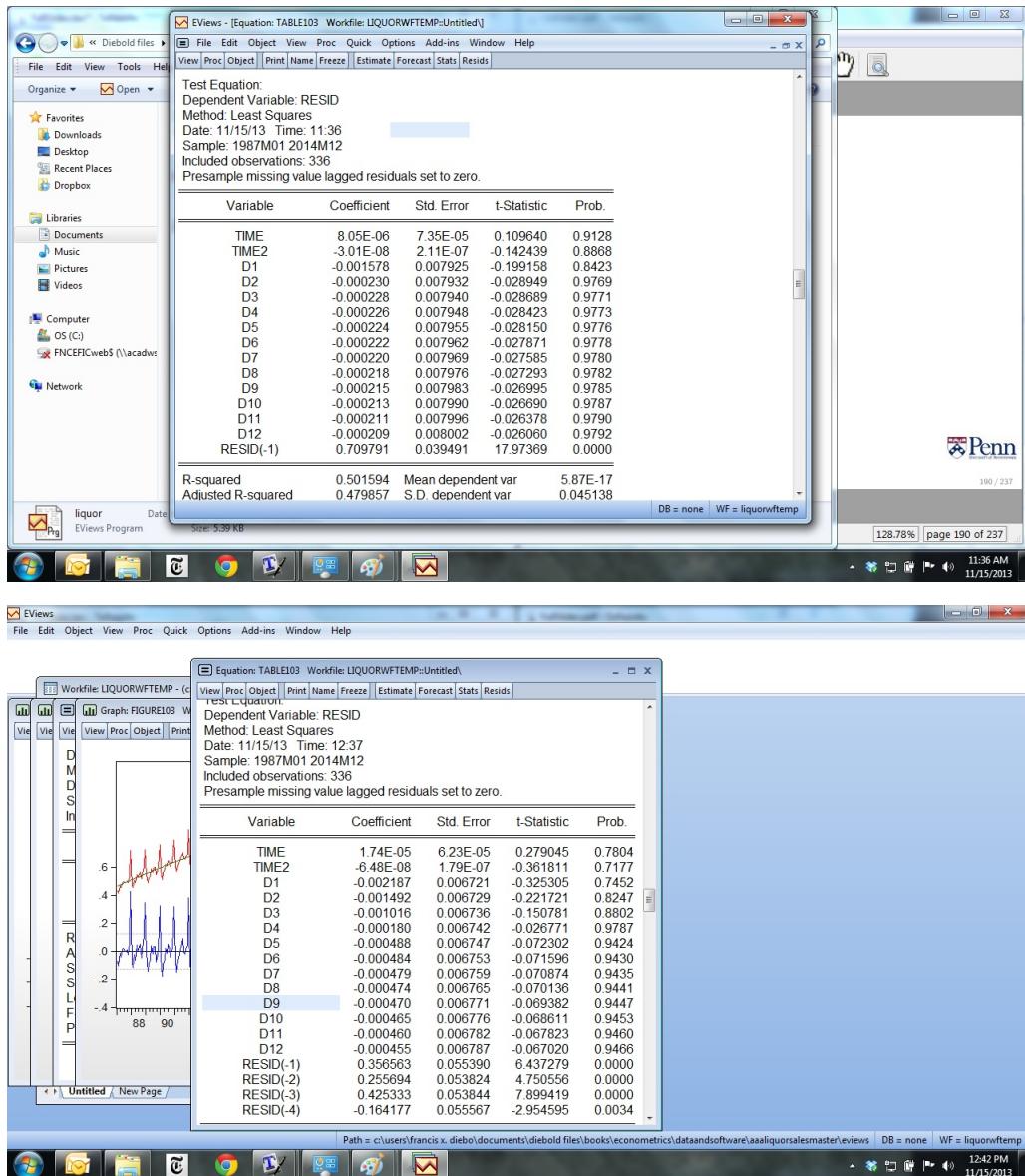
$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \dots + \phi_p \varepsilon_{t-p} + v_t$$

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

We want to test $H_0 : (\phi_1, \dots, \phi_p) = \underline{0}$ against $H_1 : (\phi_1, \dots, \phi_p) \neq \underline{0}$

- Regress $y_t \rightarrow x_t$ and obtain the residuals e_t
- Regress $e_t \rightarrow x_t, e_{t-1}, \dots, e_{t-p}$
- Examine TR^2 . In large samples $TR^2 \sim \chi_p^2$ under the null.

¹⁸Following 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.



Does this sound familiar?

BG for $AR(1)$ Disturbances

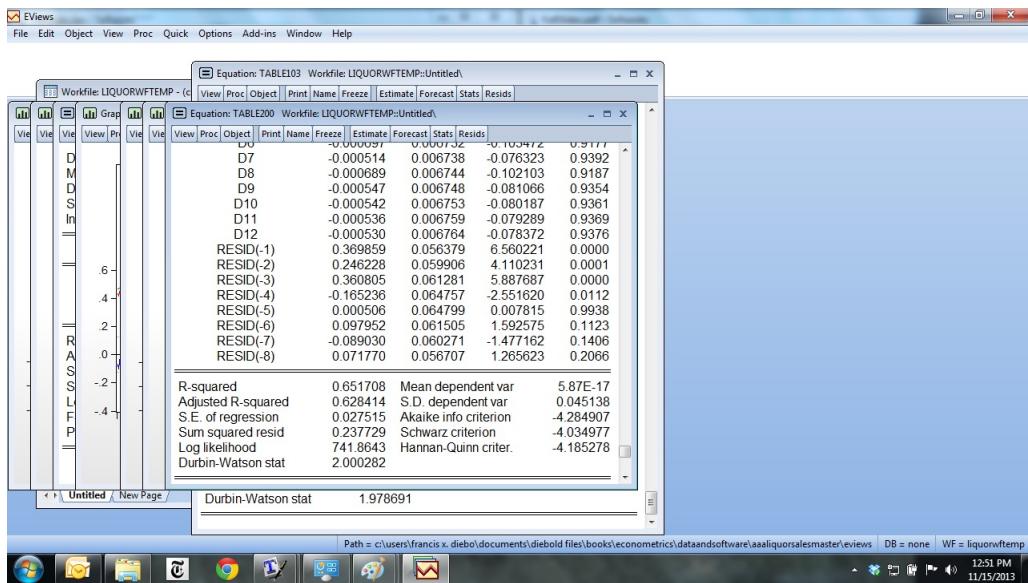
($TR^2 = 168.5$, $p = 0.0000$)

BG for $AR(4)$ Disturbances

($TR^2 = 216.7$, $p = 0.0000$)

BG for $AR(8)$ Disturbances

($TR^2 = 219.0$, $p = 0.0000$)



The Residual Correlogram

When we earlier introduced the correlogram in Chapter ***, 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.

$$\hat{\rho}_e(\tau) = \frac{\widehat{\text{cov}}(e_t, e_{t-\tau})}{\widehat{\text{var}}(e_t)} = \frac{\frac{1}{T} \sum_t e_t e_{t-\tau}}{\frac{1}{T} \sum_t e_t^2}$$

$\hat{\rho}_e(\tau)$ is the coefficient on $e_{t-\tau}$ in the regression

$$e_t \rightarrow c, e_{t-1}, \dots, e_{t-(\tau-1)}, e_{t-\tau}$$

Approximate 95% “Bartlett bands” under the $iid N$ null: $0 \pm \frac{2}{\sqrt{T}}$

Date: 10/14/12 Time: 18:32
 Sample: 1968M01 1993M12
 Included observations: 312

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
		1	0.700	0.700	154.34 0.000
		2	0.686	0.383	302.86 0.000
		3	0.725	0.369	469.36 0.000
		4	0.569	-0.141	572.36 0.000
		5	0.569	0.017	675.58 0.000
		6	0.577	0.093	782.19 0.000
		7	0.460	-0.078	850.06 0.000
		8	0.480	0.043	924.38 0.000
		9	0.466	0.030	994.46 0.000
		10	0.327	-0.188	1029.1 0.000
		11	0.364	0.019	1072.1 0.000
		12	0.355	0.089	1113.3 0.000
		13	0.225	-0.119	1129.9 0.000
		14	0.291	0.065	1157.8 0.000
		15	0.211	-0.119	1172.4 0.000
		16	0.138	-0.031	1178.7 0.000
		17	0.195	0.053	1191.4 0.000
		18	0.114	-0.027	1195.7 0.000
		19	0.055	-0.063	1196.7 0.000
		20	0.134	0.089	1202.7 0.000
		21	0.062	0.018	1204.0 0.000
		22	-0.006	-0.115	1204.0 0.000
		23	0.084	0.086	1206.4 0.000
		24	0.020	0.124	1206.8 0.000

$$Q_{BP} = T \sum_{\tau=1}^m \hat{\rho}_e^2(\tau) \sim \chi^2_{m-K} \text{ under iid } N$$

$$Q_{LB} = T(T+2) \sum_{\tau=1}^m \left(\frac{1}{T-\tau} \right) \hat{\rho}_e^2(\tau) \sim \chi^2_{m-K}$$

Residual Correlogram for Trend + Seasonal Model

14.5.3 Estimation with Serial Correlation

“Correcting for Autocorrelation”

14.5.4 Regression with Serially-Correlated Disturbances

GLS quasi-differencing, Cochrane-Orcutt iteration

- Generalized least squares
 - Transform the data such that the classical conditions hold
- Heteroskedasticity and autocorrelation consistent (HAC) s.e.’s
 - Use OLS, but calculate standard errors robustly

Recall Generalized Least Squares (*GLS*)

Consider the FIC except that we now let:

$$\varepsilon \sim N(\underline{0}, \sigma^2 \Omega)$$

The GLS estimator is:

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

Under the remaining full ideal conditions it is consistent, normally distributed with covariance matrix $\sigma^2 (X' \Omega^{-1} X)^{-1}$, and MVUE:

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

Infeasible *GLS*

(Illustrated in the Durbin-Watson *AR(1)* Environment)

$$y_t = x'_t \beta + \varepsilon_t \quad (1a)$$

$$\varepsilon_t = \phi \varepsilon_{t-1} + v_t \quad (1b)$$

$$v_t \sim iid N(0, \sigma^2) \quad (1c)$$

Suppose that you know ϕ . Then you could form:

$$\phi y_{t-1} = \phi x'_{t-1} \beta + \phi \varepsilon_{t-1} \quad (1a*)$$

$$\implies (y_t - \phi y_{t-1}) = (x'_t - \phi x'_{t-1}) \beta + (\varepsilon_t - \phi \varepsilon_{t-1}) \text{ (just (1a) - (1a*))}$$

$$\implies y_t = \phi y_{t-1} + x'_t \beta - x'_{t-1}(\phi \beta) + v_t$$

– Satisfies the classical conditions! Note the restriction.

14.5.5 Serially-Correlated Disturbances vs. Lagged Dependent Variables

Closely related. Inclusion of lagged dependent variables is the more general (and simple!) approach. *OLS* estimation.

So, two key closely-related regressions:

$$y_t \rightarrow x_t \text{ (with } AR(1) \text{ disturbances)}$$

$$y_t \rightarrow y_{t-1}, x_t, x_{t-1} \text{ (with } WN \text{ disturbances and a coef. restr.)}$$

Feasible *GLS*

- (1) Replace the unknown ϕ value with an estimate and run the OLS regression:

$$(y_t - \hat{\phi}y_{t-1}) \rightarrow (x'_t - \hat{\phi}x'_{t-1})$$

– Iterate if desired: $\hat{\beta}_1, \hat{\phi}_1, \hat{\beta}_2, \hat{\phi}_2, \dots$

- (2) Run the OLS Regression

$$y_t \rightarrow y_{t-1}, x_t, x_{t-1}$$

subject to the constraint noted earlier (or not)

- Generalizes trivially to $AR(p)$: $y_t \rightarrow y_{t-1}, \dots, y_{t-p}, x_t, x_{t-1}, \dots, x_{t-p}$
(Select p using the usual *AIC*, *SIC*, etc.)

Trend + Seasonal Model with $AR(4)$ Disturbances

Trend + Seasonal Model with $AR(4)$ Disturbances

Residual Plot

Trend + Seasonal Model with $AR(4)$ Disturbances

Residual Correlogram

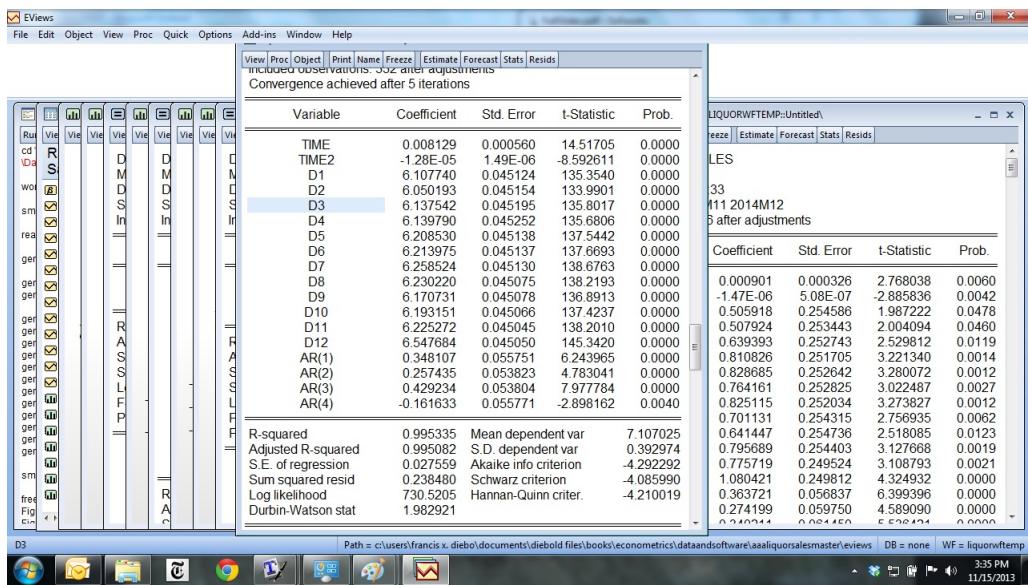


Figure 14.1: ***. ***.

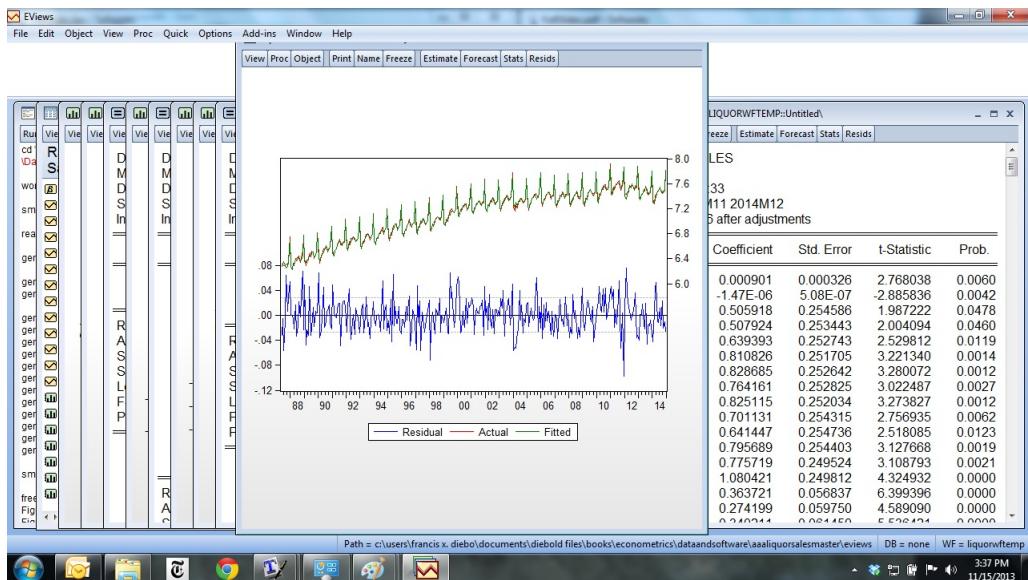


Figure 14.2: ***. ***.

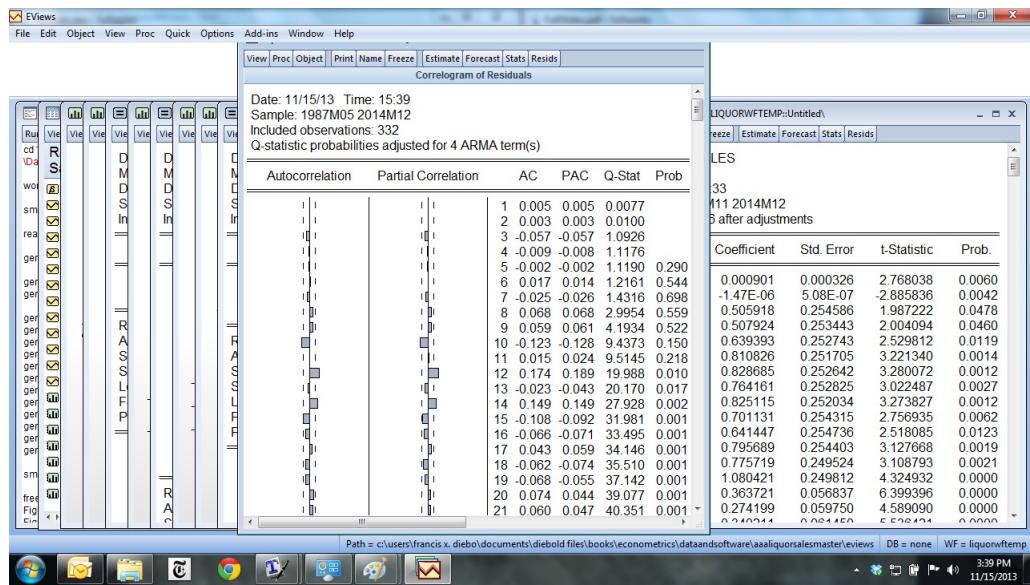
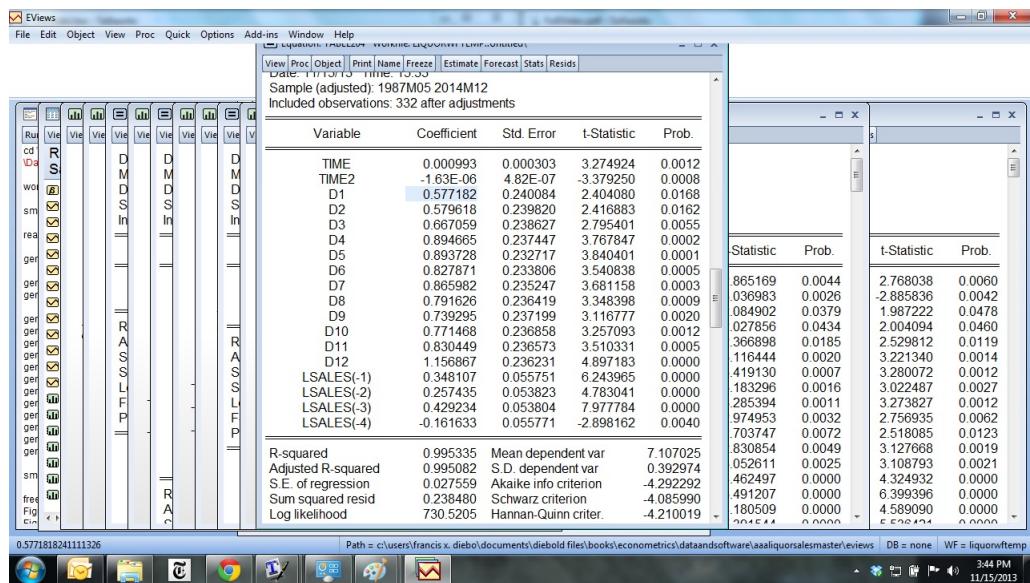


Figure 14.3: ***. ***.



Trend + Seasonal Model with Four Lags of Dep. Var.

How Did we Arrive at $AR(4)$ Dynamics?

Everything points there:

- Supported by original trend + seasonal residual correlogram
 - Supported by DW
 - Supported by BG
 - Supported by SIC pattern:

$$AR(1) = -3.797$$

$$AR(2) = -3.941$$

$$AR(3) = -4.080$$

$$AR(4) = -4.086$$

$$AR(5) = -4.071$$

$$AR(6) = -4.058$$

$$AR(7) = -4.057$$

$$AR(8) = -4.040$$

Heteroskedasticity-and-Autocorrelation Consistent (HAC) Standard Errors

Using advanced methods, one can obtain consistent standard errors (if not an efficient $\hat{\beta}$), under minimal assumptions

- “HAC standard errors”
- “Robust standard errors”
- “Newey-West standard errors”
- $\hat{\beta}$ remains unchanged at its OLS value. Is that a problem?

Trend + Seasonal Model with HAC Standard Errors

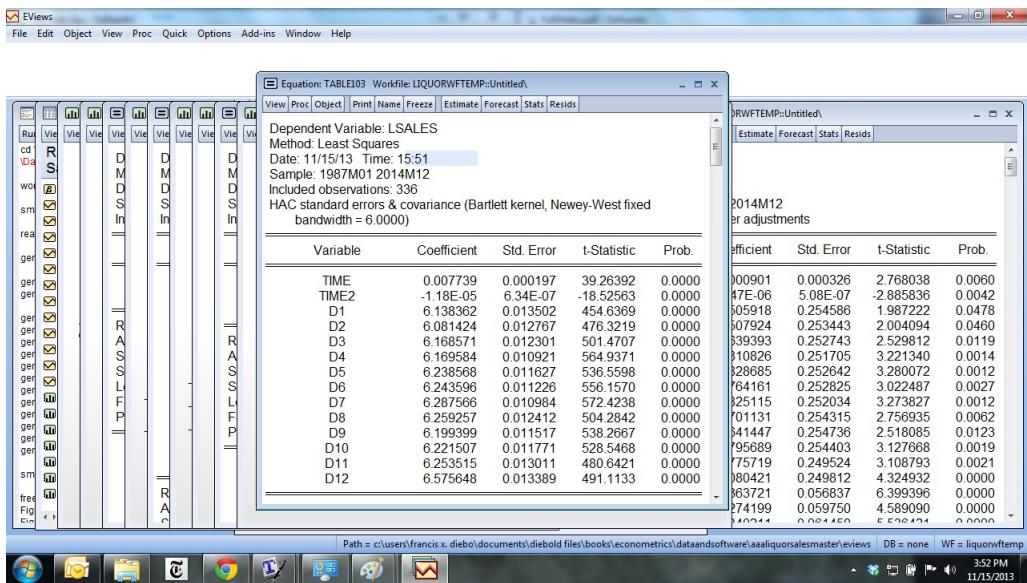


Figure 14.4: ***. ***.

14.5.6 A Full Model of Liquor Sales

We'll model monthly U.S. liquor sales. We graphed a short span of the series in Chapter *** and noted its pronounced seasonality – sales skyrocket during the Christmas season. In Figure ***, we show a longer history of liquor sales, 1968.01 - 1993.12. In Figure *** we show log liquor sales; we take logs to stabilize the variance, which grows over time.¹⁹ The variance of log liquor sales is more stable, and it's the series for which we'll build models.²⁰

Liquor sales dynamics also feature prominent trend and cyclical effects. Liquor sales trend upward, and the trend appears nonlinear in spite of the fact that we're working in logs. To handle the nonlinear trend, we adopt a quadratic trend model (in logs). The estimation results are in Table 1. The residual plot (Figure ***) shows that the fitted trend increases at a decreasing rate; both the linear and quadratic terms are highly significant. The adjusted R^2 is 89%, reflecting the fact that trend is responsible for a large part of the

¹⁹The 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.

²⁰From this point onward, for brevity we'll simply refer to “liquor sales,” but remember that we've taken logs.

variation in liquor sales. The standard error of the regression is .125; it's an estimate of the standard deviation of the error we'd expect to make in forecasting liquor sales if we accounted for trend but ignored seasonality and serial correlation. The Durbin-Watson statistic provides no evidence against the hypothesis that the regression disturbance is white noise.

The residual plot, however, shows obvious residual seasonality. The Durbin-Watson statistic missed it, evidently because it's not designed to have power against seasonal dynamics.²¹ The residual plot also suggests that there may be a cycle in the residual, although it's hard to tell (hard for the Durbin-Watson statistic as well), because the pervasive seasonality swamps the picture and makes it hard to infer much of anything.

The residual correlogram (Table 2) and its graph (Figure ***) confirm the importance of the neglected seasonality. The residual sample autocorrelation function has large spikes, far exceeding the Bartlett bands, at the seasonal displacements, 12, 24, and 36. It indicates some cyclical dynamics as well; apart from the seasonal spikes, the residual sample autocorrelation and partial autocorrelation functions oscillate, and the Ljung-Box statistic rejects the white noise null hypothesis even at very small, non-seasonal, displacements. In Table 3 we show the results of regression on quadratic trend and a full set of seasonal dummies. The quadratic trend remains highly significant. The adjusted R^2 rises to 99%, and the standard error of the regression falls to .046, which is an estimate of the standard deviation of the forecast error we expect to make if we account for trend and seasonality but ignore serial correlation. The Durbin-Watson statistic, however, has greater ability to detect serial correlation now that the residual seasonality has been accounted for, and it sounds a loud alarm.

The residual plot of Figure *** shows no seasonality, as that's now picked

²¹Recall that the Durbin-Watson test is designed to detect simple $AR(1)$ dynamics. It also has the ability to detect other sorts of dynamics, but evidently not those relevant to the present application, which are very different from a simple $AR(1)$.

up by the model, but it confirms the Durbin-Watson's warning of serial correlation. The residuals are highly persistent, and hence predictable. We show the residual correlogram in tabular and graphical form in Table *** and Figure ***. The residual sample autocorrelations oscillate and decay slowly, and they exceed the Bartlett standard errors throughout. The Ljung-Box test strongly rejects the white noise null at all displacements. Finally, the residual sample partial autocorrelations cut off at displacement 3. All of this suggests that an $AR(3)$ would provide a good approximation to the disturbance's Wold representation.

In Table 5, then, we report the results of estimating a liquor sales model with quadratic trend, seasonal dummies, and $AR(3)$ disturbances. The R^2 is now 100%, and the Durbin-Watson is fine. One inverse root of the $AR(3)$ disturbance process is estimated to be real and close to the unit circle (.95), and the other two inverse roots are a complex conjugate pair farther from the unit circle. The standard error of this regression is an estimate of the standard deviation of the forecast error we'd expect to make after modeling the residual serial correlation, as we've now done; that is, it's an estimate of the standard deviation of v .²² It's a very small .027, roughly half that obtained when we ignored serial correlation.

We show the residual plot in Figure *** and the residual correlogram in Table *** and Figure ***. The residual plot reveals no patterns; instead, the residuals look like white noise, as they should. The residual sample autocorrelations and partial autocorrelations display no patterns and are mostly inside the Bartlett bands. The Ljung-Box statistics also look good for small and moderate displacements, although their p -values decrease for longer displacements.

All things considered, the quadratic trend, seasonal dummy, $AR(3)$ specification seems tentatively adequate. We also perform a number of additional

²²Recall that v is the innovation that drives the AR process for the regression disturbance, ε .

checks. In Figure ***, we show a histogram and normality test applied to the residuals. The histogram looks symmetric, as confirmed by the skewness near zero. The residual kurtosis is a bit higher than three and causes Jarque-Bera test to reject the normality hypothesis with a p -value of .02, but the residuals nevertheless appear to be fairly well approximated by a normal distribution, even if they may have slightly fatter tails.

14.6 Vector Autoregression

Multivariate: Vector Autoregressions

Univariate $AR(p)$:

$$y_t = \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t$$

$$y_t = \phi_1 L y_t + \dots + \phi_p L^p y_t + \varepsilon_t$$

$$(I - \phi_1 L - \dots - \phi_p L^p) y_t = \varepsilon_t$$

$$\phi(L) y_t = \varepsilon_t$$

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

But what if we have more than 1 “ y ” variable?

Cross-variable interactions? Leads? Lags? Causality?

N -Variable $VAR(p)$

$$y_{1t} = \phi_{11}^1 y_{1,t-1} + \dots + \phi_{1N}^1 y_{N,t-1} + \dots + \phi_{11}^p y_{1,t-p} + \dots + \phi_{1N}^p y_{N,t-p} + \varepsilon_{1t}$$

⋮

$$y_{Nt} = \phi_{N1}^1 y_{1,t-1} + \dots + \phi_{NN}^1 y_{N,t-1} + \dots + \phi_{N1}^p y_{1,t-p} + \dots + \phi_{NN}^p y_{N,t-p} + \varepsilon_{Nt}$$

$$\begin{pmatrix} y_{1t} \\ \vdots \\ y_{Nt} \end{pmatrix} = \begin{pmatrix} \phi_{11}^1 & \dots & \phi_{1N}^1 \\ \vdots & & \vdots \\ \phi_{N1}^1 & \dots & \phi_{NN}^1 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ \vdots \\ y_{N,t-1} \end{pmatrix} + \dots + \begin{pmatrix} \phi_{11}^p & \dots & \phi_{1N}^p \\ \vdots & & \vdots \\ \phi_{N1}^p & \dots & \phi_{NN}^p \end{pmatrix} \begin{pmatrix} y_{1,t-p} \\ \vdots \\ y_{N,t-p} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{Nt} \end{pmatrix}$$

$$y_t = \Phi_1 y_{t-1} + \dots + \Phi_p y_{t-p} + \varepsilon_t$$

$$y_t = \Phi_1 L y_t + \dots + \Phi_p L^p y_t + \varepsilon_t$$

$$(I - \Phi_1 L - \dots - \Phi_p L^p) y_t = \varepsilon_t$$

$$\Phi(L) y_t = \varepsilon_t$$

$$\varepsilon_t \sim iid(0, \Sigma)$$

Estimation and Selection

Estimation: Equation-by-equation OLS

Selection: AIC, SIC

$$AIC = \frac{-2\ln L}{T} + \frac{2K}{T}$$

$$SIC = \frac{-2\ln L}{T} + \frac{K\ln T}{T}$$

The Cross-Correlation Function

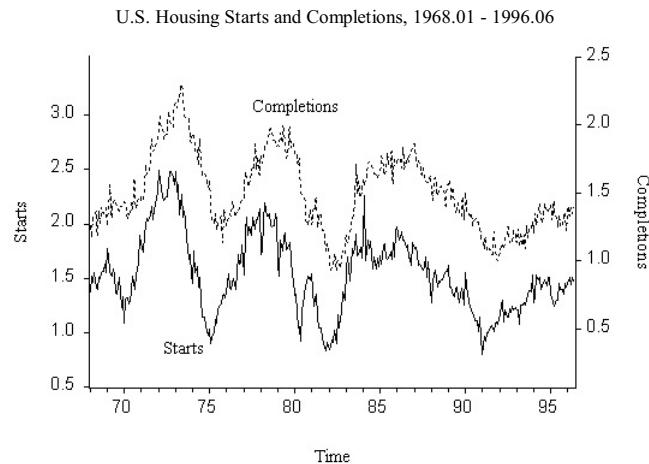
Recall the univariate autocorrelation function:

$$\rho_y(\tau) = corr(y_t, y_{t-\tau})$$

In multivariate environments we also have

the cross-correlation function:

$$\rho_{yx}(\tau) = corr(y_t, x_{t-\tau})$$



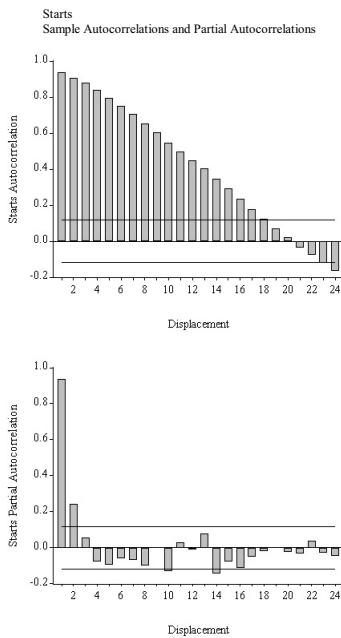
Notes to figure: The left scale is starts, and the right scale is completions.

Starts Correlogram

Sample: 1968:01 1991:12

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



Granger-Sims Causality

Bivariate case:

y_i Granger-Sims causes y_j if
 y_i has predictive content for y_j ,
over and above the past history of y_j .

Testing:

Are lags of y_i significant in the y_j equation?

Impulse-Response Functions in $AR(1)$ Case

$$y_t = \phi y_{t-1} + \varepsilon_t, \varepsilon_t \sim iid(0, \sigma^2)$$

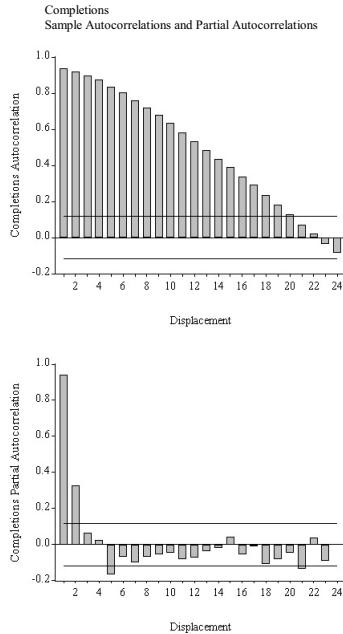
$$\implies y_t = B(L)\varepsilon_t = \varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \dots$$

Completions Correlogram

Sample: 1968:01 1991:12

Included observations: 288

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	0.939	0.939	0.059	256.61	0.000
2	0.920	0.328	0.059	504.05	0.000
3	0.896	0.066	0.059	739.19	0.000
4	0.874	0.023	0.059	963.73	0.000
5	0.834	-0.165	0.059	1168.9	0.000
6	0.802	-0.067	0.059	1359.2	0.000
7	0.761	-0.100	0.059	1531.2	0.000
8	0.721	-0.070	0.059	1686.1	0.000
9	0.677	-0.055	0.059	1823.2	0.000
10	0.633	-0.047	0.059	1943.7	0.000
11	0.583	-0.080	0.059	2046.3	0.000
12	0.533	-0.073	0.059	2132.2	0.000
13	0.483	-0.038	0.059	2203.2	0.000
14	0.434	-0.020	0.059	2260.6	0.000
15	0.390	0.041	0.059	2307.0	0.000
16	0.337	-0.057	0.059	2341.9	0.000
17	0.290	-0.008	0.059	2367.9	0.000
18	0.234	-0.109	0.059	2384.8	0.000
19	0.181	-0.082	0.059	2395.0	0.000
20	0.128	-0.047	0.059	2400.1	0.000
21	0.068	-0.133	0.059	2401.6	0.000
22	0.020	0.037	0.059	2401.7	0.000
23	-0.038	-0.092	0.059	2402.2	0.000
24	-0.087	-0.003	0.059	2404.6	0.000



$$= \varepsilon_t + \phi \varepsilon_{t-1} + \phi^2 \varepsilon_{t-2} + \dots$$

IRF is $\{1, \phi, \phi^2, \dots\}$ “dynamic response to a unit shock in ε ”

Alternatively write $\varepsilon_t = \sigma v_t$, $v_t \sim iid(0, 1)$

$$\implies y_t = \sigma v_t + (\phi \sigma) v_{t-1} + (\phi^2 \sigma) v_{t-2} + \dots$$

IRF is $\{\sigma, \phi\sigma, \phi^2\sigma, \dots\}$ “dynamic response to a one- σ shock in ε ”

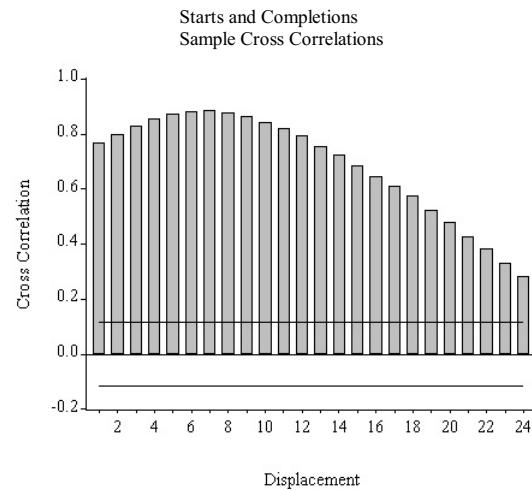
Impulse-Response Functions in $VAR(p)$ Case

$$y_t = \Phi y_{t-1} + \varepsilon_t, \varepsilon_t \sim iid(0, \Sigma)$$

$$\implies y_t = B(L)\varepsilon_t = \varepsilon_t + B_1\varepsilon_{t-1} + B_2\varepsilon_{t-2} + \dots$$

$$= \varepsilon_t + \Phi \varepsilon_{t-1} + \Phi^2 \varepsilon_{t-2} + \dots$$

But we need orthogonal shocks. Why?

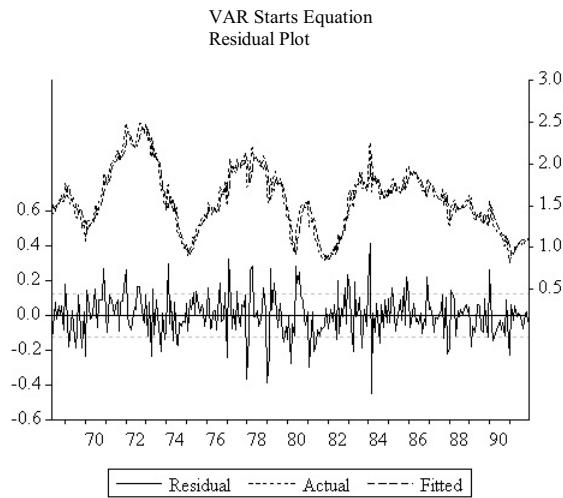


Notes to figure: We graph the sample correlation between completions at time t and starts at time $t-i$, $i = 1, 2, \dots, 24$.

VAR Starts Equation

LS // Dependent Variable is STARTS
 Sample(adjusted): 1968:05 1991:12
 Included observations: 284 after adjusting endpoints

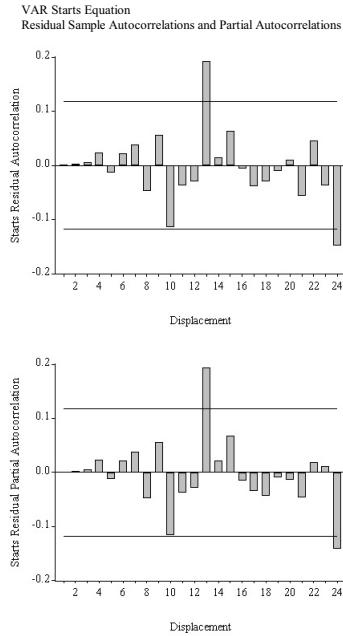
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.146871	0.044235	3.320264	0.0010
STARTS(-1)	0.659939	0.061242	10.77587	0.0000
STARTS(-2)	0.229632	0.072724	3.157587	0.0018
STARTS(-3)	0.142859	0.072655	1.966281	0.0503
STARTS(-4)	0.007806	0.066032	0.118217	0.9060
COMPS(-1)	0.031611	0.102712	0.307759	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	



VAR Starts Equation
Residual Correlogram

Sample: 1968:01 1991:12
Included observations: 284

	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



So write $\varepsilon_t = Pv_t$, $v_t \sim iid(0, I)$, where P is Cholesky factor of Σ

$$\implies y_t = Pv_t + (\Phi P)v_{t-1} + (\Phi^2 P)v_{t-2} + \dots$$

ij 'th IRF is the sequence of ij 'th elements of $\{P, \Phi P, \Phi^2 P, \dots\}$ “Dynamic response of y_i to a one- σ shock in ε_j ”

We continue working with autoregressions, but we move to a multivariate environment, which raises the possibility of cross-variable interaction.

14.6.1 Distributed Lag Models

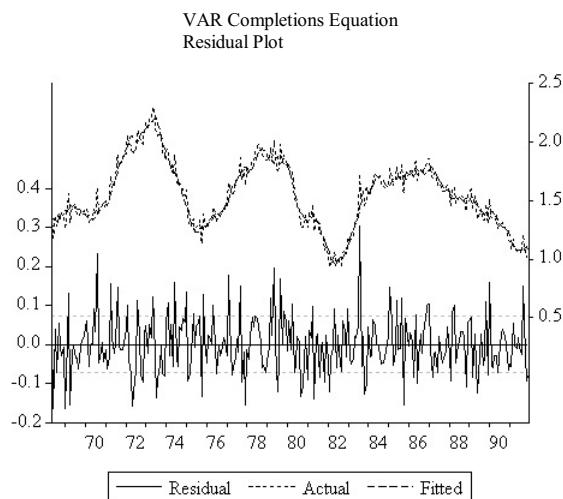
The **distributed lag model** is

$$y_t = \beta_0 + \sum_{i=1}^{N_x} \delta_i x_{t-i} + \varepsilon_t.$$

VAR Completions Equation

LS // Dependent Variable is COMPS
 Sample(adjusted): 1968:05 1991:12
 Included observations: 284 after adjusting endpoints

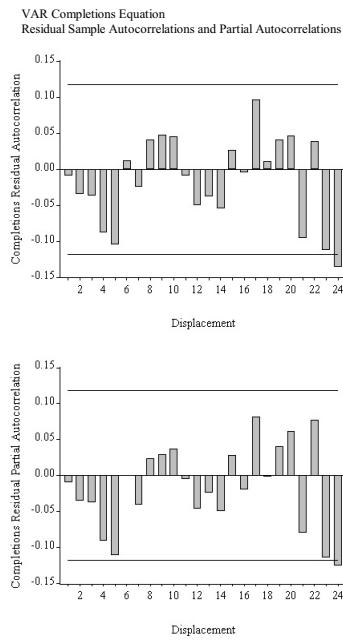
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.045347	0.025794	1.758045	0.0799
STARTS(-1)	0.074724	0.035711	2.092461	0.0373
STARTS(-2)	0.040047	0.042406	0.944377	0.3458
STARTS(-3)	0.047145	0.042366	1.112805	0.2668
STARTS(-4)	0.082331	0.038504	2.138238	0.0334
COMPS(-1)	0.236774	0.059893	3.953313	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	



VAR Completions Equation
 Residual Correlogram

Sample: 1968:01 1991:12
 Included observations: 284

	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	10.318	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



We say that y depends on a distributed lag of past x 's. The coefficients on the lagged x 's are called lag weights, and their pattern is called the lag distribution.

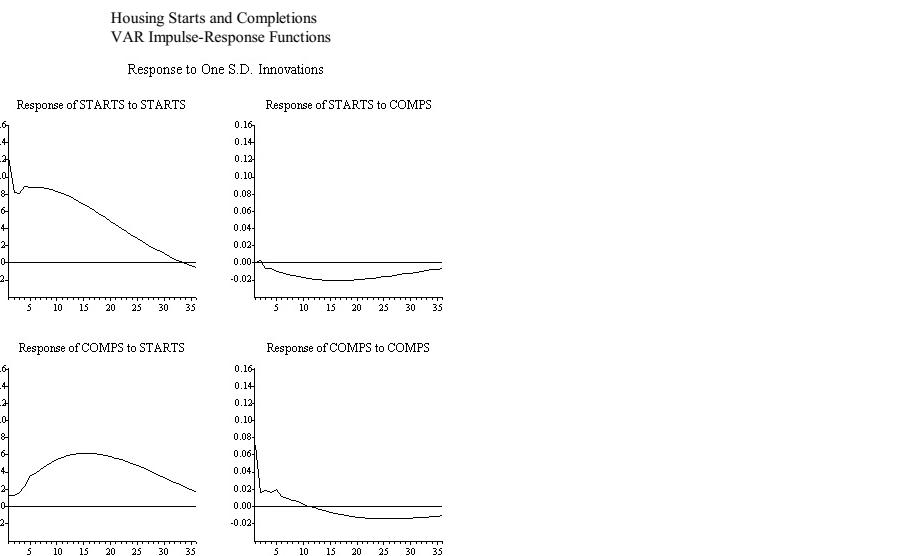
One way to estimate a distributed lag model is simply to include all N_x lags of x in the regression, which can be estimated by least squares in the usual way. In many situations, however, N_x might be quite a large number, in which case we'd have to use many degrees of freedom to estimate the model, violating the parsimony principle. Often we can recover many of those degrees of freedom without seriously worsening the model's fit by constraining the lag weights to lie on a low-order polynomial. Such **polynomial distributed lags** promote smoothness in the lag distribution and may lead to sophisticatedly simple models.

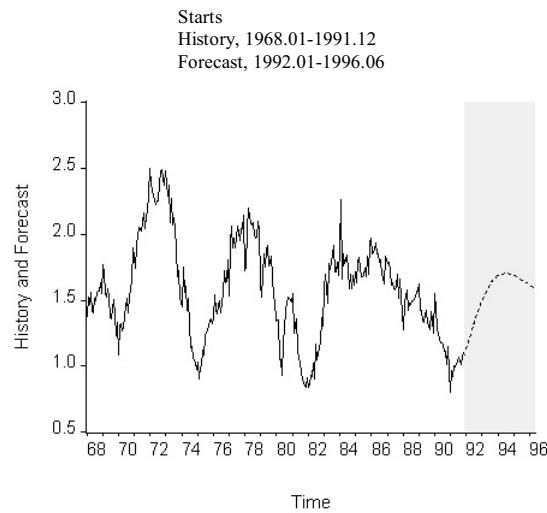
Polynomial distributed lag models are estimated by minimizing the sum of squared residuals in the usual way, subject to the constraint that the

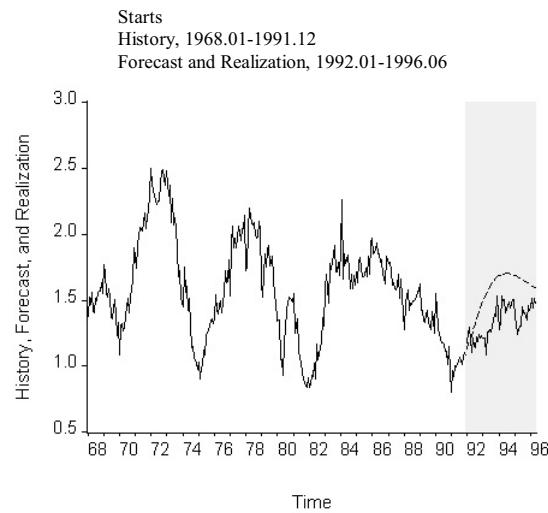
Housing Starts and Completions
Causality Tests

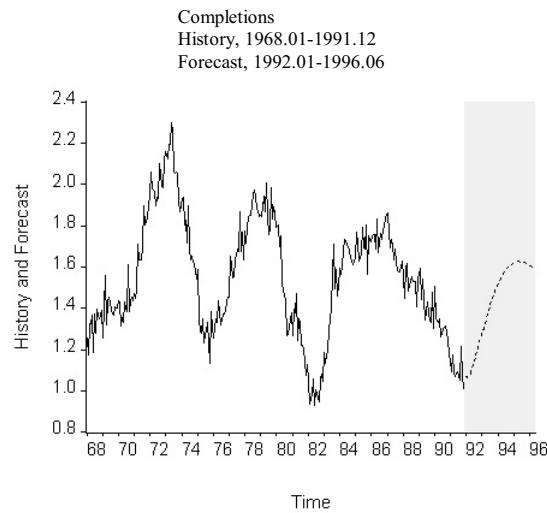
Sample: 1968:01 1991:12
Lags: 4
Obs: 284

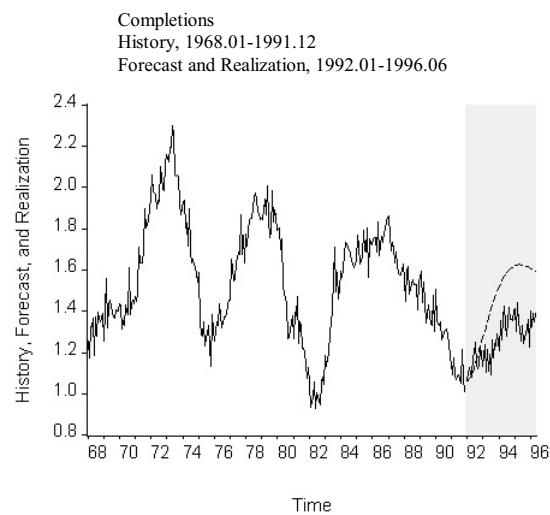
Null Hypothesis:	F-Statistic	Probability
STARTS does not Cause COMPS	26.2658	0.00000
COMPS does not Cause STARTS	2.23876	0.06511











lag weights follow a low-order polynomial whose degree must be specified. Suppose, for example, that we constrain the lag weights to follow a second-degree polynomial. Then we find the parameter estimates by solving the problem

$$\min_{\beta_0, \delta_i} \sum_{t=N_x+1}^T \left[y_t - \beta_0 - \sum_{i=1}^{N_x} \delta_i x_{t-i} \right]^2,$$

subject to

$$\delta_i = P(i) = a + bi + ci^2, \quad i = 1, \dots, N_x.$$

This converts the estimation problem from one of estimating $1 + N_x$ parameters, $\beta_0, \delta_1, \dots, \delta_{N_x}$, to one of estimating four parameters, β_0 , a , b and c . Sometimes additional constraints are imposed on the shape of the polynomial, such as $P(N_x) = 0$, which enforces the idea that the dynamics have been exhausted by lag N_x .

14.6.2 Regressions with Lagged Dependent Variables, and Regressions with *AR* Disturbances

There's something missing in distributed lag models of the form

$$y_t = \beta_0 + \sum_{i=1}^{N_x} \delta_i x_{t-i} + \varepsilon_t.$$

A multivariate model (in this case, a regression model) should relate the current value y to its own past and to the past of x . But as presently written, we've left out the past of y ! Even in distributed lag models, we always want to allow for the presence of the usual univariate dynamics. Put differently, the included regressors may not capture all the dynamics in y , which we need to model one way or another. Thus, for example, a preferable model includes

lags of the dependent variable,

$$y_t = \beta_0 + \sum_{i=1}^{N_y} \alpha_i y_{t-i} + \sum_{j=1}^{N_x} \delta_j x_{t-j} + \varepsilon_t.$$

This model, a **distributed lag regression model with lagged dependent variables**, is closely related to, but not exactly the same as, the rational distributed lag model introduced earlier. (Why?) You can think of it as arising by beginning with a univariate autoregressive model for y , and then introducing additional explanatory variables. If the lagged y 's don't play a role, as assessed with the usual tests, we can always delete them, but we never want to eliminate from the outset the possibility that lagged dependent variables play a role. Lagged dependent variables absorb residual serial correlation.

Alternatively, we can capture own-variable dynamics in distributed-lag regression models by using a **distributed-lag regression model with AR disturbances**. Recall that our $AR(p)$ models are equivalent to regression models, with only a constant regressor, and with $AR(p)$ disturbances,

$$\begin{aligned} y_t &= \beta_0 + \varepsilon_t \\ \varepsilon_t &= \frac{1}{\Phi(L)} v_t \\ v_t &\sim WN(0, \sigma^2). \end{aligned}$$

We want to begin with the univariate model as a baseline, and then generalize it to allow for multivariate interaction, resulting in models such as

$$\begin{aligned} y_t &= \beta_0 + \sum_{i=1}^{N_x} \delta_i x_{t-i} + \varepsilon_t \\ \varepsilon_t &= \frac{1}{\Phi(L)} v_t \\ v_t &\sim WN(0, \sigma^2). \end{aligned}$$

Regressions with *AR* disturbances make clear that regression (a statistical and econometric tool with a long tradition) and the *AR* model of time-series dynamics (a more recent innovation) are not at all competitors; rather, when used appropriately they can be highly complementary.

It turns out that the distributed-lag regression model with autoregressive disturbances – a great workhorse in econometrics – is a special case of the more general model with lags of both y and x and white noise disturbances. To see this, let's take a simple example:

$$y_t = \beta_0 + \beta_1 x_{t-1} + \varepsilon_t$$

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

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

In lag operator notation, we write the *AR*(1) regression disturbance as

$$(1 - \phi L)\varepsilon_t = v_t,$$

or

$$\varepsilon_t = \frac{1}{(1 - \phi L)} v_t.$$

Thus we can rewrite the regression model as

$$y_t = \beta_0 + \beta_1 x_{t-1} + \frac{1}{(1 - \phi L)} v_t.$$

Now multiply both sides by $(1 - \phi L)$ to get

$$(1 - \phi L)y_t = (1 - \phi)\beta_0 + \beta_1(1 - \phi)Lx_{t-1} + v_t,$$

or

$$y_t = \phi y_{t-1} + (1 - \phi)\beta_0 + \beta_1 x_{t-1} - \phi\beta_1 x_{t-2} + v_t.$$

Thus a model with one lag of x on the right and *AR*(1) disturbances is equiv-

alent to a model with y_{t-1} , x_{t-1} , and x_{t-2} on the right-hand side and white noise errors, *subject to the restriction* that the coefficient on the second lag of x_{t-2} is the negative of the product of the coefficients on y_{t-1} and x_{t-1} . Thus, distributed lag regressions with lagged dependent variables are more general than distributed lag regressions with dynamic disturbances. In practice, the important thing is to allow for own-variable dynamics *somewhat*, in order to account for dynamics in y not explained by the right-hand-side variables. Whether we do so by including lagged dependent variables or by allowing for *ARMA* disturbances can occasionally be important, but usually it's a comparatively minor issue.

14.6.3 Vector Autoregressions

A univariate autoregression involves one variable. In a univariate autoregression of order p , we regress a variable on p lags of itself. In contrast, a multivariate autoregression – that is, a vector autoregression, or *VAR* – involves N variables. In an N -variable **vector autoregression of order p** , or *VAR*(p), we estimate N different equations. In each equation, we regress the relevant left-hand-side variable on p lags of itself, *and p lags of every other variable*.²³ 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

$$y_{1,t} = \phi_{11}y_{1,t-1} + \phi_{12}y_{2,t-1} + \varepsilon_{1,t}$$

²³Trends, seasonals, and other exogenous variables may also be included, as long as they're all included in every equation.

$$y_{2,t} = \phi_{21}y_{1,t-1} + \phi_{22}y_{2,t-1} + \varepsilon_{2,t}.$$

Each variable depends on one lag of the other variable in addition to one lag of itself; that's one obvious source of multivariate interaction captured by the *VAR*. 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. 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.²⁴

When fitting *VARs* to data, we use the Schwarz and Akaike criteria, just as in the univariate case. The formulas differ, however, because we're now working with a multivariate system of equations rather than a single equation. To get an *AIC* or *SIC* value for a *VAR* system, we could add up the equation-

²⁴For an exposition of seemingly unrelated regression, see Pindyck and Rubinfeld (1997).

by-equation *AICs* or *SICs*, but unfortunately, doing so is appropriate only if the innovations are uncorrelated across equations, which is a very special and unusual situation. Instead, explicitly multivariate versions of the *AIC* and *SIC* – and more advanced formulas – are required that account for cross-equation innovation correlation. It's beyond the scope of this book to derive and present those formulas, because they involve unavoidable use of matrix algebra, but fortunately we don't need to. They're pre-programmed in many computer packages, and we interpret the *AIC* and *SIC* values computed for *VARs* of various orders in exactly the same way as in the univariate case: we select that order p such that the *AIC* or *SIC* is minimized.

14.6.4 Predictive Causality

There's an important statistical notion of causality that's intimately related to forecasting and naturally introduced in the context of *VARs*. It is based on two key principles: first, cause should occur before effect, and second, a causal series should contain information useful for forecasting that is not available in the other series (including the past history of the variable being forecast). In the unrestricted *VARs* that we've studied thus far, *everything* causes everything else, because lags of every variable appear on the right of every equation. Cause precedes effect because the right-hand-side variables are lagged, and each variable is useful in forecasting every other variable.

We stress from the outset that the notion of **predictive causality** contains little if any information about causality in the philosophical sense. Rather, the statement " y_i causes y_j " is just shorthand for the more precise, but long-winded, statement, " y_i contains useful information for predicting y_j (in the linear least squares sense), over and above the past histories of the other variables in the system." To save space, we simply say that y_i causes y_j .

To understand what predictive causality means in the context of a $VAR(p)$,

consider the j -th equation of the N -equation system, which has y_j on the left and p lags of each of the N variables on the right. If y_i causes y_j , then at least one of the lags of y_i that appear on the right side of the y_j equation must have a nonzero coefficient.

It's also useful to consider the opposite situation, in which y_i does not cause y_j . In that case, all of the lags of that y_i that appear on the right side of the y_j equation must have zero coefficients.²⁵ 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 ?.” 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,

²⁵Note 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 .

we might want to *impose* it, by omitting certain lags of certain variables from certain equations.

Various types of causality hypotheses are sometimes entertained. In any equation (the j -th, say), we've already discussed testing the simple noncausality hypothesis that:

- (a) No lags of variable i aid in one-step-ahead prediction of variable j .

We can broaden the idea, however. Sometimes we test stronger noncausality hypotheses such as:

- (b) No lags of a *set* of other variables aid in one-step-ahead prediction of variable j .
- (b) No lags of *any other variables* aid in one-step-ahead prediction of variable j .

All of hypotheses (a), (b) and (c) amount to assertions that various coefficients are zero. Finally, sometimes we test noncausality hypotheses that involve more than one equation, such as:

- (b) No variable in a set A causes any variable in a set B , in which case we say that the variables in A are block non-causal for those in B .

This particular noncausality hypothesis corresponds to exclusion restrictions that hold simultaneously in a number of equations. Again, however, standard test procedures are applicable.

14.6.5 Impulse-Response Functions

The **impulse-response function** is another device that helps us to learn about the dynamic properties of vector autoregressions. We'll introduce it first in the *univariate* context, and then we'll move to *VARs*. The question of interest is simple and direct: How does a unit innovation to a series affect

it, now and in the future? To answer the question, we simply read off the coefficients in the moving average representation of the process.

We're used to normalizing the coefficient on ε_t to unity in moving-average representations, but we don't have to do so; more generally, we can write

$$y_t = b_0\varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \dots$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

The additional generality introduces ambiguity, however, because we can always multiply and divide every ε_t by an arbitrary constant m , yielding an equivalent model but with different parameters and innovations,

$$y_t = (b_0m) \left(\frac{1}{m}\varepsilon_t \right) + (b_1m) \left(\frac{1}{m}\varepsilon_{t-1} \right) + (b_2m) \left(\frac{1}{m}\varepsilon_{t-2} \right) + \dots$$

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

or

$$y_t = b'_0\varepsilon'_t + b'_1\varepsilon'_{t-1} + b'_2\varepsilon'_{t-2} + \dots$$

$$\varepsilon'_t \sim WN(0, \frac{\sigma^2}{m^2}),$$

where $b'_i = b_i m$ and $\varepsilon'_t = \frac{\varepsilon_t}{m}$.

To remove the ambiguity, we must set a value of m . Typically we set $m = 1$, which yields the standard form of the moving average representation. For impulse-response analysis, however, a different normalization turns out to be particularly convenient; we choose $m = \sigma$, which yields

$$y_t = (b_0\sigma) \left(\frac{1}{\sigma}\varepsilon_t \right) + (b_1\sigma) \left(\frac{1}{\sigma}\varepsilon_{t-1} \right) + (b_2\sigma) \left(\frac{1}{\sigma}\varepsilon_{t-2} \right) + \dots$$

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

or

$$y_t = b'_0\varepsilon'_t + b'_1\varepsilon'_{t-1} + b'_2\varepsilon'_{t-2} + \dots$$

$$\varepsilon'_t \sim WN(0, 1),$$

where $b'_i = b_i\sigma$ and $\varepsilon'_t = \frac{\varepsilon_t}{\sigma}$. Taking $m = \sigma$ converts shocks to “standard deviation units,” because a unit shock to ε'_t corresponds to a one standard deviation shock to ε_t .

To make matters concrete, consider the univariate $AR(1)$ process,

$$y_t = \phi y_{t-1} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

The standard moving average form is

$$y_t = \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \dots$$

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

and the equivalent representation in standard deviation units is

$$y_t = b_0\varepsilon'_t + b_1\varepsilon'_{t-1} + b_2\varepsilon'_{t-2} + \dots$$

$$\varepsilon'_t \sim WN(0, 1)$$

where $b_i = \phi^i\sigma$ and $\varepsilon'_t = \frac{\varepsilon_t}{\sigma}$. The impulse-response function is $\{b_0, b_1, \dots\}$. The parameter b_0 is the contemporaneous effect of a unit shock to ε'_t , or equivalently a one standard deviation shock to ε_t ; as must be the case, then, $b_0 = \sigma$. Note well that b_0 gives the immediate effect of the shock at time t , when it hits. The parameter b_1 , which multiplies ε'_{t-1} , gives the effect of the shock one period later, and so on. The full set of impulse-response coefficients, $\{b_0, b_1, \dots\}$, tracks the complete dynamic response of y to the shock.

Now we consider the multivariate case. The idea is the same, but there are more shocks to track. The key question is, “How does a unit shock to ε_i affect y_j , now and in the future, for all the various combinations of i and j ?”

Consider, for example, the bivariate $VAR(1)$,

$$y_{1t} = \phi_{11}y_{1,t-1} + \phi_{12}y_{2,t-1} + \varepsilon_{1t}$$

$$y_{2t} = \phi_{21}y_{1,t-1} + \phi_{22}y_{2,t-1} + \varepsilon_{2t}$$

$$\varepsilon_{1,t} \sim WN(0, \sigma_1^2)$$

$$\varepsilon_{2,t} \sim WN(0, \sigma_2^2)$$

$$cov(\varepsilon_1, \varepsilon_2) = \sigma_{12}.$$

The standard moving average representation, obtained by back substitution, is

$$y_{1t} = \varepsilon_{1t} + \phi_{11}\varepsilon_{1,t-1} + \phi_{12}\varepsilon_{2,t-1} + \dots$$

$$y_{2t} = \varepsilon_{2t} + \phi_{21}\varepsilon_{1,t-1} + \phi_{22}\varepsilon_{2,t-1} + \dots$$

$$\varepsilon_{1,t} \sim WN(0, \sigma_1^2)$$

$$\varepsilon_{2,t} \sim WN(0, \sigma_2^2)$$

$$cov(\varepsilon_1, \varepsilon_2) = \sigma_{12}.$$

Just as in the univariate case, it proves fruitful to adopt a different normalization of the moving average representation for impulse-response analysis. The multivariate analog of our univariate normalization by σ is called normalization by the Cholesky factor.²⁶ The resulting VAR moving average representation has a number of useful properties that parallel the univariate case precisely. First, the innovations of the transformed system are in standard deviation units. Second, although the current innovations in the standard representation have unit coefficients, the current innovations in the normalized representation have non-unit coefficients. In fact, the first equation has only one current innovation, ε_{1t} . (The other has a zero coefficient.) The second equation has both current innovations. Thus, the ordering of the

²⁶For detailed discussion and derivation of this advanced topic, see Hamilton (1994).

variables can matter.²⁷

If y_1 is ordered first, the normalized representation is

$$\begin{aligned}y_{1,t} &= b_{11}^0 \varepsilon'_{1,t} + b_{11}^1 \varepsilon'_{1,t-1} + b_{12}^1 \varepsilon'_{2,t-1} + \dots \\y_{2,t} &= b_{21}^0 \varepsilon'_{1,t} + b_{22}^0 \varepsilon'_{2,t} + b_{21}^1 \varepsilon'_{1,t-1} + b_{22}^1 \varepsilon'_{2,t-1} + \dots \\\varepsilon'_{1,t} &\sim WN(0, 1) \\\varepsilon'_{2,t} &\sim WN(0, 1) \\cov(\varepsilon'_1, \varepsilon'_2) &= 0.\end{aligned}$$

Alternatively, if y_2 ordered first, the normalized representation is

$$\begin{aligned}y_{2,t} &= b_{22}^0 \varepsilon'_{2,t} + b_{21}^1 \varepsilon'_{1,t-1} + b_{22}^1 \varepsilon'_{2,t-1} + \dots \\y_{1,t} &= b_{11}^0 \varepsilon'_{1,t} + b_{12}^0 \varepsilon'_{2,t} + b_{11}^1 \varepsilon'_{1,t-1} + b_{12}^1 \varepsilon'_{2,t-1} + \dots \\\varepsilon'_{1,t} &\sim WN(0, 1) \\\varepsilon'_{2,t} &\sim WN(0, 1) \\cov(\varepsilon'_1, \varepsilon'_2) &= 0.\end{aligned}$$

Finally, the normalization adopted yields a zero covariance between the disturbances of the transformed system. This is crucial, because it lets us perform the experiment of interest – shocking one variable in isolation of the others, which we can do if the innovations are uncorrelated but can't do if they're correlated, as in the original unnormalized representation.

After normalizing the system, for a given ordering, say y_1 first, we compute four sets of impulse-response functions for the bivariate model: response of y_1 to a unit normalized innovation to y_1 , $\{ b_{11}^0, b_{11}^1, b_{11}^2, \dots \}$, response of y_1 to a unit normalized innovation to y_2 , $\{ b_{12}^1, b_{12}^2, \dots \}$, response of y_2 to a unit

²⁷In higher-dimensional *VAR*'s, the equation that's first in the ordering has only one current innovation, ε'_{1t} . The equation that's second has only current innovations ε'_{1t} and ε'_{2t} , the equation that's third has only current innovations ε'_{1t} , ε'_{2t} and ε'_{3t} , and so on.

normalized innovation to y_2 , $\{ b_{22}^0, b_{22}^1, b_{22}^2, \dots \}$, and response of y_2 to a unit normalized innovation to y_1 , $\{ b_{21}^0, b_{21}^1, b_{21}^2, \dots \}$. Typically we examine the set of impulse-response functions graphically. Often it turns out that impulse-response functions aren't sensitive to ordering, but the only way to be sure is to check.²⁸

In practical applications of impulse-response analysis, we simply replace unknown parameters by estimates, which immediately yields point estimates of the impulse-response functions. Getting confidence intervals for impulse-response functions is trickier, however, and adequate procedures are still under development.

14.6.6 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 show housing starts and completions in Figure ***. 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.

We show the starts correlogram in Table *** and Figure ***. 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 *** and Figure ***, 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

²⁸Note well that the issues of normalization and ordering only affect impulse-response analysis.

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.

The cross-correlation function (Figure ****) for housing starts and completions is very revealing. Starts and completions are highly correlated at all displacements, and a clear pattern emerges as well: although the contemporaneous correlation is high (.78), completions are maximally correlated with starts lagged by roughly 6-12 months (around .90). Again, this makes good sense in light of the time it takes to build a house.

Now we proceed to model starts and completions. We need to select the order, p , of our $VAR(p)$. Based on exploration using multivariate versions of *SIC* and *AIC*, we adopt a $VAR(4)$.

First consider the starts equation (Table ***), residual plot (Figure ***), and residual correlogram (Table ***, Figure ***). The explanatory power of the model is good, as judged by the R^2 as well as the plots of actual and fitted values, and the residuals appear white, as judged by the residual sample autocorrelations, partial autocorrelations, and Ljung-Box statistics. Note as well that no lag of completions has a significant effect on starts, which makes sense – we obviously expect starts to cause completions, but not conversely. The completions equation (Table ***), residual plot (Figure ***), and residual correlogram (Table ***, Figure ***) appear similarly good. Lagged starts, moreover, most definitely have a significant effect on completions.

Table *** 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 is the fact that 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**.

In order to get a feel for the dynamics of the estimated *VAR*, we compute impulse-response functions. We present results for starts first in the ordering, so that a current innovation to starts affects only current starts, but the results are robust to reversal of the ordering.

In Figure ***, we display the impulse-response functions. First let's consider the own-variable impulse responses, that is, the effects of a starts innovation on subsequent starts or a completions innovation on subsequent completions; the effects are similar. In each case, the impulse response is large and decays in a slow, approximately monotonic fashion. In contrast, the cross-variable impulse responses are very different. An innovation to starts produces no movement in completions at first, but the effect gradually builds and becomes large, peaking at about fourteen months. (It takes time to build houses.) An innovation to completions, however, produces little movement in starts at any time.

14.7 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. $\gamma(t, \tau) = \alpha\tau$
- 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. White noise can be simulated on a computer using **random number generators**, which are available in most statistics, econometrics and forecasting packages.

- a. Simulate a Gaussian white noise realization of length 200. Call the white noise ε_t . Compute the correlogram. Discuss.
- b. Form the distributed lag $y_t = \varepsilon_t + .9\varepsilon_{t-1}$, $t = 2, 3, \dots, 200$. Compute the sample autocorrelations and partial autocorrelations. Discuss.
- c. Let $y_1 = 1$ and $y_t = .9y_{t-1} + \varepsilon_t$, $t = 2, 3, \dots, 200$. Compute the sample autocorrelations and partial autocorrelations. Discuss.
4. (Sample autocorrelation functions for trending series)

A tell-tale sign of the slowly-evolving nonstationarity associated with trend is a sample autocorrelation function that damps extremely slowly.

- a. Find three trending series, compute their sample autocorrelation functions, and report your results. Discuss.

- b. Fit appropriate trend models, obtain the model residuals, compute their sample autocorrelation functions, and report your results. Discuss.
5. (Sample autocorrelation functions for seasonal series)
A tell-tale sign of seasonality is a sample autocorrelation function with sharp peaks at the seasonal displacements (4, 8, 12, etc. for quarterly data, 12, 24, 36, etc. for monthly data, and so on).
 - a. Find a series with both trend and seasonal variation. Compute its sample autocorrelation function. Discuss.
 - b. Detrend the series. Discuss.
 - c. Compute the sample autocorrelation function of the detrended series. Discuss.
 - d. Seasonally adjust the detrended series. Discuss.
 - e. Compute the sample autocorrelation function of the detrended, seasonally-adjusted series. Discuss.
6. (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.

7. (Serially correlated disturbances vs. lagged dependent variables)

Estimate the quadratic trend model for log liquor sales with seasonal dummies and three lags of the dependent variable included directly. Discuss your results and compare them to those we obtained when we instead allowed for $AR(3)$ disturbances in the regression.

8. (Liquor sales model selection using AIC and SIC)

Use the AIC and SIC to assess the necessity and desirability of including trend and seasonal components in the liquor sales model.

a. Display the AIC and SIC for a variety of specifications of trend and seasonality. Which would you select using the AIC? SIC? Do the AIC and SIC select the same model? If not, which do you prefer?

b. Discuss the estimation results and residual plot from your preferred model, and perform a correlogram analysis of the residuals. Discuss, in particular, the patterns of the sample autocorrelations and partial autocorrelations, and their statistical significance.

c. How, if at all, are your results different from those reported in the text? Are the differences important? Why or why not?

9. (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 correlogram, Durbin's h test, or the Breusch-Godfrey test? Why?
10. (Assessing the adequacy of the liquor sales model trend specification)

Critique the liquor sales model that we adopted (log liquor sales with quadratic trend, seasonal dummies, and $AR(3)$ disturbances).

- a. If the trend is not a good approximation to the actual trend in the series, would it greatly affect short-run forecasts? Long-run forecasts?
- b. Fit and assess the adequacy of a model with log-linear trend.
- c. How might you fit and assess the adequacy of a broken linear trend? How might you decide on the location of the break point?
- d. Recall our assertion that best practice requires using a χ^2_{m-k} distribution rather than a χ^2_m distribution to assess the significance of Q -statistics for model residuals, where m is the number of autocorrelations included in the Box-Pierce statistic and k is the number of parameters estimated. In several places in this chapter, we failed to heed this advice when evaluating the liquor sales model. If we were instead to compare the residual Q -statistic p -values to a χ^2_{m-k} distribution, how, if at all, would our assessment of the model's adequacy change?
- e. Return to the log-quadratic trend model with seasonal dummies, allow for $AR(p)$ disturbances, and do a systematic selection of p and q using

the *AIC* and *SIC*. Do *AIC* and *SIC* select the same model? If not, which do you prefer? If your preferred model differs from the *AR(3)* that we used, replicate the analysis in the text using your preferred model, and discuss your results.

11. (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?

14.8 Notes

For concise and insightful discussion of random number generation, as well as a variety of numerical and computational techniques, see Press *et al.* (1992)***.

The idea that regression models with serially correlated disturbances are more restrictive than other sorts of transfer function models has a long history in econometrics and engineering and is highlighted in a memorably-titled paper, "Serial Correlation as a Convenient Simplification, not a Nuisance," by Hendry and Mizon (1978)***.

? is the pioneering paper in vector autoregressive econometric modeling. Predictive causality is often called Granger-Sims causality, after **Granger**

(1969) and Sims (1972), who build on earlier work by the mathematician Norbert Weiner. Lütkepohl (1991) is a good reference on VAR analysis and forecasting.

14.9 Christopher A. Sims

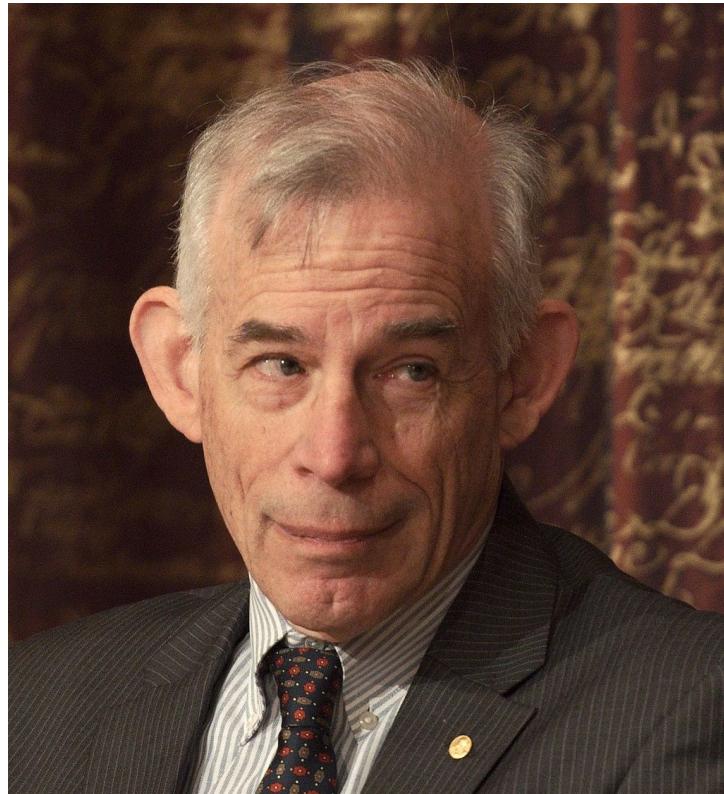


Figure 14.5: Christopher Sims

29

²⁹Nobel Prize 2011; press conference with the laureates of the Nobel prizes in chemistry and physics and the memorial prize in economic sciences at the Royal Swedish Academy of Sciences. Date: 7 December 2011. Source/Author: Holger Motzkau Wikipedia/Wikimedia Commons (cc-by-sa-3.0). The file is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license. The file was made possible through the pool of technology at Wikimedia Sverige. This photo was taken at the Beijer hall of the Royal Swedish Academy of Sciences in Stockholm.

Figure 1
A Rigid Cyclical Pattern

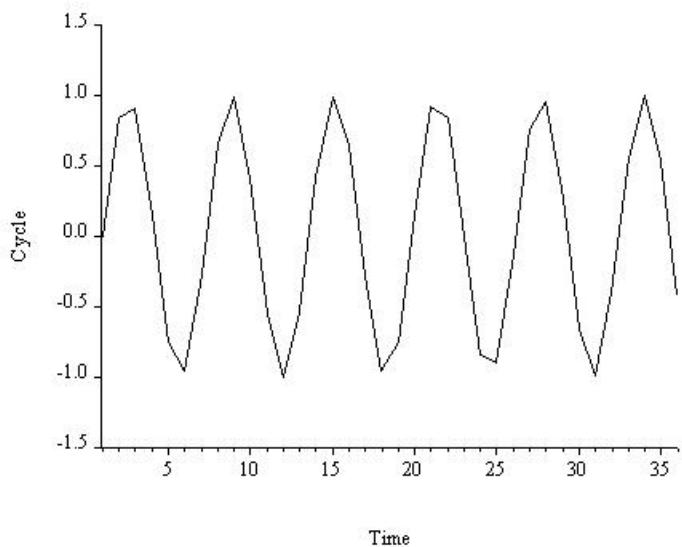
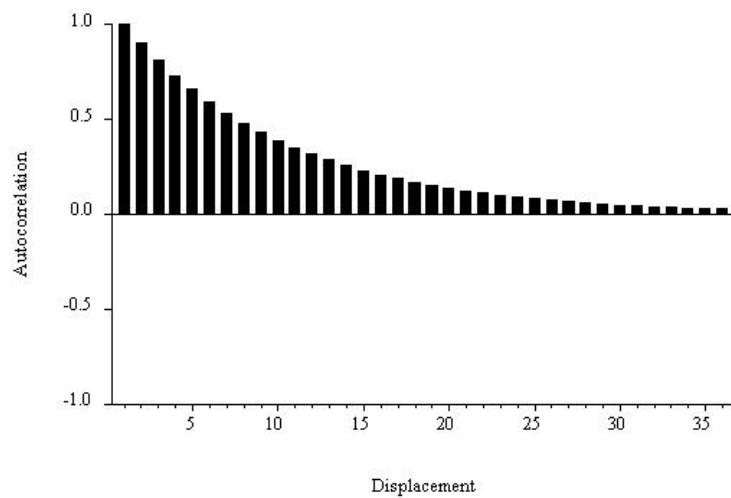


Figure 2
Autocorrelation Function, One-Sided Gradual Damping



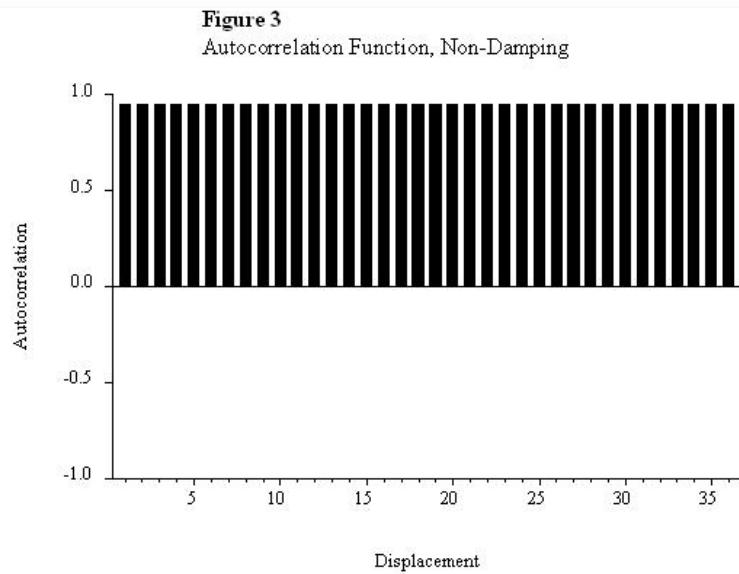


Figure 4
Autocorrelation Function, Gradual Damped Oscillation

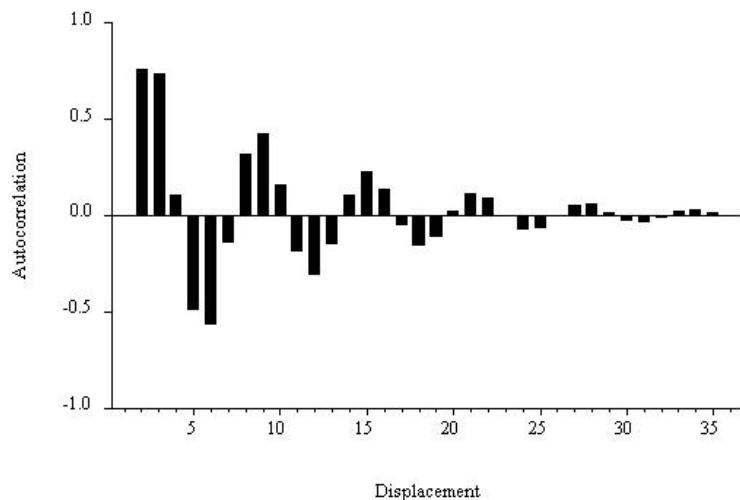


Figure 5
Autocorrelation Function, Sharp Cutoff

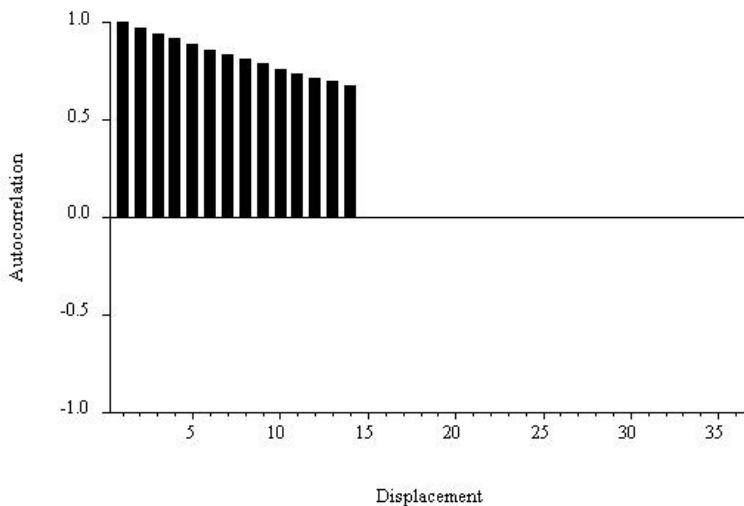


Figure 6
Realization of White Noise Process

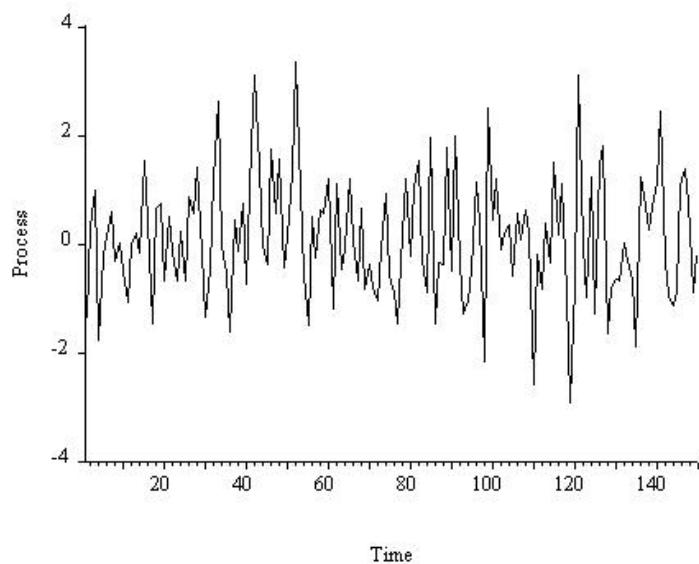


Figure 7
Population Autocorrelation Function
White Noise Process

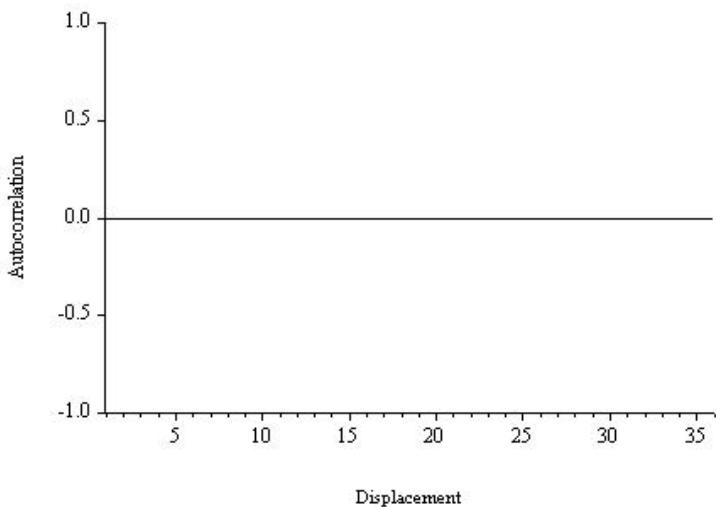
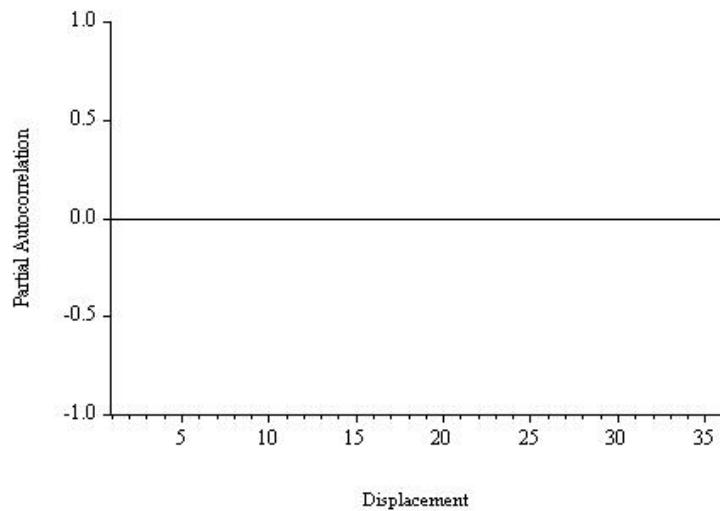


Figure 8
Population Partial Autocorrelation Function
White Noise Process



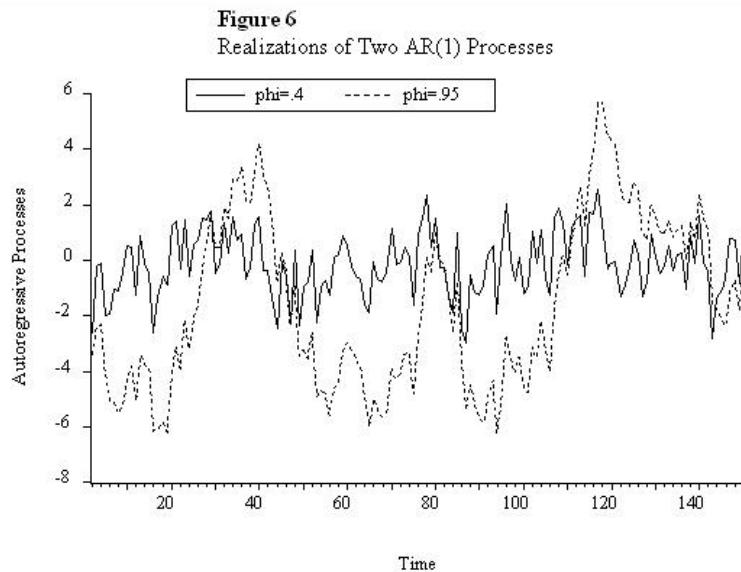
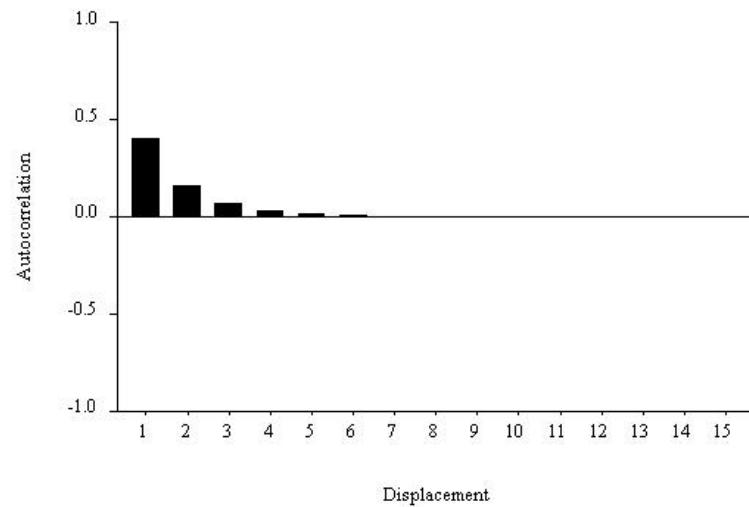
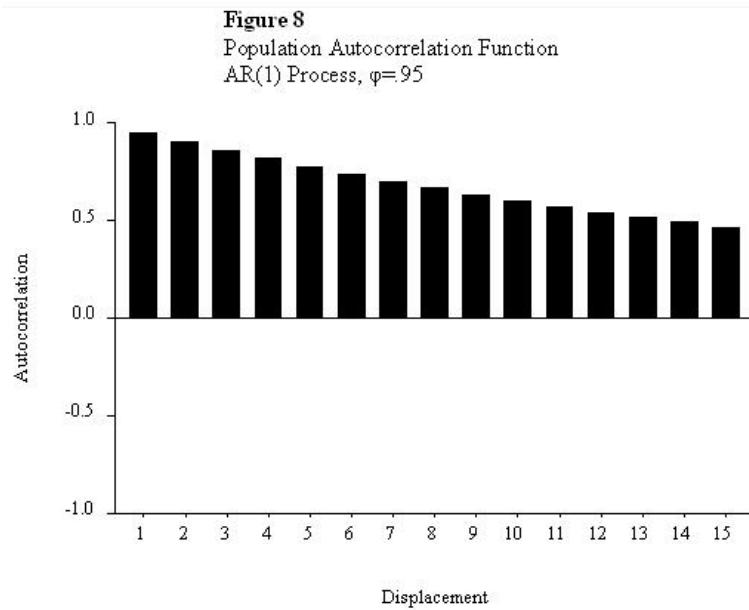
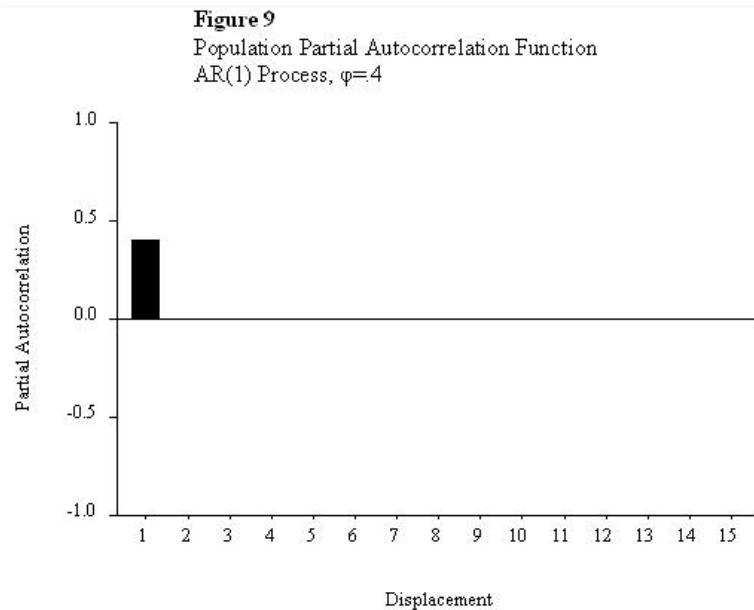


Figure 7
Population Autocorrelation Function
AR(1) Process, $\varphi=4$







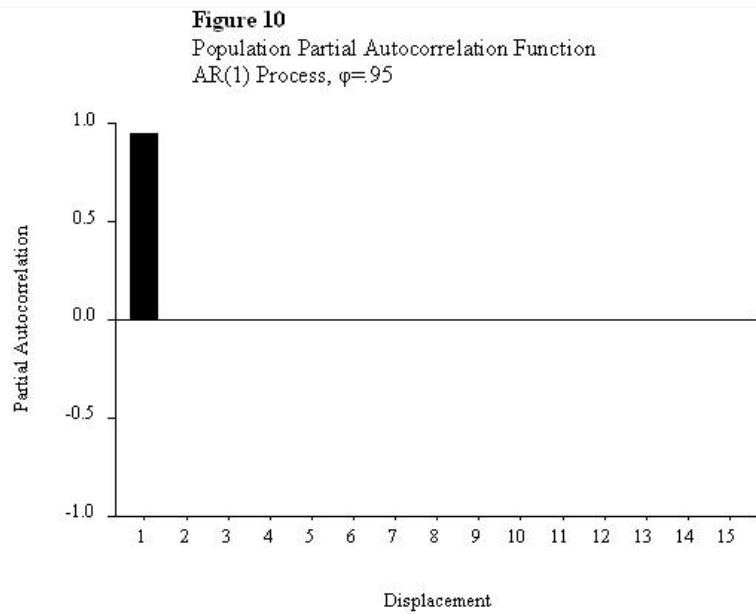
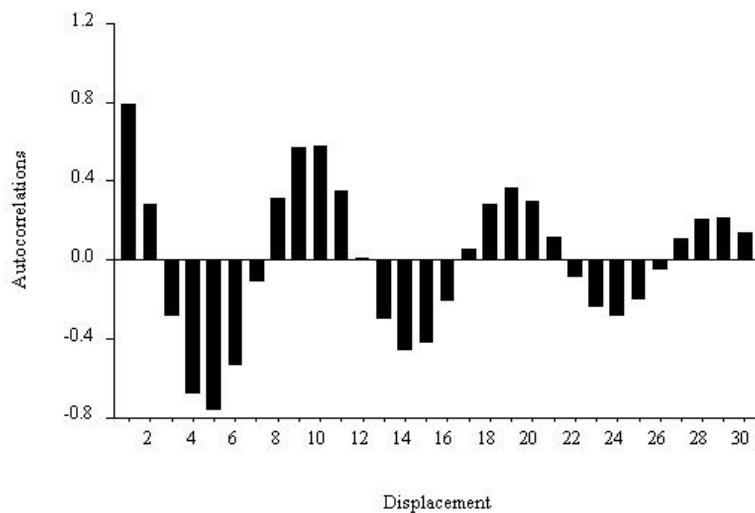


Figure 11
Population Autocorrelation Function
AR(2) Process with Complex Roots



Chapter 15

Structural Change

Much of econometrics is about *failure* of one or more of the FIC. Here we consider another failure: structural change.

Recall the full ideal conditions.

Here we deal with violation of the assumption that the coefficients, β , are fixed.

The dummy variables that we already studied effectively allow for structural change in the cross section (across groups). But structural change is of special relevance in time series. It can be abrupt (e.g., new legislation) or gradual (Lucas critique, learning, ...).

Structural change has many interesting connections to nonlinearity, non-normality and outliers.

Structural change is related to nonlinearity, because structural change is actually a *type* of nonlinearity.

Structural change is related to outliers, because outliers are a kind of structural change. For example, dummying out an outlier amounts to incorporating 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.

15.1 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:

$$y_t \rightarrow c, , TIME_t, x_t, TIME_t * x_t.$$

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.

15.2 Sharp Parameter Breaks

15.2.1 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^* .

The Simplest Case

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:

$$y_t = (\beta_1^1 + (\beta_1^2 - \beta_1^1)D_t) + (\beta_2^1 + (\beta_2^2 - \beta_2^1)D_t)x_t + \varepsilon_t$$

We simply run:

$$y_t \rightarrow c, D_t, x_t, D_t \times x_t$$

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 $D_t \times x_t$, for which the “F” statistic is distributed χ^2 asymptotically (and F in finite samples under normality).

The General Case

Under the no-break null, the so-called Chow breakpoint test statistic,

$$Chow = \frac{(e'e - (e'_1 e_1 + e'_2 e_2))/K}{(e'_1 e_1 + e'_2 e_2)/(T - 2K)},$$

is distributed F in finite samples (under normality) and χ^2 asymptotically.

15.2.2 Endogenously-Selected Breaks

Thus far we have (unrealistically) assumed that the potential break date is known. In practice, of course, potential break dates are 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 of course not χ^2 asymptotically,

let alone F in finite samples, as for the traditional Chow breakpoint test statistic. Rather, the distribution is that of the *maximum* of many draws of such Chow test statistics, which will be centered far to the right of the distribution of any single draw.

The test statistic is

$$MaxChow = \max_{\tau_1 \leq \tau \leq \tau_2} Chow(\tau),$$

where τ denotes sample fraction (typically we take $\tau_1 = .15$ and $\tau_2 = .85$). The distribution of $MaxChow$ has been tabulated.

15.3 Recursive Estimation

$$\begin{aligned} y_t &= \sum_{k=1}^K \beta_k x_{kt} + \varepsilon_t \\ \varepsilon_t &\sim iidN(0, \sigma^2), \end{aligned}$$

$t = 1, \dots, T$.

OLS estimation uses the full sample, $t = 1, \dots, T$.

Recursive least squares uses an expanding sample.

Begin with the first K observations and estimate the model.

Then estimate using the first $K + 1$ observations, and so on.

At the end we have a set of recursive parameter estimates:

$\hat{\beta}_{k,t}$, for $k = 1, \dots, K$ and $t = K, \dots, T$.

15.3.1 Recursive Residuals

At each t , $t = K, \dots, T - 1$, compute a 1-step forecast,

$$\hat{y}_{t+1,t} = \sum_{k=1}^K \hat{\beta}_{kt} x_{k,t+1}.$$

The corresponding forecast errors, or recursive residuals, are

$$\hat{e}_{t+1,t} = y_{t+1} - \hat{y}_{t+1,t}.$$

$$\hat{e}_{t+1,t} \sim N(0, \sigma^2 r_t)$$

where $r_t > 1$ for all t

15.3.2 Standardized Recursive Residuals and CUSUM

$$w_{t+1,t} \equiv \frac{\hat{e}_{t+1,t}}{\sigma \sqrt{r_t}},$$

$t = K, \dots, T - 1$.

Under the maintained assumptions,

$$w_{t+1,t} \sim iidN(0, 1).$$

Then

$$CUSUM_{t^*} \equiv \sum_{t=K}^{t^*} w_{t+1,t}, \quad t^* = K, \dots, T - 1$$

is just a sum of *iid* $N(0, 1)$'s (i.e. a Gaussian random walk).

15.4 Liquor Sales

- 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

15.5 Exercises, Problems and Complements

1. (Dummy Variables, Again and Again)

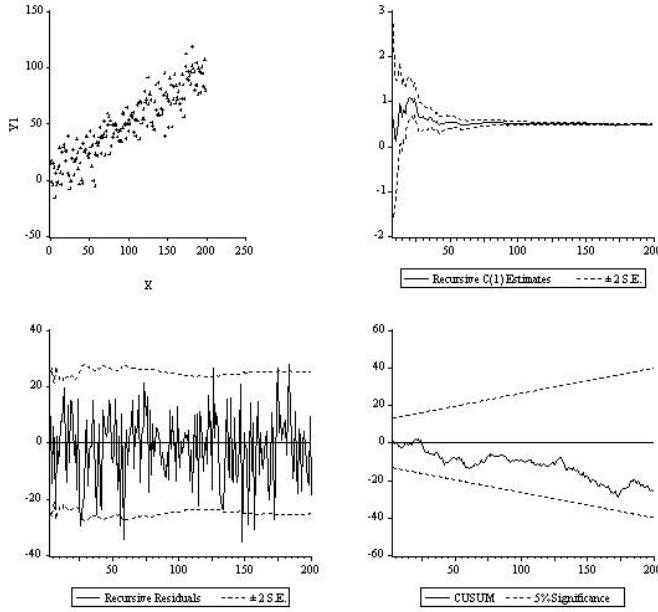


Figure 15.1: Recursive Analysis, Constant Parameter

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.

2. Regime Switching I: Observed-Regime Threshold Model

$$y_t = \begin{cases} c^{(u)} + \phi^{(u)} y_{t-1} + \varepsilon_t^{(u)}, & \theta^{(u)} < y_{t-d} \\ c^{(m)} + \phi^{(m)} y_{t-1} + \varepsilon_t^{(m)}, & \theta^{(l)} < y_{t-d} < \theta^{(u)} \\ c^{(l)} + \phi^{(l)} y_{t-1} + \varepsilon_t^{(l)}, & \theta^{(l)} > y_{t-d} \end{cases}$$

3. Regime Switching II: Markov-Switching Model

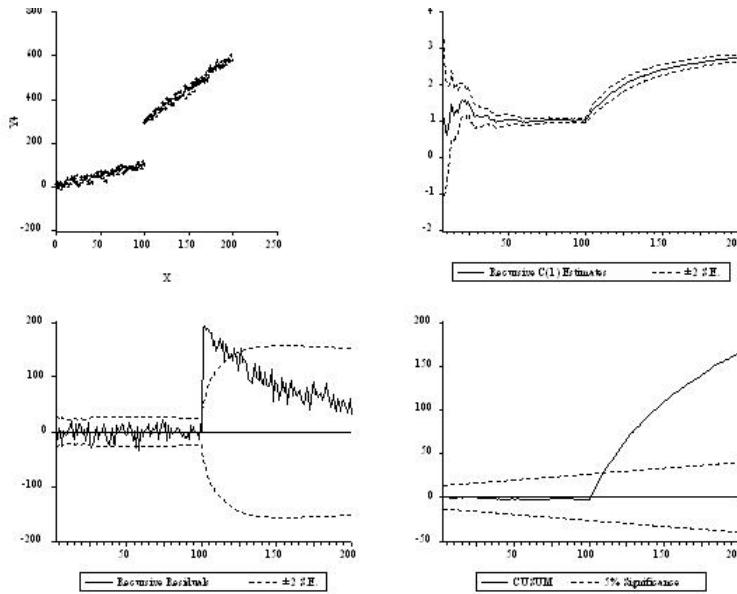


Figure 15.2: Recursive Analysis, Breaking Parameter

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:

$$f(y_t | s_t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y_t - x'_t \beta_{s_t})^2}{2\sigma^2}\right).$$

4. Rolling Regression for Generic Structural Change

15.6 Notes

Chapter 16

Heteroskedasticity in Time Series

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.

16.1 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.”

16.2 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.

16.3 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.

16.3.1 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 $\gamma\neq0$ 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.”

16.3.2 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.

16.3.3 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

16.3.4 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.

16.3.5 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.

16.4 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, $\varepsilon_{\frac{t}{\sigma_t}} = v_t \sim iidN(0, 1)$.

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.

16.5 Stock Market Volatility

We model and forecast the volatility of daily returns on *** from *** through ***, excluding holidays, for a total of *** observations. We estimate using observations ***, and then we forecast observations ***.

In Figure *** we plot the daily returns, r_t . There is no visual evidence of serial correlation in the returns, but there *is* evidence of serial correlation in the *amplitude* of the returns. That is, volatility appears to cluster: large changes tend to be followed by large changes, and small by small, *of either sign*. In Figure *** we show the histogram and related statistics for r_t . The mean daily return is slightly positive. Moreover, the returns are approximately symmetric (only slightly left skewed) but highly leptokurtic.

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.

The Jarque-Bera statistic indicates decisive rejection of normality. In Figure *** we show the correlogram for r_t . The sample autocorrelations are tiny and usually insignificant relative to the Bartlett standard errors, yet the autocorrelation function shows some evidence of a systematic cyclical pattern, and the Q statistics (not shown), which cumulate the information across all displacements, reject the null of weak white noise. Despite the weak serial correlation evidently present in the returns, we will proceed for now as if returns were weak white noise, which is approximately, if not exactly, the case.¹⁷ In Figure *** we plot r_t^2 . The volatility clustering is even more evident than it was in the time series plot of returns. Perhaps the strongest evidence of all comes from the correlogram of r_t^2 , which we show in Figure ***: all sample autocorrelations of r_t^2 are positive, overwhelmingly larger than those of the returns themselves, and statistically significant. As a crude first pass at modeling the stock market volatility, we fit an $AR(5)$ model directly to r_t^2 ; the results appear in Table ***. It is interesting to note that the t -statistics on the lagged squared returns are often significant, even at long lags, yet the R^2 of the regression is low, reflecting the fact that r_t^2 is a very noisy volatility proxy. As a more sophisticated second pass at modeling the return volatility, we fit an $ARCH(5)$ model to r_t ; the results appear in Table ***. The lagged squared returns appear significant even at long lags. The correlogram of squared standardized residuals shown in Figure ***, however, displays some remaining systematic behavior, indicating that the $ARCH(5)$ model fails to capture all of the volatility dynamics, potentially because even longer lags are needed.¹⁸ In Table *** we show the results of fitting a $GARCH(1, 1)$ model. All of the parameter estimates are highly statistically significant, and the “ $ARCH$ coefficient” (α) and “ $GARCH$ coefficient” (β) sum to a value near unity (.987), with β substantially larger than α , as is commonly found

¹⁷In the Exercises, Problems and Complements at the end of this chapter we model the conditional mean, as well as the conditional variance, of returns.

¹⁸In the Exercises, Problems and Complements at the end of this chapter we also examine $ARCH(p)$ models with $p > 5$.

for financial asset returns. We show the correlogram of squared standardized *GARCH*(1, 1) residuals in Figure ***. All sample autocorrelations are tiny and inside the Bartlett bands, and they display noticeably less evidence of any systematic pattern than for the squared standardized *ARCH*(5) residuals. In Figure *** we show the time series of estimated conditional standard deviations implied by the estimated *GARCH*(1, 1) model. Clearly, volatility fluctuates a great deal and is highly persistent. For comparison we show in Figure *** the series of exponentially smoothed r_t^2 , computed using a standard smoothing parameter of .05.¹⁹ Clearly the *GARCH* and exponential smoothing volatility estimates behave similarly, although not at all identically. The difference reflects the fact that the *GARCH* smoothing parameter is effectively estimated by the method of maximum likelihood, whereas the exponential smoothing parameter is set rather arbitrarily. Now, using the model estimated using observations ***, we generate a forecast of the conditional standard deviation for the out-of-sample observations ***. We show the results in Figure ***. The forecast period begins just following a volatility burst, so it is not surprising that the forecast calls for gradual volatility reduction. For greater understanding, in Figure *** we show both a longer history and a longer forecast. Clearly the forecast conditional standard deviation is reverting exponentially to the unconditional standard deviation (**), per the formula discussed earlier.

Heteroskedasticity in Time Series

Key Fact 1: Stock Returns are Approximately Serially Uncorrelated

Key Fact 2: Returns are Unconditionally Non-Gaussian

Unconditional Volatility Measures

Variance: $\sigma^2 = E(r_t - \mu)^2$ (or standard deviation: σ)

¹⁹For comparability with the earlier-computed *GARCH* estimated conditional standard deviation, we actually show the square root of exponentially smoothed r_t^2 .

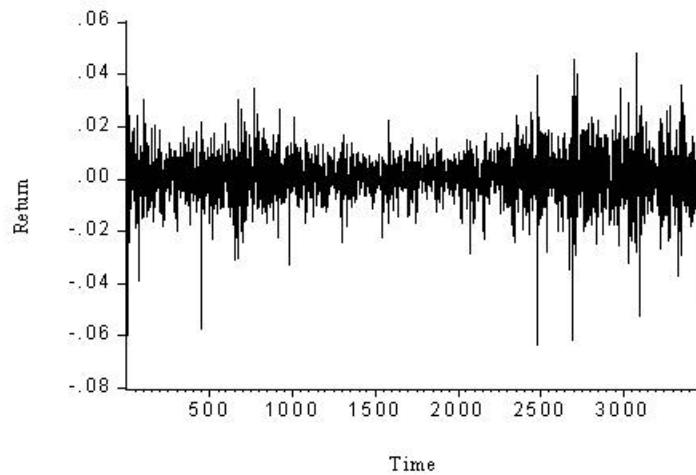


Figure 16.1: Time Series of Daily NYSE Returns.

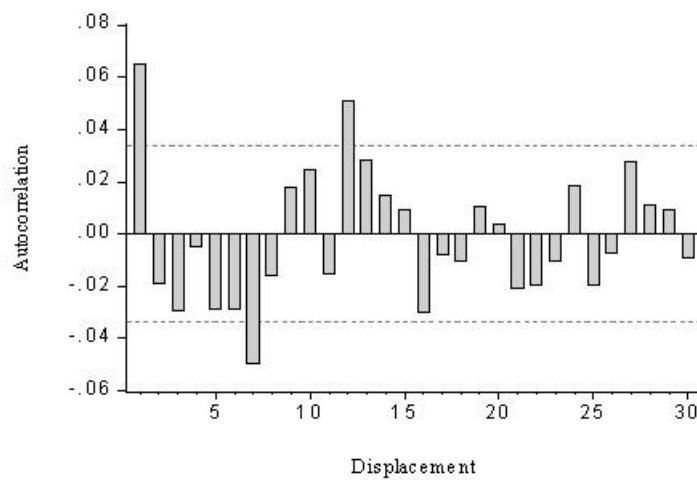


Figure 16.2: Correlogram of Daily Stock Market Returns.

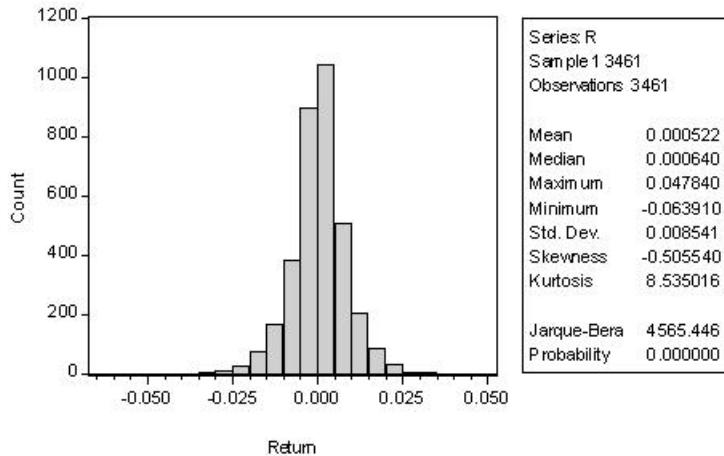


Figure 16.3: Histogram and Statistics for Daily NYSE Returns.

Mean Absolute Deviation: $MAD = E|r_t - \mu|$

Interquartile Range: $IQR = 75\% - 25\%$

Outlier probability: $P|r_t - \mu| > 5\sigma$ (for example)

Tail index: γ s.t. $P(r_t > r) = k r^{-\gamma}$

Kurtosis: $K = E(r - \mu)^4 / \sigma^4$

$p\%$ Value at Risk (VaR^p): x s.t. $P(r_t < x) = p$

Key Fact 3: Returns are Conditionally Heteroskedastic I

Key Fact 3: Returns are Conditionally Heteroskedastic II

Background: Financial Economics Changes Fundmentally

When Volatility is Dynamic

- Risk management
- Portfolio allocation
- Asset pricing

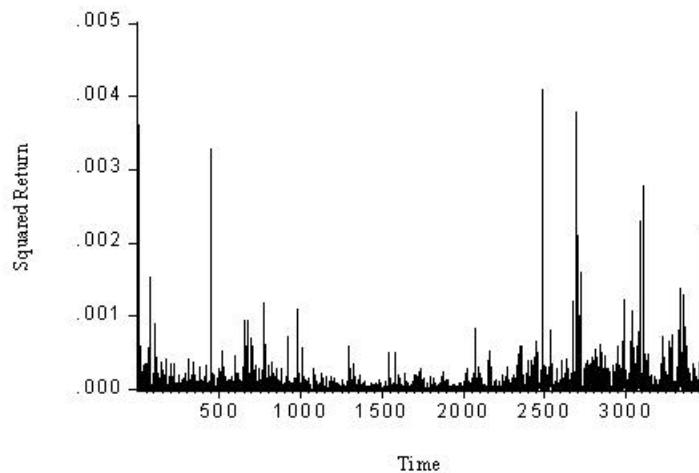


Figure 16.4: Time Series of Daily Squared NYSE Returns

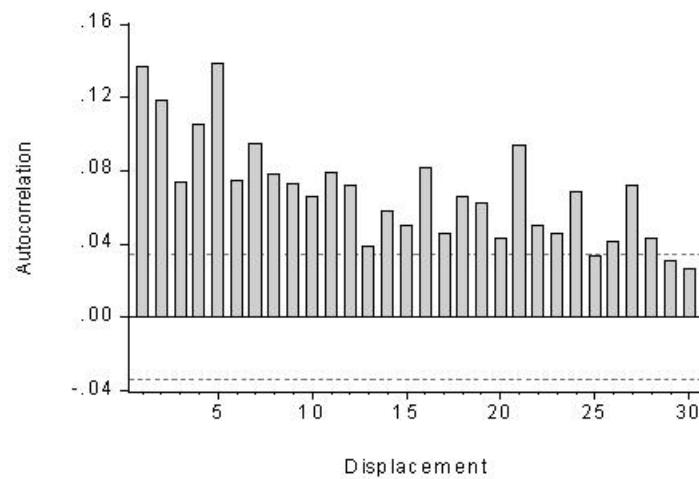


Figure 16.5: Correlogram of Daily Squared NYSE Returns.

- Hedging

- Trading

Asset Pricing I: Sharpe Ratios

Standard Sharpe:

$$\frac{E(r_{it} - r_{ft})}{\sigma}$$

Conditional Sharpe:

$$\frac{E(r_{it} - r_{ft})}{\sigma_t}$$

Asset Pricing II: CAPM Standard CAPM:

$$(r_{it} - r_{ft}) = \alpha + \beta(r_{mt} - r_{ft})$$

$$\beta = \frac{\text{cov}((r_{it} - r_{ft}), (r_{mt} - r_{ft}))}{\text{var}(r_{mt} - r_{ft})}$$

Conditional CAPM:

$$\beta_t = \frac{\text{cov}_t((r_{it} - r_{ft}), (r_{mt} - r_{ft}))}{\text{var}_t(r_{mt} - r_{ft})}$$

Asset Pricing III: Derivatives

Black-Scholes:

$$C = N(d_1)S - N(d_2)Ke^{-r\tau}$$

$$d_1 = \frac{\ln(S/K) + (r + \sigma^2/2)\tau}{\sigma\sqrt{\tau}}$$

$$d_2 = \frac{\ln(S/K) + (r - \sigma^2/2)\tau}{\sigma\sqrt{\tau}}$$

$$P_C = BS(\sigma, \dots)$$

(Standard Black-Scholes options pricing)

Completely different when σ varies!

Conditional Return Distributions

$$f(r_t) \text{ vs. } f(r_t | \Omega_{t-1})$$

$$\text{Key 1: } E(r_t | \Omega_{t-1})$$

Are returns conditional mean independent? Arguably yes.

Returns are (arguably) approximately serially uncorrelated, and (arguably) approximately free of additional non-linear conditional mean dependence.

Conditional Return Distributions, Continued

$$\text{Key 2: } \text{var}(r_t | \Omega_{t-1}) = E((r_t - \mu)^2 | \Omega_{t-1})$$

Are returns conditional variance independent? No way!

Squared returns serially correlated, often with very slow decay.

Linear Models (e.g., AR(1))

$$r_t = \phi r_{t-1} + \varepsilon_t$$

$$\varepsilon_t \sim iid(0, \sigma^2), \quad |\phi| < 1$$

Uncond. mean: $E(r_t) = 0$ (constant)

Uncond. variance: $E(r_t^2) = \sigma^2 / (1 - \phi^2)$ (constant)

Cond. mean: $E(r_t | \Omega_{t-1}) = \phi r_{t-1}$ (varies)

Cond. variance: $E([r_t - E(r_t | \Omega_{t-1})]^2 | \Omega_{t-1}) = \sigma^2$ (constant)

– Conditional mean adapts, but conditional variance does not

ARCH(1) Process

$$r_t | \Omega_{t-1} \sim N(0, h_t)$$

$$h_t = \omega + \alpha r_{t-1}^2$$

$$E(r_t) = 0$$

$$E(r_t^2) = \frac{\omega}{(1 - \alpha)}$$

$$E(r_t | \Omega_{t-1}) = 0$$

$$E([r_t - E(r_t | \Omega_{t-1})]^2 | \Omega_{t-1}) = \omega + \alpha r_{t-1}^2$$

GARCH(1,1) Process (“Generalized ARCH”)

$$r_t | \Omega_{t-1} \sim N(0, h_t)$$

$$h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1}$$

$$E(r_t) = 0$$

$$E(r_t^2) = \frac{\omega}{(1 - \alpha - \beta)}$$

$$E(r_t | \Omega_{t-1}) = 0$$

$$E([r_t - E(r_t | \Omega_{t-1})]^2 | \Omega_{t-1}) = \omega + \alpha r_{t-1}^2 + \beta h_{t-1}$$

Well-defined and covariance stationary if

$$0 < \alpha < 1, 0 < \beta < 1, \alpha + \beta < 1$$

GARCH(1,1) and Exponential Smoothing

Exponential smoothing recursion:

$$\hat{\sigma}_t^2 = \lambda \hat{\sigma}_{t-1}^2 + (1 - \lambda) r_t^2$$

$$\implies \hat{\sigma}_t^2 = (1 - \lambda) \sum_j \lambda^j r_{t-j}^2$$

But in GARCH(1,1) we have:

$$h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1}$$

$$h_t = \frac{\omega}{1 - \beta} + \alpha \sum \beta^{j-1} r_{t-j}^2$$

Unified Theoretical Framework

- Volatility dynamics (of course, by construction)
- Volatility clustering produces unconditional leptokurtosis
- Temporal aggregation reduces the leptokurtosis

Tractable Empirical Framework

$$L(\theta; r_1, \dots, r_T) = f(r_T | \Omega_{T-1}; \theta) f((r_{T-1} | \Omega_{T-2}; \theta) \dots,$$

$$\text{where } \theta = (\omega, \alpha, \beta)'$$

If the conditional densities are Gaussian,

$$f(r_t | \Omega_{t-1}; \theta) = \frac{1}{\sqrt{2\pi}} h_t(\theta)^{-1/2} \exp \left(-\frac{1}{2} \frac{r_t^2}{h_t(\theta)} \right),$$

- Explanatory variables in the variance equation: GARCH-X
- Fat-tailed conditional densities: t-GARCH
- Asymmetric response and the leverage effect: T-GARCH
- Regression with GARCH disturbances
- Time-varying risk premia: GARCH-M

so

$$\ln L = \text{const} - \frac{1}{2} \sum_t \ln h_t(\theta) - \frac{1}{2} \sum_t \frac{r_t^2}{h_t(\theta)}$$

Variations on the GARCH Theme

Explanatory variables in the Variance Equation: GARCH-X

$$h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1} + \gamma z_t$$

where z is a positive explanatory variable

Fat-Tailed Conditional Densities: t-GARCH

If r is conditionally Gaussian, then

$$r_t = \sqrt{h_t} N(0, 1)$$

But often with high-frequency data,

$$\frac{r_t}{\sqrt{h_t}} \sim \text{leptokurtic}$$

So take:

$$r_t = \sqrt{h_t} \frac{t_d}{std(t_d)}$$

and treat d as another parameter to be estimated

Asymmetric Response and the Leverage Effect: T-GARCH

Standard GARCH: $h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1}$

T-GARCH: $h_t = \omega + \alpha r_{t-1}^2 + \gamma r_{t-1}^2 D_{t-1} + \beta h_{t-1}$

$$D_t = \begin{cases} 1 & \text{if } r_t < 0 \\ 0 & \text{otherwise} \end{cases}$$

positive return (good news): α effect on volatility

negative return (bad news): $\alpha + \gamma$ effect on volatility

$\gamma \neq 0$: Asymmetric news response

$\gamma > 0$: “Leverage effect”

Regression with GARCH Disturbances

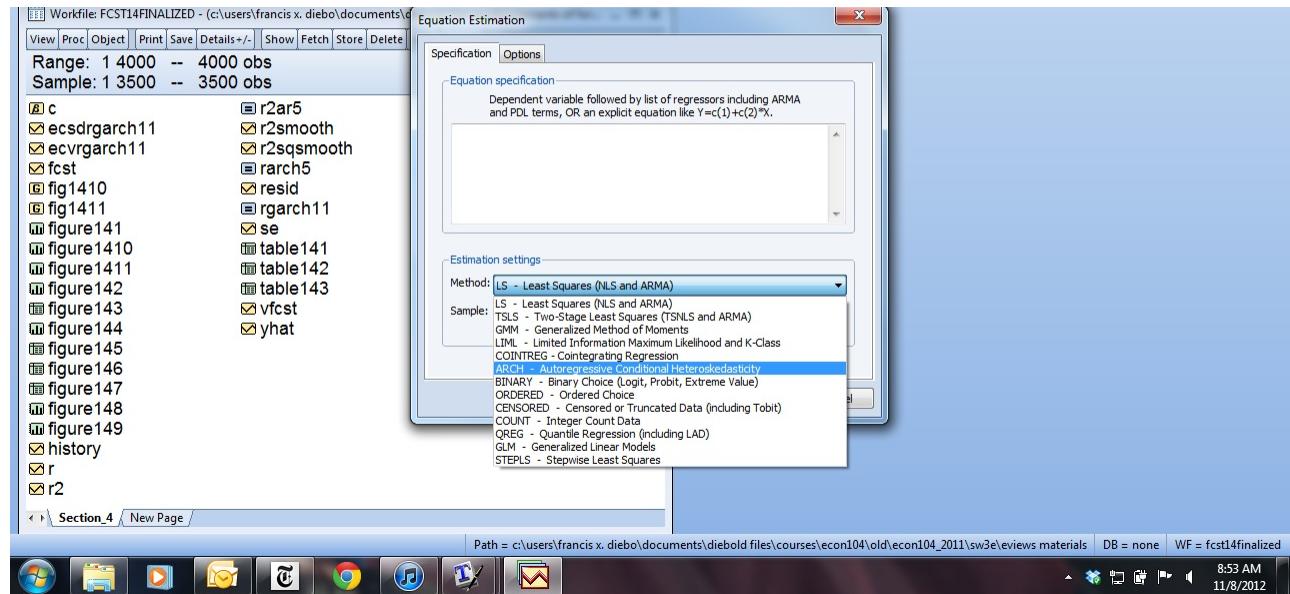
$$y_t = x_t' \beta + \varepsilon_t$$

$$\varepsilon_t | \Omega_{t-1} \sim N(0, h_t)$$

Time-Varying Risk Premia: GARCH-M

Standard GARCH regression model:

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



$$\varepsilon_t | \Omega_{t-1} \sim N(0, h_t)$$

GARCH-M model is a special case:

$$y_t = x_t' \beta + \gamma h_t + \varepsilon_t$$

$$\varepsilon_t | \Omega_{t-1} \sim N(0, h_t)$$

Back to Empirical Work – “Standard” GARCH(1,1)

GARCH(1,1)

GARCH(1,1)

GARCH(1,1)

GARCH(1,1)

A Useful Specification Diagnostic

$$r_t | \Omega_{t-1} \sim N(0, h_t)$$

$$r_t = \sqrt{h_t} \varepsilon_t, \quad \varepsilon_t \sim iidN(0, 1)$$

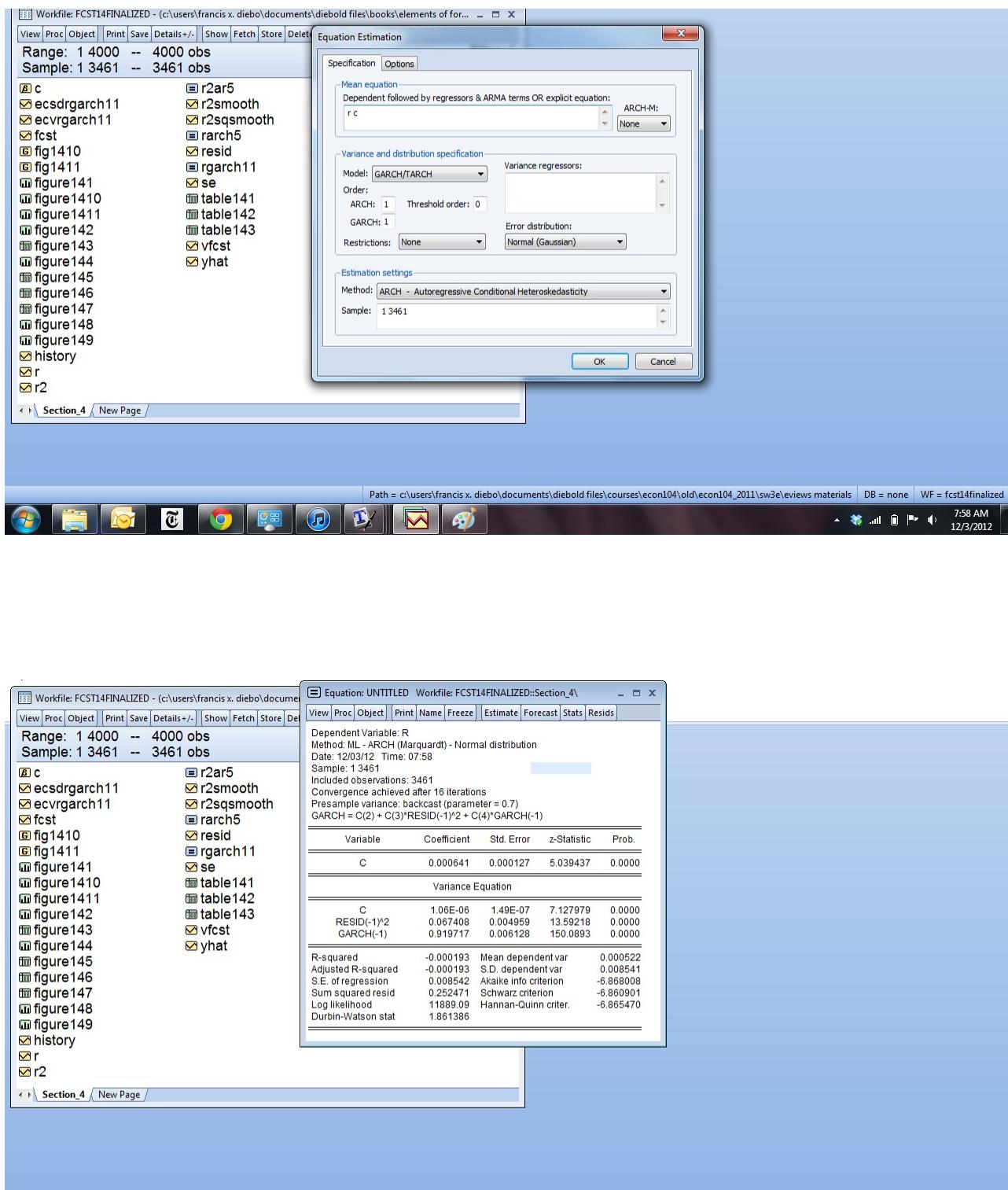


Figure 16.6: GARCH(1,1) Estimation, Daily NYSE Returns.

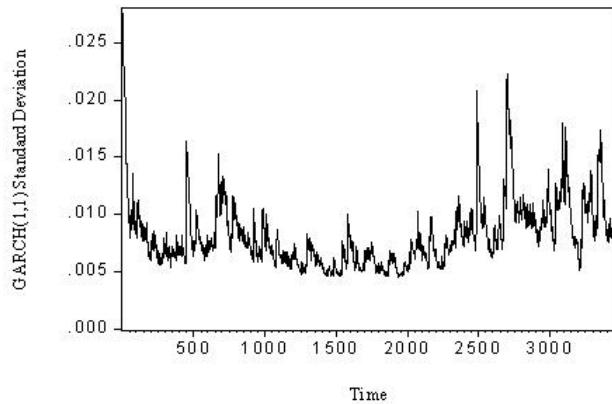


Figure 16.7: Estimated Conditional Standard Deviation, Daily NYSE Returns.

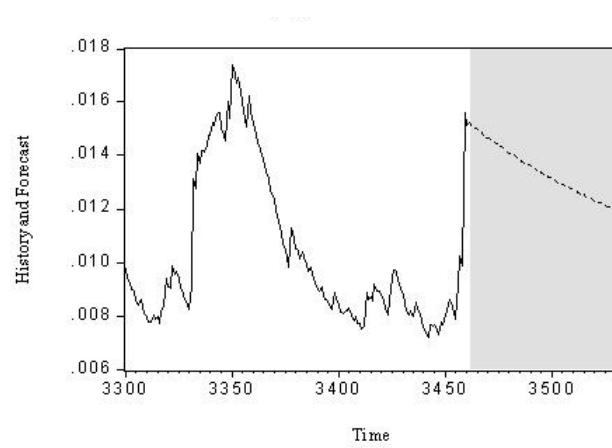


Figure 16.8: Conditional Standard Deviation, History and Forecast, Daily NYSE Returns.

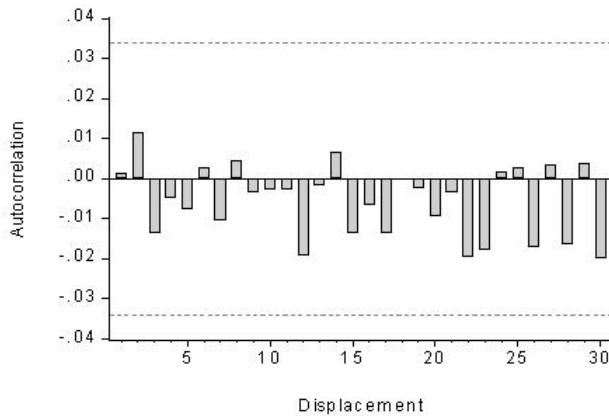


Figure 16.9: Correlogram of Squared Standardized GARCH(1,1) Residuals, Daily NYSE Returns.

$$\frac{r_t}{\sqrt{h_t}} = \varepsilon_t, \quad \varepsilon_t \sim iidN(0, 1)$$

Infeasible: examine ε_t . iid? Gaussian?

Feasible: examine $\hat{\varepsilon}_t = r_t / \sqrt{\hat{h}_t}$. iid? Gaussian?

Key deviation from iid is volatility dynamics. So examine correlogram of squared standardized returns, $\hat{\varepsilon}_t^2$

GARCH(1,1)

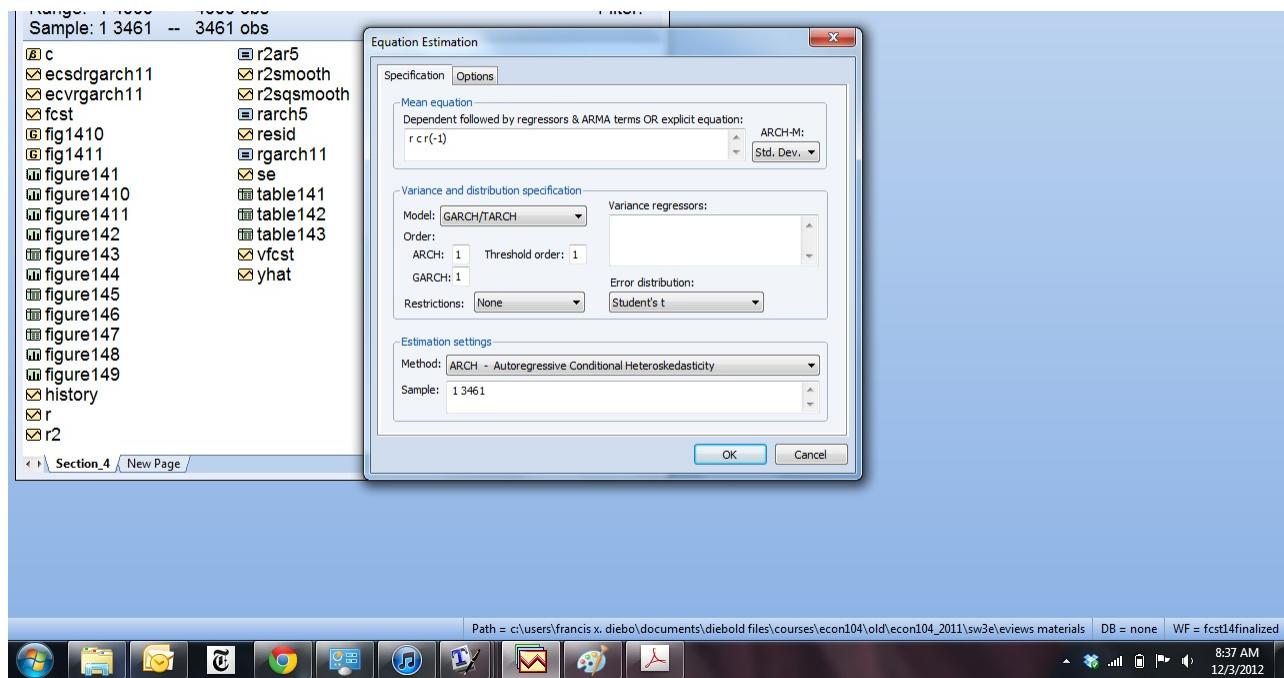
“Fancy” GARCH(1,1)

“Fancy” GARCH(1,1)

16.6 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 dynamics)

In the application in the text we noted that NYSE stock returns appeared to have some weak conditional mean dynamics, yet we ignored them and proceeded directly to model volatility.

- a. Instead, first fit autoregressive models using the SIC to guide order selection, and then fit GARCH models to the residuals. Redo the entire empirical analysis reported in the text in this way, and discuss any important differences in the results.
- b. Consider instead the simultaneous estimation of all parameters of AR(p)-GARCH models. That is, estimate regression models where the regressors are lagged dependent variables and the disturbances

Dependent Variable: R				
Method: ML - ARCH (Marquardt) - Student's t distribution				
Date: 04/10/12 Time: 13:48				
Sample (adjusted): 2 3461				
Included observations: 3460 after adjustments				
Convergence achieved after 19 iterations				
Presample variance: backcast (parameter = 0.7)				
GARCH = C(4) + C(5)*RESID(-1)^2 + C(6)*RESID(-1)^2*(RESID(-1)<0)				
+ C(7)*GARCH(-1)				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
@SQRT(GARCH)	0.083360	0.053138	1.568753	0.1167
C	1.28E-05	0.000372	0.034443	0.9725
R(-1)	0.073763	0.017611	4.188535	0.0000
Variance Equation				
C	1.03E-06	2.23E-07	4.628790	0.0000
RESID(-1)^2	0.014945	0.009765	1.530473	0.1259
RESID(-1)^2*(RESID(-1)<0)	0.094014	0.014945	6.290700	0.0000
GARCH(-1)	0.922745	0.009129	101.0741	0.0000
T-DIST. DOF	5.531579	0.478432	11.56188	0.0000

Figure 16.10: AR(1) Returns with Threshold t-GARCH(1,1)-in Mean.

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?

16.7 Notes

Chapter 17

Endogeneity

Recall the full ideal conditions.

We have thus far been assuming that X is non-stochastic and fixed in repeated samples. (We acknowledged that X is stochastic in autoregressive contexts, but in that case it's "almost" nonstochastic insofar as X contains only lagged dependent variables, which are predetermined if not fully exogenous.) Most results continue to hold if we allow X_t to be stochastic but uncorrelated with ε_t (as when X_t contains only lagged dependent variables); that is, if $E(X'\varepsilon) = 0$.¹

But a host of complications can produce $E(X'\varepsilon) \neq 0$.

In this chapter we confront those complications.

17.1 A Key Subtlety: Causal vs. Non-Causal Regression

There are actually two "flavors" of consistency of concern in econometrics. They happen to coincide under the FIC, but they generally diverge when the the FIC fail.

¹ In fact we can even weaken the $corr(X, \varepsilon) = 0$ slightly, as we did for autoregressive models.

17.1.1 Causal Predictive Modeling and T-Consistency

Consider a standard linear regression setting with K regressors and sample size N .

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 .

T-consistency is the standard econometric notion of consistency.

OLS is T-consistent under the FIC.

OLS is generally not T-consistent without the FIC.

And Remember How Stringent the FIC Are!

1. The fitted model is:

$$y = X\beta + \varepsilon$$

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

and it matches the true data-generating process.

- (a) The relationship, if any, is truly linear, with no omitted variables, no measurement error, etc.
- (b) The coefficients, β , are fixed.
- (c) $\varepsilon \sim N$.
- (d) The ε 's have constant variance σ^2 .
- (e) The ε 's are uncorrelated.
2. There is no redundancy among the variables contained in X , so that $X'X$ is non-singular.

3. X is a non-stochastic matrix, fixed in repeated samples (old style), or
 X is a stochastic matrix such that $E(\varepsilon|X) = 0$ (new style).

17.1.2 Non-Causal Predictive Modeling and P-Consistency

Again consider a standard linear regression setting with K regressors and sample size N .

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!*

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

The distinction between P-consistency and T-consistency is 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 $\hat{\beta}$) 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 full ideal conditions. 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 nonzero 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. (Almost trivially easy to obtain.)
2. T-consistency is consistency for a causal predictive effect. (Notoriously difficult to obtain reliably.)

To link to our wage equation example, predicting someone's wage based on knowledge of her education is straightforward, regardless of whether the FIC hold. Predicting the extra wage accruing to people treated with an extra year of education (i.e., predicting the effects of such a policy) is quite another matter, however, requiring the FIC.

17.1.4 An Example of Correlation Without Causality

To take a simple example, suppose that y and x are in fact causally *unrelated*, so that the true treatment effect of x on y is 0 by construction. But suppose that x is also highly correlated with an unobserved variable z that *does* cause y . Then y and x will be correlated due to their joint dependence on z , and that correlation can be used to predict y given x , despite the fact that, by construction, x treatments (interventions) will have no effect on y .

Consider a thought experiment. Let the DGP be $y_i = z_i + \varepsilon_i$, and suppose also that there exists a variable x such that $\text{corr}(x, z) > 0$. Consider running the regression $y \rightarrow x$. $\hat{\beta}_{OLS}$ P-consistent, as always. But it's not $\hat{\beta}_{OLS}$ T-consistent, because the omitted variable z_i is in ε_i , so that the regressor and disturbance in the fitted regression are correlated. The fitted regression coefficient on x will be non-zero and may be very large, even asymptotically, despite the fact that the true causal impact of x on y is zero, by construction.

17.2 Causes of $E(X'\varepsilon) \neq 0$

17.2.1 Omitted Variables

We already introduced the problem above.

If an omitted variable (“**confounding variable**”) is correlated with an included variable, then stop omitting it! But of course that’s easier said than done. We simply may never know about various omitted variables, or we may suspect them but be unable to measure them.

There may also be problems of “over-controlling.”

Over-Controlling I: Controls Can Induce Multicollinearity

When a previously-omitted variable z is included, and z is highly correlated with x , we encounter potentially severe multicollinearity. This means that we can’t estimate the causal effect of x on y efficiently without a huge sample, although we may be able to estimate the *joint* causal effect of x and z on y . Effectively, inclusion of additional controls may replace potentially-severe inconsistency with potentially-severe inefficiency.

Over-Controlling II: “Controls” Can Introduce More Endogeneity!

The following example is of course absurd, but that’s the point, and it teaches the lesson perfectly.²

Suppose you’re interested in how the quantity of shoes sold varies with price. You fit

$$\text{left shoe sales} = a + b \text{price} + \text{error}.$$

right shoe sales, subsumed in the error, is surely correlated with price, so you ‘control’ for it and run

$$\text{left shoe sales} = a + b \text{Price} + c(\text{right shoe sales}) + \text{error}.$$

²Credit goes to Tom Rothenberg via John Cochrane.

Was that a good idea? The problem here is that the “control” variable (*right shoe sales*) is itself massively endogenous, and almost-perfectly locked to the LHS variable, so that its inclusion actually *worsens* matters, making it basically *impossible* for the regression to determine the effect of price movement on quantity sold (whether left or right!).

17.2.2 Simultaneity

Suppose that y and x are jointly determined, as for example in simultaneous determination of price and quantity in market equilibrium. We may write $y_t = x'_t \beta + \varepsilon$, but note that the ε shocks affect not only y but also x . That is, ε is correlated with x , violating the FIC.

17.2.3 Measurement Error

Measurement errors in x may be correlated with y and hence ε . For example, those people in the extreme tails of the distribution of y (that is, those people who got extremely large positive or negative ε shocks) may have their x 's measured particularly well or poorly.

17.2.4 Sample Selection

17.3 Instrumental Variables

17.3.1 The Basic Idea and the IV Estimator

17.3.2 Instrument Strength and Exogeneity

We want instruments that are both “strong” or “relevant” (highly-correlated with x) and exogenous (uncorrelated with ε). There is a tradeoff. A strong but slightly-endogenous instrument might nevertheless be highly valuable, as might a weak but completely-exogenous instrument.

Weak Instruments

“Nelson-Startz disease”

“Slightly-Endogenous” Instruments

17.3.3 Sources of Instruments

Randomized Experiments

Randomized experiments, sometimes called randomized control trials (RCT’s), are the gold standard. Randomistas.

Natural Experiments

Thought Experiments (Structural Econometric Models)

Require many assumptions. But if the assumptions are credible, they can be used to assess the effects of a wide variety of treatments. “Counterfactuals.”

Time

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} .

17.4 Additional Useful Strategies

17.4.1 Regression Discontinuity Designs

17.4.2 Differences of Differences

17.4.3 Matching

17.5 “Graphical Models”

- Directed Acyclical Graphs
 - Conditional Independence
 - “Do” calculus. Limitation to recursive systems.
 - Bayes nets

17.6 Internal and External Validity

Even randomized experiments have issues. They reveal the treatment effect only for the precise experiment performed. Put differently, if done well, they enjoy internal validity, but there is no guarantee of external validity. Even slight differences in experiments can produce different results. For example, there can be large differences between “open RCT’s” and “double-blind” RCT’s. See <http://boringdevelopment.com/2014/04/09/a-torpedo-aimed-straight-at-h-m-s-randomista/>

17.7 Exercises, Problems and Complements

1. Moment conditions, GMM, and IV.
2. Big data.
 - (a) “Big Data” is typically *found data* – basically “digital exhaust.”

That is, it is non-experimental, so econometric methods are necessary for estimating causal effects.

- (b) Some Big Data, however, is experimental. Google, for example, does many experiments to see how ad click-through rates depend on ad placement.

17.8 Notes

Angrist et al. on experiments. Wolpin on structural models.

Pearl, White

Jim Stock's sleuthing on origins of IV.

Chapter 18

Nonstationarity

18.1 Nonstationary Series

Thus far we've handled nonstationarities, such as trend, using deterministic components. Now we consider an alternative, stochastic, approach. Stochastic trend is important insofar as it sometimes provides a good description of certain business, economic and financial time series, and it has a number of special properties and implications. As we'll see, for example, if we knew for sure that a series had a stochastic trend, then we'd want to difference the series and then fit a stationary model to the difference.¹ The strategy of differencing to achieve stationarity contrasts with the approach of earlier chapters, in which we worked in levels and included deterministic trends. In practice, it's sometimes very difficult to decide whether trend is best modeled as deterministic or stochastic, and the decision is an important part of the science – and art – of building forecasting models.

18.1.1 Stochastic Trends and Forecasting

Consider an $AR(p)$ process,

$$\Phi(L)y_t = \varepsilon_t,$$

¹We speak of modeling in “differences,” as opposed to “levels.” We also use “differences” and “changes” interchangeably.

with all the autoregressive roots on or outside the unit circle, and at most one autoregressive root on the unit circle. We say that y has a **unit autoregressive root**, or simply a **unit root**, if one of the p roots of its autoregressive lag-operator polynomial is 1, in which case we can factor the autoregressive lag-operator polynomial as

$$\Phi(L) = \Phi'(L)(1 - L),$$

where $\Phi'(L)$ is of degree $p - 1$. Thus y is really an $AR(p - 1)$ process in differences, because

$$\Phi'(L)(1 - L)y_t = \varepsilon_t$$

is simply

$$\Phi'(L)\Delta y_t = \varepsilon_t.$$

Note that y is not covariance stationary, because one of the roots of its autoregressive lag-operator polynomial is on the unit circle, whereas covariance stationarity requires all roots to be outside the unit circle. Δy , however, is a covariance stationary and invertible $AR(p - 1)$ process.

You may recall from calculus that we can “undo” an integral by taking a derivative. By analogy, we say that a nonstationary series is **integrated** if its nonstationarity is appropriately “undone” by differencing. If only one difference is required (as with the series y above), we say that the series is integrated of order one, or $I(1)$ (pronounced “eye-one”) for short. More generally, if d differences are required, the series is $I(d)$. The order of integration equals the number of autoregressive unit roots. In practice $I(0)$ and $I(1)$ processes are by far the most important cases, which is why we restricted the discussion above to allow for at most one unit root.² To get a feel for the behavior of $I(1)$ processes, let’s take a simple and very important example, the **random walk**, which is nothing more than an $AR(1)$ process with a unit

² $I(2)$ series sometimes, but rarely, arise, and orders of integration greater than two are almost unheard of.

coefficient,

$$y_t = y_{t-1} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

The random walk is not covariance stationary, because the *AR*(1) coefficient is not less than one. In particular, it doesn't display mean reversion; in contrast to a stationary *AR*(1), it wanders up and down randomly, as its name suggests, with no tendency to return to any particular point. Although the random walk is somewhat ill-behaved, its first difference is the ultimate well-behaved series: zero-mean white noise.

As an illustration, we show a random walk realization of length 300, as well as its first difference, in Figure 1.³ The difference of the random walk is white noise, which vibrates randomly. In contrast, the level of the random walk, which is the cumulative sum of the white noise changes, wanders aimlessly and persistently.

Now let's consider a **random walk with drift**,

$$y_t = \delta + y_{t-1} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

Note that the random walk with drift is effectively a model of trend, because on average it grows each period by the drift, δ . Thus the drift parameter plays the same role as the slope parameter in our earlier model of linear deterministic trend. We call the random walk with drift (and of course also the random walk without drift) a model of **stochastic trend**, because the trend is driven by stochastic shocks, in contrast to the **deterministic trends** considered in Chapter ??.

Just as the random walk has no particular level to which it returns, so too the random walk with drift has no particular trend to which it returns. If a

³The random walk was simulated on a computer with $y_1 = 1$ and $N(0, 1)$ innovations.

shock lowers the value of a random walk, for example, there is no tendency for it to necessarily rise again – we expect it to stay permanently lower. Similarly, if a shock moves the value of a random walk with drift below the currently projected trend, there's no tendency for it to return – the trend simply begins anew from the series' new location. Thus shocks to random walks have completely permanent effects; a unit shock forever moves the expected future path of the series by one unit, regardless of the presence of drift.

For illustration, we show in Figure 2 a realization of a random walk with drift, in levels and differences. As before, the sample size is 300 and $y_1 = 1$. The innovations are $N(0, 1)$ white noise and the drift is $\delta = .3$ per period, so the differences are white noise with a mean of .3. It's hard to notice the nonzero mean in the difference, because the stochastic trend in the level, which is the cumulative sum of $N(.3, 1)$ white noise, dominates the scale.

Let's study the properties of random walks in greater detail. The random walk is

$$\begin{aligned} y_t &= y_{t-1} + \varepsilon_t \\ \varepsilon_t &\sim WN(0, \sigma^2). \end{aligned}$$

Assuming the process started at some time 0 with value

$$y_0,$$

we can write it as

$$y_t = y_0 + \sum_{i=1}^t \varepsilon_i.$$

Immediately,

$$E(y_t) = y_0$$

and

$$var(y_t) = t\sigma^2.$$

In particular note that

$$\lim_{t \rightarrow \infty} \text{var}(y_t) = \infty,$$

so that the variance grows continuously rather than converging to some finite unconditional variance.

Now consider the random walk with drift. The process is

$$y_t = \delta + y_{t-1} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

Assuming the process started at some time 0 with value

$$y_0,$$

we have

$$y_t = t\delta + y_0 + \sum_{i=1}^t \varepsilon_i.$$

Immediately

$$E(y_t) = y_0 + t\delta$$

and

$$\text{var}(y_t) = t\sigma^2.$$

As with the simple random walk, then, the random walk with drift also has the property that

$$\lim_{t \rightarrow \infty} \text{var}(y_t) = \infty.$$

Just as white noise is the simplest $I(0)$ process, the random walk is the simplest $I(1)$ process. And just as $I(0)$ processes with richer dynamics than

white noise can be constructed by transforming white noise, so too can $I(1)$ processes with richer dynamics than the random walk be obtained by transforming the random walk. We're led immediately to the model,

$$\Phi(L)(1 - L)y_t = c + \varepsilon_t$$

or

$$(1 - L)y_t = c\Phi^{-1}(1) + \Phi^{-1}(L)\varepsilon_t,$$

where

$$\Phi(L) = 1 - \Phi_1L - \dots - \Phi_pL^p$$

and all the roots of $\Phi(L)$ are outside the unit circle. *ARI* stands for autoregressive *integrated* process. The $ARI(p, 1)$ process is just a stationary and invertible $AR(p)$ process in first differences.

More generally, we can work with the model,

$$\Phi(L)(1 - L)^d y_t = c + \varepsilon_t$$

or

$$(1 - L)^d y_t = c\Phi^{-1}(1) + \Phi^{-1}(L)\varepsilon_t,$$

where

$$\Phi(L) = 1 - \Phi_1L - \dots - \Phi_pL^p$$

and all the roots of the lag operator polynomial are outside the unit circle. The $ARI(p, d)$ process is a stationary and invertible $AR(p)$ after differencing d times. In practice, $d = 0$ and $d = 1$ are by far the most important cases. When $d = 0$, y is covariance stationary, or $I(0)$, with mean $c\Phi^{-1}(1)$. When $d = 1$, y is $I(1)$ with drift, or stochastic linear trend, of $c\Phi^{-1}(1)$ per period.

It turns out that more complicated $ARI(p, 1)$ processes behave like random walks in certain key respects. First, $ARI(p, 1)$ processes are appropriately made stationary by differencing. Second, shocks to $ARI(p, 1)$ processes

have permanent effects.⁴ Third, the variance of an $ARI(p, 1)$ process grows without bound as time progresses. The special properties of $I(1)$ series, associated with the fact that innovations have permanent effects, have important implications for forecasting. As regards point forecasting, the permanence of shocks means that optimal forecasts, even at very long horizons, don't completely revert to a mean or a trend. And as regards interval and density forecasting, the fact that the variance of an $I(1)$ process approaches infinity as time progresses means that the uncertainty associated with our forecasts, which translates into the width of interval forecasts and the spread of density forecasts, increases without bound as the forecast horizon grows.⁵

Let's see how all this works in the context of a simple random walk, which is an $AR(1)$ process with a unit coefficient. Recall that for the $AR(1)$ process,

$$y_t = \phi y_{t-1} + \varepsilon_t$$

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

the optimal forecast is

$$y_{T+h,T} = \phi^h y_T.$$

Thus in the random walk case of $\phi = 1$, the optimal forecast is simply the current value, regardless of horizon. This makes clear the way that the permanence of shocks to random walk processes affects forecasts: any shock that moves the series up or down today also moves the optimal forecast up or down, at all horizons. In particular, the effects of shocks don't wash out as the forecast horizon lengthens, because the series does not revert to a mean.

In Figure 3, we illustrate the important differences in forecasts from deterministic-trend and stochastic-trend models for U.S. GNP per capita.⁶ We show GNP per capita 1869-1933, followed by the forecasts from the best-fitting

⁴In contrast to random walks, however, the long-run effect of a unit shock to an $ARI(p, 1)$ process may be greater or less than unity, depending on the parameters of the process.

⁵This is true even if we ignore parameter estimation uncertainty.

⁶The GNP per capita data are in logarithms. See Diebold and Senhadji (1996) for details.

deterministic-trend and stochastic-trend models, 1934-1993, made in 1933. The best-fitting deterministic-trend model is an $AR(2)$ in levels with linear trend, and the best-fitting stochastic-trend model is an $AR(1)$ in differences (that is, an $ARI(1,1)$) with a drift.⁷ Because 1932 and 1933 were years of severe recession, the forecasts are made from a position well below trend. The forecast from the deterministic-trend model reverts to trend quickly, in sharp contrast to that from the stochastic-trend model, which remains permanently lower. As it happens, the forecast from the deterministic-trend model turns out to be distinctly better in this case, as shown in Figure 4, which includes the realization.

18.1.2 Unit-Roots: Estimation and Testing

Least-Squares Regression with Unit Roots

The properties of least squares estimators in models with unit roots are of interest to us, because they have implications for forecasting. We'll use a random walk for illustration, but the results carry over to general $ARI(p, 1)$ processes. Suppose that y is a random walk, so that

$$y_t = y_{t-1} + \varepsilon_t,$$

but we don't know that the autoregressive coefficient is one, so we estimate the $AR(1)$ model,

$$y_t = \phi y_{t-1} + \varepsilon_t.$$

Two key and offsetting properties of the least squares estimator emerge: superconsistency and bias.

First we consider superconsistency. In the unit root case of $\phi = 1$, the difference between

$$\hat{\phi}_{LS}$$

⁷Note well that the two dashed lines are two different point extrapolation forecasts, not an interval forecast.

and 1 vanishes quickly as the sample size (T) grows; in fact, it shrinks like $\frac{1}{T}$. Thus, $T(\hat{\phi}_{LS} - 1)$ converges to a non-degenerate random variable. In contrast, in the covariance stationary case of

$$|\phi| < 1,$$

the difference between

$$\hat{\phi}_{LS}$$

and ϕ shrinks more slowly, like $\frac{1}{\sqrt{T}}$, so that $\sqrt{T}(\hat{\phi}_{LS} - \phi)$ converges to a non-degenerate random variable. We call the extra-fast convergence in the unit root case **superconsistency**; we say that the least squares estimator of a unit root is superconsistent. Now we consider bias. It can be shown that the least-squares estimator,

$$\hat{\phi}_{LS},$$

is biased downward, so that if the true value of ϕ is ϕ^* , the expected value of

$$\hat{\phi}_{LS}$$

is less than ϕ^* .⁸ Other things the same, the larger is the true value of ϕ , the larger the bias, so the bias is worst in the unit root case. The bias is also larger if an intercept is included in the regression, and larger still if a trend is included. The bias vanishes as the sample size grows, as the estimate converges to the true population value, but the bias can be sizeable in samples of the size that concern us.

Superconsistency and bias have offsetting effects as regards forecasting. Superconsistency is helpful; it means that the sampling uncertainty in our parameter estimates vanishes unusually quickly as sample size grows. Bias, in contrast, is harmful, because badly biased parameter estimates can translate

⁸The bias in the least-squares estimator in the unit-root and near-unit-root cases was studied by Dickey (1976) and Fuller (1976), and is sometimes called the Dickey-Fuller bias.

into poor forecasts. The superconsistency associated with unit roots guarantees that bias vanishes quickly as sample size grows, but it may nevertheless be highly relevant in small samples.

Effects of Unit Roots on the Sample Autocorrelation and Partial Autocorrelation Functions

If a series has a unit root, its autocorrelation function isn't well-defined in population, because its variance is infinite. But the *sample* autocorrelation function can of course be mechanically computed in the usual way, because the computer software doesn't know or care whether the data being fed into it have a unit root. The sample autocorrelation function will tend to damp extremely slowly; loosely speaking, we say that it fails to damp. The reason is that, because a random walk fails to revert to any population mean, any given sample path will tend to wander above and below its sample mean for long periods of time, leading to very large positive sample autocorrelations, even at long displacements. The sample partial autocorrelation function of a unit root process, in contrast, will damp quickly: it will tend to be very large and close to one at displacement 1, but will tend to be smaller and decay quickly thereafter.

If the properties of the sample autocorrelations and partial autocorrelations of unit root processes appear rather exotic, the properties of the sample autocorrelations and partial autocorrelations of *differences* of unit root processes are much more familiar. That's because the first difference of an $I(1)$ process, by definition, is covariance stationary and invertible.

We illustrate the properties of sample autocorrelations and partial autocorrelations of levels and differences of unit root processes in Figures 5 and 6. In Figure 5 we show the correlogram of our simulated random walk. The sample autocorrelations fail to damp, and the sample partial autocorrelation is huge at displacement 1, but tiny thereafter. In Figure 6, we show the

correlogram of the first difference of the random walk. All the sample autocorrelations and partial autocorrelations are insignificantly different from zero, as expected, because the first difference of a random walk is white noise.

Unit Root Tests

In light of the special properties of series with unit roots, it's sometimes of interest to test for their presence, with an eye toward the desirability of imposing them, by differencing the data, if they seem to be present. Let's start with the simple $AR(1)$ process,

$$y_t = \phi y_{t-1} + \varepsilon_t$$

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

We can regress y_t on y_{t-1} , and then use the standard t -test for testing

$$\phi = 1,$$

$$\hat{\tau} = \frac{\hat{\phi} - 1}{s \sqrt{\frac{1}{\sum_{t=2}^T y_{t-1}^2}}},$$

where s is the standard error of the regression. Note that the $\hat{\tau}$ statistic is *not* the t -statistic computed automatically by regression packages; the standard t -statistic is for the null of a zero coefficient, whereas $\hat{\tau}$ is the t -statistic for a *unit* coefficient. A simple trick, however, coaxes standard software into printing $\hat{\tau}$ automatically. Simply rewrite the first-order autoregression as

$$y_t - y_{t-1} = (\phi - 1)y_{t-1} + \varepsilon_t.$$

Thus, $\hat{\tau}$ is the usual t -statistic in a regression of the first *difference* of y on the first *lag* of y .

A key result is that, in the unit root case,

$$\hat{\tau}$$

does *not* have the t -distribution. Instead it has a special distribution now called the **Dickey-Fuller distribution**, named for two statisticians who studied it extensively in the 1970s and 1980s. Fuller (1976) presents tables of the percentage points of the distribution of $\hat{\tau}$, which we'll call the Dickey-Fuller statistic, under the null hypothesis of a unit root. Because we're only allowing for roots on or outside the unit circle, a one-sided test is appropriate.

Thus far we've shown how to test the null hypothesis of a random walk with no drift against the alternative of a zero-mean, covariance-stationary $AR(1)$. Now we allow for a nonzero mean, μ , under the alternative hypothesis, which is of potential importance because business and economic data can rarely be assumed to have zero mean. Under the alternative hypothesis, the process becomes a covariance stationary $AR(1)$ process *in deviations from the mean*,

$$(y_t - \mu) = \phi(y_{t-1} - \mu) + \varepsilon_t,$$

which we can rewrite as

$$y_t = \alpha + \phi y_{t-1} + \varepsilon_t,$$

where

$$\alpha = \mu(1 - \phi).$$

If we knew μ , we could simply center the data and proceed as before. In practice, of course, μ must be estimated along with the other parameters. Although α vanishes under the unit-root null hypothesis of

$$\phi = 1,$$

it is nevertheless present under the alternative hypothesis, and so we include an intercept in the regression. The distribution of the corresponding Dickey-Fuller statistic, $\hat{\tau}_\mu$, has been tabulated under the null hypothesis of $(\alpha, \phi) = (0, 1)$; tables appear in Fuller (1976).

Finally, let's allow for deterministic linear trend under the alternative hypothesis, by writing the $AR(1)$ in deviations from a linear trend,

$$(y_t - a - bTIME_t) = \phi(y_{t-1} - a - bTIME_{t-1}) + \varepsilon_t,$$

or

$$y_t = \alpha + \beta TIME_t + \phi y_{t-1} + \varepsilon_t,$$

where

$$\alpha = a(1 - \phi) + b\phi$$

and

$$\beta = b(1 - \phi).$$

Under the unit root hypothesis that

$$\phi = 1,$$

we have a random walk with drift,

$$y_t = b + y_{t-1} + \varepsilon_t,$$

which is a stochastic trend, but under the deterministic-trend alternative hypothesis both the intercept and the trend enter and so they must be included in the regression. The random walk with drift is a null hypothesis that frequently arises in economic applications; stationary deviations from linear trend are a natural alternative. The distribution of the Dickey-Fuller statistic

$$\hat{\tau}_\tau,$$

which allows for linear trend under the alternative hypothesis, has been tabulated under the unit root null hypothesis by Fuller (1976).

Now we generalize the test to allow for higher-order autoregressive dynamics. Consider the $AR(p)$ process

$$y_t + \sum_{j=1}^p \phi_j y_{t-j} = \varepsilon_t,$$

which we rewrite as

$$y_t = \rho_1 y_{t-1} + \sum_{j=2}^p \rho_j (y_{t-j+1} - y_{t-j}) + \varepsilon_t,$$

where $p \geq 2$, $\rho_1 = -\sum_{j=1}^p \phi_j$, and $\rho_i = \sum_{j=i}^p \phi_j$, $i = 2, \dots, p$. If there is a unit root, then $\rho_1 = 1$, and y is simply an $AR(p-1)$ in first differences. The Dickey-Fuller statistic for the null hypothesis of $\rho_1 = 1$ has the same asymptotic distribution as $\hat{\tau}$. Thus, the results for the $AR(1)$ process generalize (asymptotically) in a straightforward manner to higher-order processes.

To allow for a nonzero mean in the $AR(p)$ case, write

$$(y_t - \mu) + \sum_{j=1}^p \phi_j (y_{t-j} - \mu) = \varepsilon_t,$$

or

$$y_t = \alpha + \rho_1 y_{t-1} + \sum_{j=2}^p \rho_j (y_{t-j+1} - y_{t-j}) + \varepsilon_t,$$

where

$$\alpha = \mu(1 + \sum_{j=1}^p \phi_j),$$

and the other parameters are as above. Under the null hypothesis of a unit

root, the intercept vanishes, because in that case

$$\sum_{j=1}^p \phi_j = -1.$$

The distribution of the Dickey-Fuller statistic for testing

$$\rho_1 = 1$$

in this regression is asymptotically identical to that of

$$\hat{\tau}_\mu.$$

Finally, to allow for linear trend under the alternative hypothesis, write

$$(y_t - a - bTIME_t) + \sum_{j=1}^p \phi_j(y_{t-j} - a - bTIME_{t-j}) = \varepsilon_t,$$

which we rewrite as

$$y_t = k_1 + k_2TIME_t + \rho_1 y_{t-1} + \sum_{j=2}^p \rho_j(y_{t-j+1} - y_{t-j}) + \varepsilon_t,$$

where

$$k_1 = a(1 + \sum_{i=1}^p \phi_i) - b \sum_{i=1}^p i\phi_i$$

and

$$k_2 = b(1 + \sum_{i=1}^p \phi_i).$$

Under the null hypothesis,

$$k_1 = -b \sum_{i=1}^p i\phi_i$$

and

$$k_2 = 0.$$

The Dickey-Fuller statistic for the hypothesis that

$$\rho_1 = 1$$

has the

$$\hat{\tau}_\tau$$

distribution asymptotically.

New tests, with better power than the Dickey-Fuller tests in certain situations, have been proposed recently.⁹ But power and size problems will always plague unit root tests; power problems, because the relevant alternative hypotheses are typically very close to the null hypothesis, and size problems, because we should include infinitely many augmentation lags in principle but we can't in practice.

Thus, although unit root tests are sometimes useful, don't be fooled into thinking they're the end of the story as regards the decision of whether to specify models in levels or differences. For example, the fact that we can't reject a unit root doesn't necessarily mean that we should impose it – the power of unit root tests against alternative hypotheses near the null hypothesis, which are the relevant alternatives, is likely to be low. On the other hand, it may sometimes be desirable to impose a unit root even when the true root is less than one, if the true root is nevertheless very close to one, because the Dickey-Fuller bias plagues estimation in levels. We need to use introspection and theory, in addition to formal tests, to guide the difficult decision of whether to impose unit roots, and we need to compare the forecasting performance of different models with and without unit roots imposed.

In certain respects, the most important part of unit root theory for fore-

⁹See Elliott, Rothenberg and Stock (1996), Dickey and Gonzalez-Farias (1992), and the comparisons in Pantula, Gonzalez-Farias and Fuller (1994).

casting concerns estimation, not testing. It's important for forecasters to understand the effects of unit roots on consistency and small-sample bias. Such understanding, for example, leads to the insight that at least *asymptotically* we're probably better off estimating forecasting models in levels with trends included, because then we'll get an accurate approximation to the dynamics in the data regardless of the true state of the world, unit root or no unit root. If there's no unit root, then of course it's desirable to work in levels, and if there is a unit root, the estimated largest root will converge appropriately to unity, and at a fast rate. On the other hand, differencing is appropriate only in the unit root case, and inappropriate differencing can be harmful, even asymptotically.

18.1.3 Smoothing

We bumped into the idea of time series smoothing early on, when we introduced simple moving-average smoothers as ways of estimating trend.¹⁰ Now we introduce additional smoothing techniques and show how they can be used to produce forecasts.

Smoothing techniques, as traditionally implemented, have a different flavor than the modern model-based methods that we've used in this book. Smoothing techniques, for example, don't require "best-fitting models," and they don't generally produce "optimal forecasts." Rather, they're simply a way to tell a computer to draw a smooth line through data, just as we'd do with a pencil, and to extrapolate the smooth line in a reasonable and replicable way.

When using smoothing techniques, we make no attempt to find the model that best fits the data; rather, we force a prespecified model on the data. Some academics turn their nose at smoothing techniques for that reason, but such behavior reflects a shallow understanding of key aspects of applied fore-

¹⁰See the Exercises, Problems and Complements of Chapter ??.

casting – smoothing techniques have been used productively for many years, and for good reason. They’re most useful in situations when model-based methods can’t, or shouldn’t, be used. First, available samples of data are sometimes very small. Suppose, for example, that we must produce a forecast based on a sample of historical data containing only four observations. This scenario sounds extreme, and it is, but such scenarios arise occasionally in certain important applications, as when forecasting the sales of a newly-introduced product. In such cases, available degrees of freedom are so limited as to render any estimated model of dubious value. Smoothing techniques, in contrast, require no estimation, or minimal estimation.

Second, the forecasting task is sometimes immense. Suppose, for example, that each week we must forecast the prices of 10,000 inputs to a manufacturing process. Again, such situations are extreme, but they do occur in practice – think of how many parts there are in a large airplane. In such situations, even if historical data are plentiful (and of course, they might not be) , there is simply no way to provide the tender loving care required for estimation and maintenance of 10,000 different forecasting models. Smoothing techniques, in contrast, require little attention. They’re one example of what are sometimes called “automatic” forecasting methods and are often useful for forecasting voluminous, high-frequency, data.

Finally, smoothing techniques *do* produce optimal forecasts under certain conditions, which turn out to be intimately related to the presence of unit roots in the series being forecast. That’s why we waited until now to introduce them. Moreover, fancier approaches produce optimal forecasts only under certain conditions as well, such as correct specification of the forecasting model. As we’ve stressed throughout, all our models are approximations, and all are surely false. Any procedure with a successful track record in practice is worthy of serious consideration, and smoothing techniques do have successful track records in the situations sketched above.

Moving Average Smoothing, Revisited

As a precursor to the more sophisticated smoothing techniques that we'll soon introduce, recall the workings of simple moving-average smoothers. Denote the original data by

$$y_{t=1}^T$$

and the smoothed data by

$$\bar{y}_t.$$

Then the two-sided moving average is

$$\bar{y}_t = (2m + 1)^{-1} \sum_{i=-m}^m y_{t-i},$$

the one-sided moving average is

$$\bar{y}_t = (m + 1)^{-1} \sum_{i=0}^m y_{t-i},$$

and the one-sided weighted moving average is

$$\bar{y}_t = \sum_{i=0}^m w_i y_{t-i}.$$

The standard one-sided moving average corresponds to a one-sided weighted moving average with all weights equal to

$$(m + 1)^{-1}.$$

The user must choose the smoothing parameter, m ; the larger is m , the more smoothing is done.

One-sided weighted moving averages turn out to be very useful in practice. The one-sided structure means that at any time t , we need only current and past data for computation of the time- t smoothed value, which means

that it can be implemented in real time. The weighting, moreover, enables flexibility in the way that we discount the past. Often, for example, we want to discount the distant past more heavily than the recent past. Exponential smoothing, to which we now turn, is a particularly useful and convenient way of implementing such a moving average.

Exponential Smoothing

Exponential smoothing, also called **simple exponential smoothing**, or **single exponential smoothing**, is what's called an **exponentially weighted moving average**, for reasons that will be apparent soon. The basic framework is simple. Imagine that a series c_0 is a random walk,

$$c_{0t} = c_{0,t-1} + \eta_t$$

$$\eta_t \sim WN(0, \sigma_\eta^2),$$

in which case the level of c_0 wanders randomly up and down, and the best forecast of any future value is simply the current value. Suppose, however, that we don't see c_0 ; instead, we see y , which is c_0 plus white noise,¹¹

$$y_t = c_{0t} + \varepsilon_t,$$

where ε is uncorrelated with η at all leads and lags. Then our optimal forecast of any future y is just our optimal forecast of future c_0 , which is current c_0 , plus our optimal forecast of future ε , which is 0. The problem, of course, is that we don't know current c_0 , the current "local level." We do know current and past y , however, which should contain information about current c_0 . When the data-generating process is as written above, exponential smoothing constructs the optimal estimate of c_0 – and hence the optimal forecast of any future value of y – on the basis of current and past y . When the data-

¹¹We can think of the added white noise as measurement error.

generating process is not as written above, the exponential smoothing forecast may not be optimal, but recent work suggests that exponential smoothing remains optimal or nearly-optimal under surprisingly broad circumstances.¹²

As is common, we state the exponential smoothing procedure as an algorithm for converting the observed series, $y_{t=1}^T$, into a smoothed series, $\bar{y}_{t=1}^T$, and forecasts, \bar{y}

$$\hat{y}_{T+h,T} :$$

(1) Initialize at $t=1$:

$$\bar{y}_1 = y_1.$$

(2) Update:

$$\bar{y}_t = \alpha y_t + (1 - \alpha)\bar{y}_{t-1}, t = 2, \dots, T.$$

(3) Forecast:

$$\hat{y}_{T+h,T} = \bar{y}_T.$$

Referring to the level of c_0 , we call \bar{y}_t the estimate of the *level* at time t . The smoothing parameter α is in the unit interval, $\alpha \in [0, 1]$. The smaller is α the smoother the estimated level. As α approaches 0, the smoothed series approaches constancy, and as α approaches 1, the smoothed series approaches point-by-point interpolation. Typically, the more observations we have per unit of calendar time, the more smoothing we need; thus we'd smooth weekly data (52 observations per year) more than quarterly data (4 observations per year). There is no substitute, however, for a trial-and-error approach involving a variety of values of the smoothing parameter.

It's not obvious at first that the algorithm we just described delivers a one-sided moving average with exponentially declining weights. To convince yourself, start with the basic recursion,

$$\bar{y}_t = \alpha y_t + (1 - \alpha)\bar{y}_{t-1},$$

¹²See, in particular, Chatfield et al. (2001).

and substitute backward for \bar{y}_t , which yields

$$\bar{y}_t = \sum_{j=0}^{t-1} w_j y_{t-j},$$

where

$$w_j = \alpha(1 - \alpha)^j.$$

Suppose, for example, that $\alpha=.5$. Then

$$w_0 = .5(1 - .5)^0 = .5$$

$$w_1 = .5(1 - .5) = .25$$

$$w_2 = .5(1 - .5)^2 = .125,$$

and so forth. Thus moving average weights decline exponentially, as claimed.

Notice that exponential smoothing has a recursive structure, which can be very convenient when data are voluminous. At any time t , the new time t estimate of the level, \bar{y}_t , is a function only of the previously-computed estimate, \bar{y}_{t-1} , and the new observation, y_t . Thus there's no need to re-smooth the entire dataset as new data arrive.

Holt-Winters Smoothing

Now imagine that we have not only a slowly-evolving local level, but also a trend with a slowly-evolving local slope,

$$y_t = c_{0t} + c_{1t} \text{TIME}_t + \varepsilon_t$$

$$c_{0t} = c_{0,t-1} + \eta_t$$

$$c_{1t} = c_{1,t-1} + \nu_t,$$

where all the disturbances are orthogonal at all leads and lags. Then the optimal smoothing algorithm, named Holt-Winters smoothing after the re-

searchers who worked it out in the 1950s and 1960s, is

(1) Initialize at $t = 2$:

$$\bar{y}_2 = y_2$$

$$F_2 = y_2 - y_1.$$

(2) Update:

$$\bar{y}_t = \alpha y_t + (1 - \alpha) (\bar{y}_{t-1} + F_{t-1}), 0 < \alpha < 1$$

$$F_t = \beta (\bar{y}_t - \bar{y}_{t-1}) + (1 - \beta) F_{t-1}, 0 < \beta < 1$$

$$t = 3, 4, \dots, T.$$

(3) Forecast:

$$\hat{y}_{T+h,T} = \bar{y}_T + hF_T.$$

\bar{y}_t is the estimated, or smoothed, level at time t , and F_t is the estimated slope at time t . The parameter α controls smoothing of the level, and β controls smoothing of the slope. The h -step-ahead forecast simply takes the estimated level at time T and augments it with h times the estimated slope at time T .

Again, note that although we've displayed the data-generating process for which Holt-Winters smoothing produces optimal forecasts, when we apply Holt-Winters we don't assume that the data are actually generated by that process. We hope, however, that the actual data-generating process is close to the one for which Holt-Winters is optimal, in which case the Holt-Winters forecasts may be close to optimal.

Holt-Winters Smoothing with Seasonality

We can augment the Holt-Winters smoothing algorithm to allow for seasonality with period s . The algorithm becomes:

(1) Initialize at $t = s$:

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

$$F_s = 0$$

$$G_j = \frac{y_j}{\left(\frac{1}{s} \sum_{t=1}^s y_t\right)}, j = 1, 2, \dots, s.$$

(2) Update:

$$\bar{y}_t = \alpha (y_t - G_{t-s}) + (1 - \alpha) (\bar{y}_{t-1} + F_{t-1}), 0 < \alpha < 1$$

$$F_t = \beta (\bar{y}_t - \bar{y}_{t-1}) + (1 - \beta) F_{t-1}, 0 < \beta < 1$$

$$G_t = \gamma (y_t - \bar{y}_t) + (1 - \gamma) G_{t-s}, 0 < \gamma < 1$$

$$t = s + 1, \dots, T.$$

(3) Forecast:

$$\hat{y}_{T+h,T} = \bar{y}_T + hF_T + G_{T+h-s}, h = 1, 2, \dots, s,$$

$$\hat{y}_{T+h,T} = \bar{y}_T + hF_T + G_{T+h-2s}, h = s+1, s+2, \dots, 2s,$$

etc.

The only thing new is the recursion for the seasonal, with smoothing parameter γ .

Forecasting with Smoothing Techniques

Regardless of which smoothing technique we use, the basic paradigm is the same. We plug data into an algorithm that smooths the data and lets us generate point forecasts. The resulting point forecasts are optimal for certain data-generating processes, as we indicated for simple exponential smoothing and Holt-Winters smoothing without seasonality. In practice, of course, we don't know if the actual data-generating process is close to the one for which

the adopted smoothing technique is optimal; instead, we just swallow hard and proceed. That's the main contrast with the model-based approach, in which we typically spend a lot of time trying to find a “good” specification.

The “one-size-fits-all” flavor of the smoothing approach has its costs, because surely one size does *not* fit all, but it also has benefits in that no, or just a few, parameters need be estimated. Sometimes we simply set the smoothing parameter values based upon our knowledge of the properties of the series being considered, and sometimes we select parameter values that provide the best h-step-ahead forecasts under the relevant loss function. For example, under 1-step-ahead squared-error loss, if the sample size is large enough so that we're willing to entertain estimation of the smoothing parameters, we can estimate them as,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{t=m+1}^T (y_t - \hat{y}_{t-1,t})^2,$$

where m is an integer large enough such that the start-up values of the smoothing algorithm have little effect.

In closing this section, we note that smoothing techniques, as typically implemented, produce point forecasts only. They may produce optimal point forecasts for certain special data-generating processes, but typically we don't assume that those special data-generating processes are the truth. Instead, the smoothing techniques are used as “black boxes” to produce point forecasts, with no attempt to exploit the stochastic structure of the data to find a best-fitting model, which could be used to produce interval or density forecasts in addition to point forecasts.

Random walk:

$$y_t = y_{t-1} + \varepsilon_t$$

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

Just a simple special case of $AR(1)$ $\phi = 1$

Random Walk with Drift

$$y_t = \delta + y_{t-1} + \varepsilon_t$$

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

$$y_t = t\delta + y_0 + \sum_{i=1}^t \varepsilon_i$$

$$E(y_t) = y_0 + t\delta$$

$$var(y_t) = t\sigma^2$$

$$\lim_{t \rightarrow \infty} var(y_t) = \infty$$

Recall Properties of $AR(1)$ with $|\phi| < 1$

- Shocks ε_t have persistent but not permanent effects

$$y_t = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j} \quad (\text{note } \phi_j \rightarrow 0)$$

- Series y_t varies but not too extremely

$$\text{var}(y_t) = \frac{\sigma^2}{1 - \phi^2} \quad (\text{note } \text{var}(y_t) < \infty)$$

- Autocorrelations $\rho(\tau)$ nonzero but decay to zero

$$\rho(\tau) = \phi^\tau \quad (\text{note } \phi^\tau \rightarrow 0)$$

Properties of the Random Walk ($AR(1)$) With $|\phi| = 1$)

- Shocks have permanent effects

$$y_t = y_0 + \sum_{j=0}^{t-1} \varepsilon_{t-j}$$

- Series is infinitely variable

$$E(y_t) = y_0$$

$$\text{var}(y_t) = t\sigma^2$$

$$\lim_{t \rightarrow \infty} \text{var}(y_t) = \infty$$

- Autocorrelations $\rho(\tau)$ do not decay

$$\rho(\tau) \approx 1 \quad (\text{formally not defined})$$

A Key Insight Regarding the Random Walk

- Level series y_t is non-stationary (of course)

- Differenced series y_t is stationary (indeed white noise)!

$$\Delta y_t = \varepsilon_t$$

A series is called $I(d)$ if it is non-stationary in levels but is appropriately made stationary by differencing d times.

Random walk is the key $I(1)$ process.

Other $I(1)$ processes are similar. Why?

The Beveridge-Nelson Decomposition

$$y_t \sim I(1) \implies y_t = x_t + z_t$$

x_t = random walk

z_t = covariance stationary

Hence the random walk is the key ingredient for all $I(1)$ processes.

The Beveridge-Nelson decomposition implies that shocks to any $I(1)$ process have some permanent effect, as with a random walk. But the effects are not *completely* permanent,

unless the process is a pure random walk.

$I(1)$ Processes and “Unit Roots”

Random walk is an $I(1)$ $AR(1)$ process:

$$y_t = y_{t-1} + \varepsilon_t$$

$$\underbrace{(1 - L)}_{\deg 1} y_t = \varepsilon_t$$

One (unit) root, $L = 1$

Δy_t is standard covariance-stationary WN

More general $I(1)$ $AR(p)$ process:

$$\underbrace{\Phi(L)}_{\deg p} y_t = \varepsilon_t$$

$$\underbrace{[\Phi'(L)(1 - L)]}_{(\deg p-1)(\deg 1)} y_t = \varepsilon_t$$

$p - 1$ stationary roots, one unit root

Δy_t is standard covariance stationary $AR(p - 1)$

Unit Root Distribution for the AR(1) Process

Key issue (hypothesis) in economics:

$I(1)$ vs. $I(0)$, unit root vs. stationary process

When $|\phi| < 1$,

$$\sqrt{T}(\hat{\phi}_{LS} - \phi) \xrightarrow{d} N$$

When $\phi = 1$,

$$T(\hat{\phi}_{LS} - 1) \xrightarrow{d} DF$$

Superconsistent

Nonstandard limiting distribution

Downward finite-sample bias (“Dickey-Fuller bias”)

Studentized Statistic

$$\hat{\tau} = \frac{\hat{\phi} - 1}{s \sqrt{\frac{1}{\sum y_{t-1}^2}}}$$

Not t in finite samples

Not $N(0, 1)$ asymptotically

Trick:

Don't run $y_t \rightarrow y_{t-1}$

Instead run $\Delta y_t \rightarrow y_{t-1}$

AR(1) With Nonzero Mean Under the Alternative

$$(y_t - \mu) = \phi(y_{t-1} - \mu) + \varepsilon_t$$

$$y_t = \alpha + \phi y_{t-1} + \varepsilon_t$$

$$\text{where } \alpha = \mu(1 - \phi)$$

Random walk null vs. mean-reverting alternative

Studentized statistic $\hat{\tau}_\mu$

AR(1) With Trend Under the Alternative

$$(y_t - a - bt) = \phi(y_{t-1} - a - b(t-1)) + \varepsilon_t$$

$$y_t = \alpha + \beta t + \phi y_{t-1} + \varepsilon_t$$

$$\text{where } \alpha = a(1 - \phi) + b\phi \text{ and } \beta = b(1 - \phi)$$

$$H_0 : \phi = 1 \text{ (unit root)}$$

$$H_1 : \phi < 1 \text{ (stationary root)}$$

Studentized statistic $\hat{\tau}_\tau$

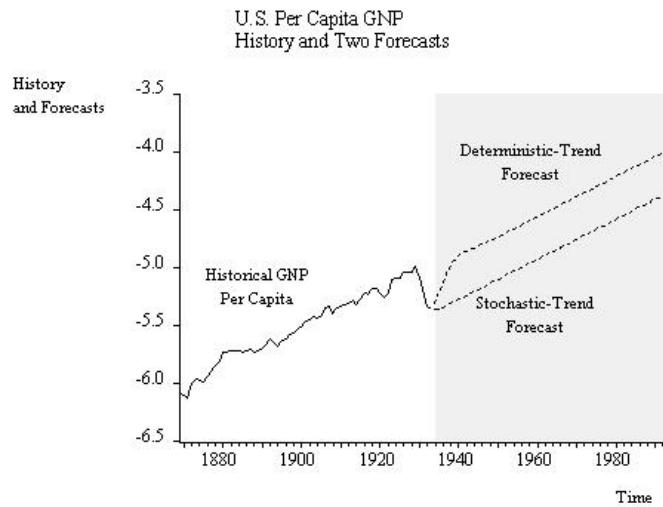
“Random walk with drift” vs. “stat. AR(1) around linear trend”

“Stochastic trend” vs. “deterministic trend”

Stochastic Trend vs. Deterministic Trend

$$AR(p)$$

$$y_t + \sum_{j=1}^p \phi_j y_{t-j} = \varepsilon_t$$



$$y_t = \rho_1 y_{t-1} + \sum_{j=2}^p \rho_j (y_{t-j+1} - y_{t-j}) + \varepsilon_t$$

where $p \geq 2$, $\rho_1 = -\sum_{j=1}^p \phi_j$, and $\rho_i = \sum_{j=i}^p \phi_j$, $i = 2, \dots, p$

Studentized statistic $\hat{\tau}$ is still relevant

$AR(p)$ With Nonzero Mean Under the Alternative

$$(y_t - \mu) + \sum_{j=1}^p \phi_j (y_{t-j} - \mu) = \varepsilon_t$$

$$y_t = \alpha + \rho_1 y_{t-1} + \sum_{j=2}^p \rho_j (y_{t-j+1} - y_{t-j}) + \varepsilon_t$$

$$\text{where } \alpha = \mu(1 + \sum_{j=1}^p \phi_j)$$

Studentized statistic $\hat{\tau}_\mu$ is still relevant

$AR(p)$ With Trend Under the Alternative

$$(y_t - a - bt) + \sum_{j=1}^p \phi_j(y_{t-j} - a - b(t-j)) = \varepsilon_t$$

$$\begin{aligned} y_t &= k_1 + k_2 t + \rho_1 y_{t-1} + \sum_{j=2}^p \rho_j(y_{t-j+1} - y_{t-j}) + \varepsilon_t \\ k_1 &= a \left(1 + \sum_{i=1}^p \phi_i \right) - b \sum_{i=1}^p i \phi_i \\ k_2 &= b \left(1 + \sum_{i=1}^p \phi_i \right) \end{aligned}$$

Under the null hypothesis, $k_1 = -b \sum_{i=1}^p i \phi_i$ and $k_2 = 0$

Studentized statistic $\hat{\tau}_\tau$ is still relevant

“Trick Form” of ADF in the General $AR(p)$ Case

$$(y_t - y_{t-1}) = (\rho_1 - 1)y_{t-1} + \sum_{j=2}^{k-1} \rho_j(y_{t-j+1} - y_{t-j}) + \varepsilon_t$$

- Unit root corresponds to $(\rho_1 - 1) = 0$
- Use standard automatically-computed t -statistic
(which of course does not have the t -distribution)

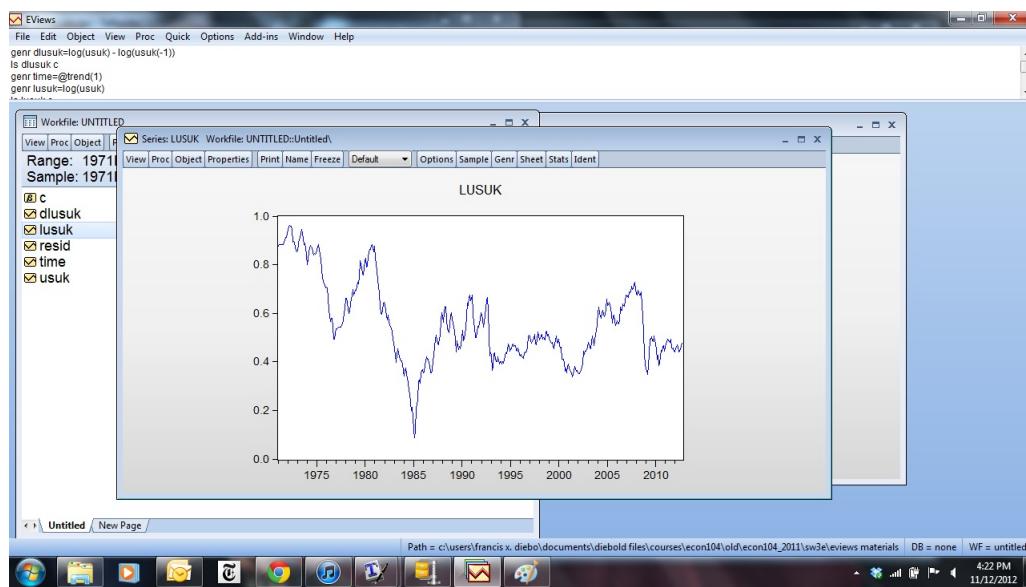
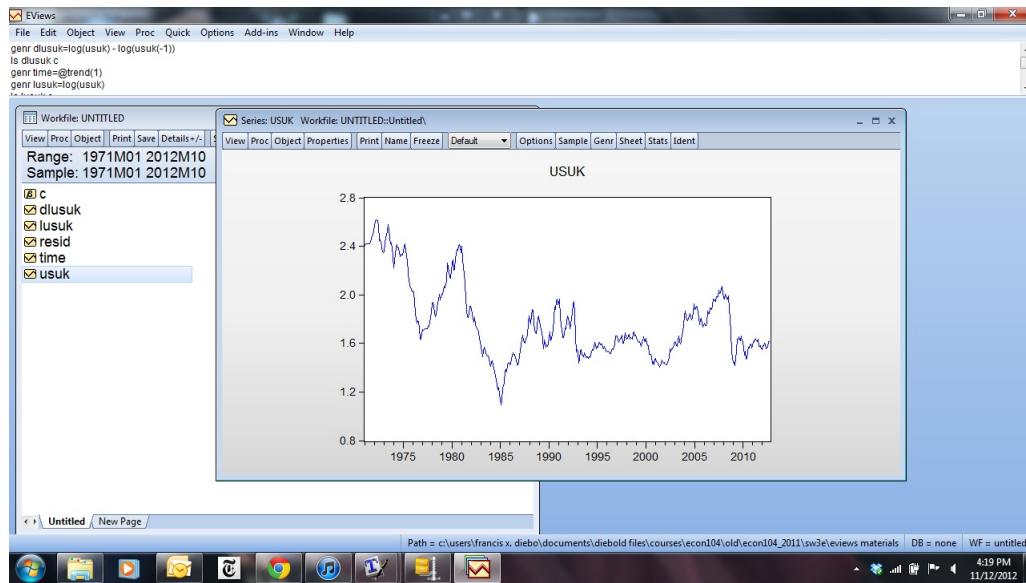
USD/GBP Exchange Rate, 1971.01-2012.10

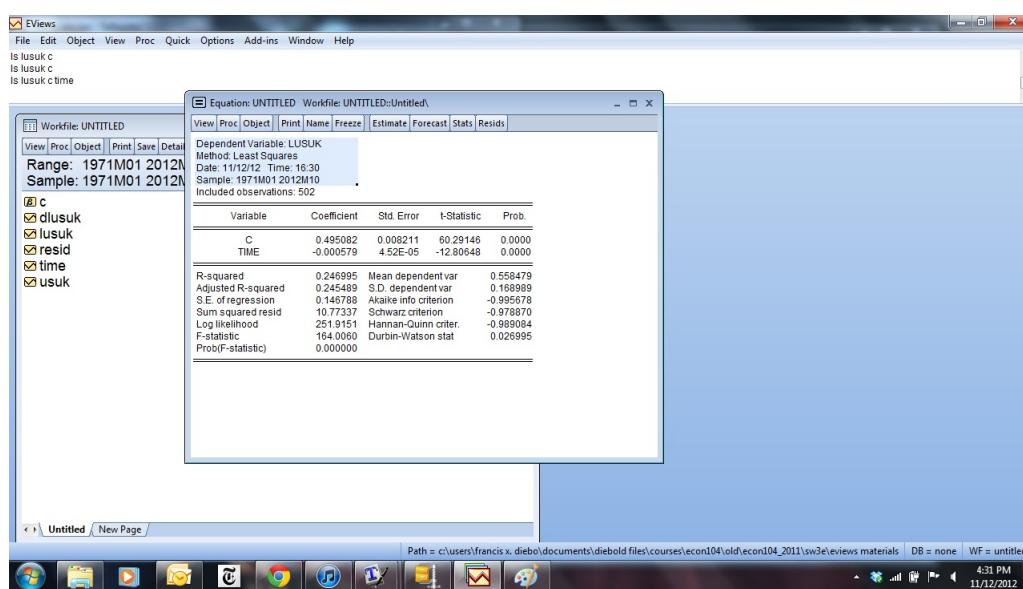
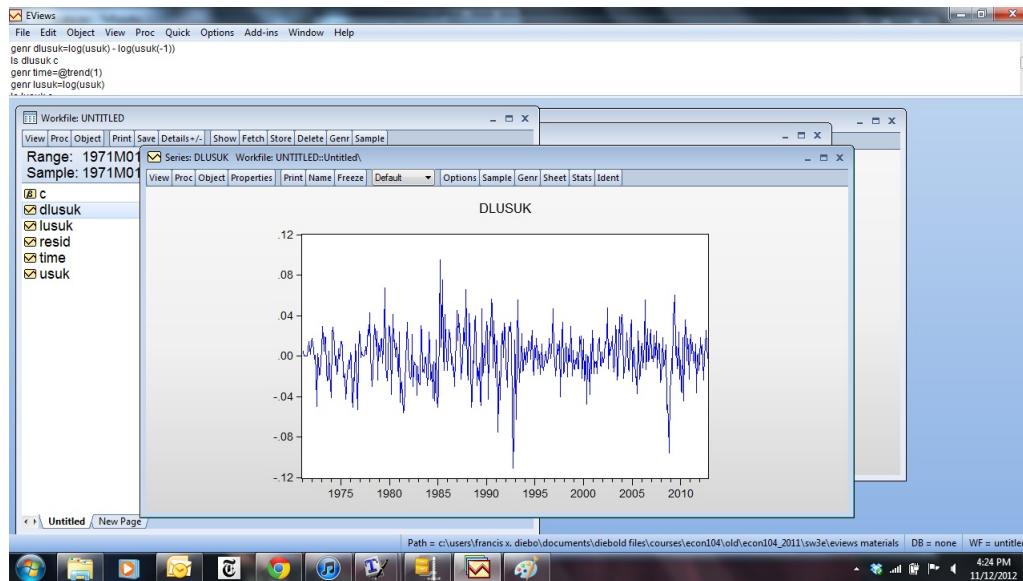
Log USD/GBP Exchange Rate, 1971.01-2012.10

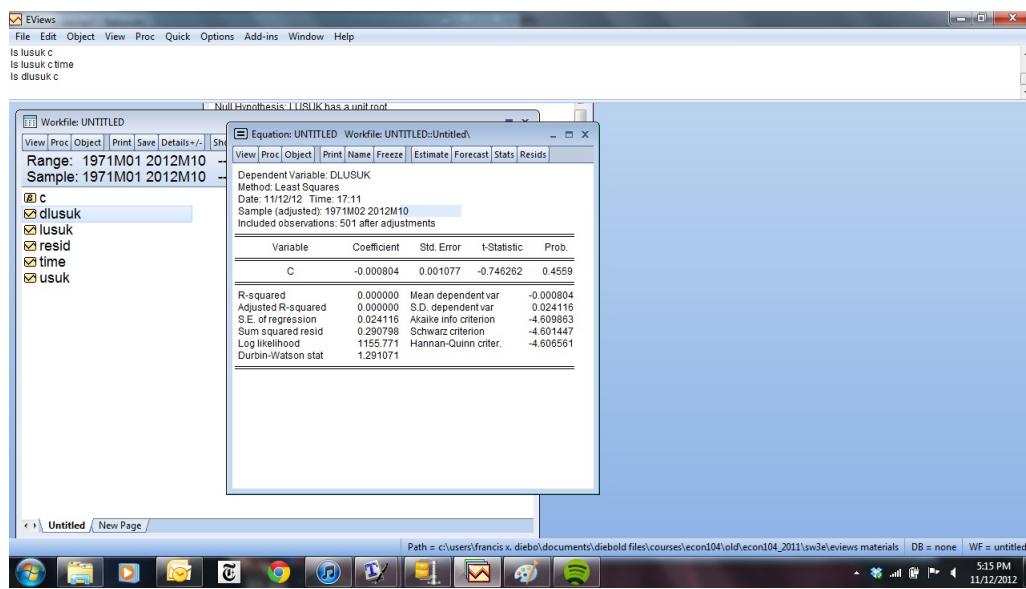
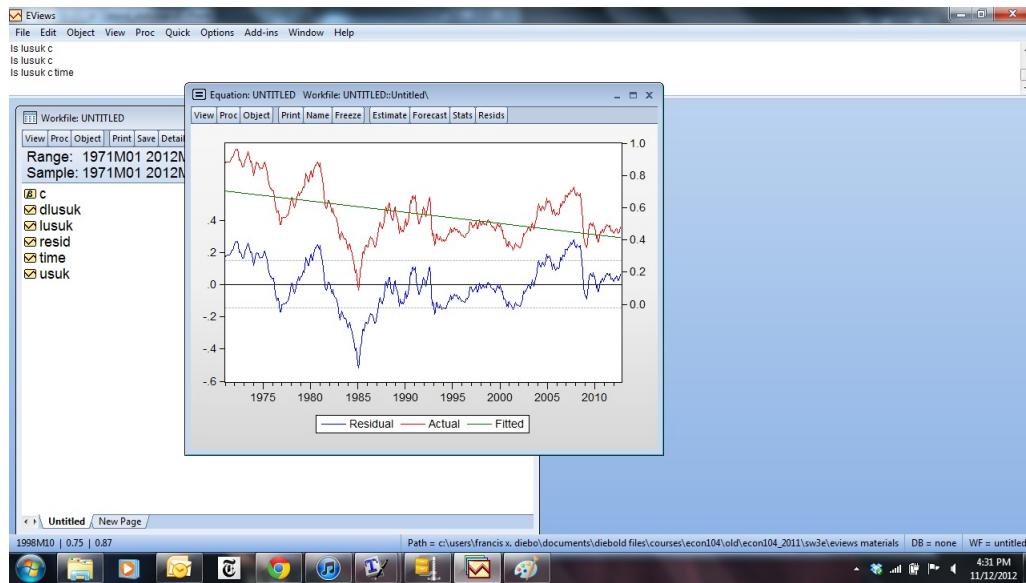
Change in USD/GBP Exchange Rate, 1971.01-2012.10

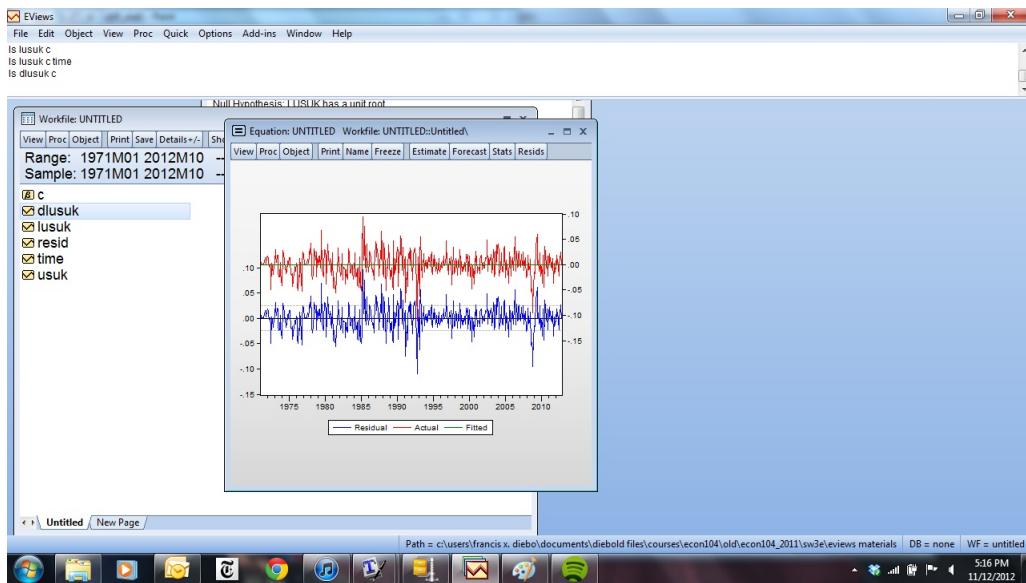
Trend-Stationary Model

Trend-Stationary Model









Difference-Stationary Model (Random Walk With Drift)

Difference-Stationary Model (Random Walk With Drift)

DF Tests – Option Screen

ADF Test, Allowing for Trend Under the Alternative

Multivariate Problem: Spurious Time-Series Regressions

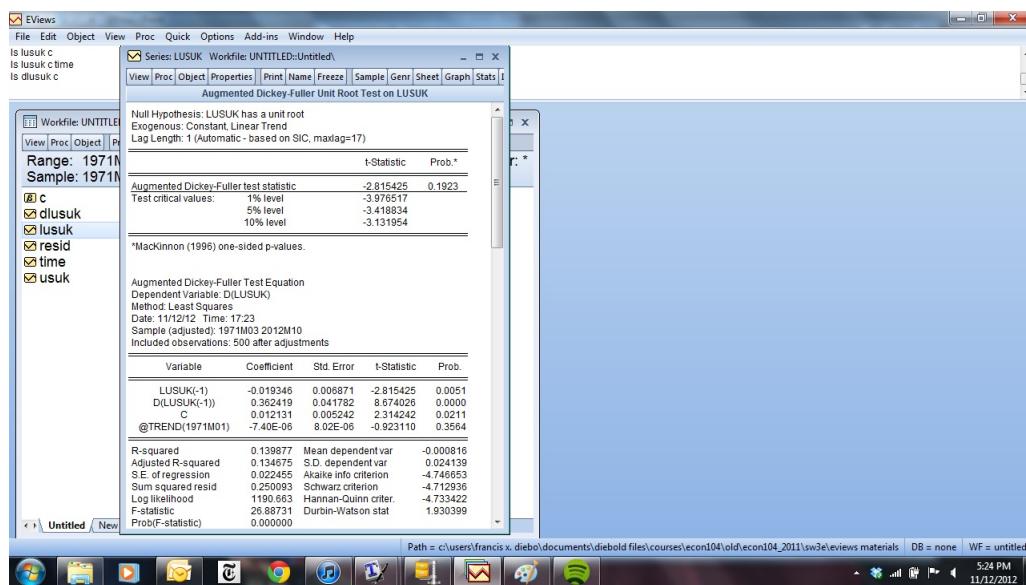
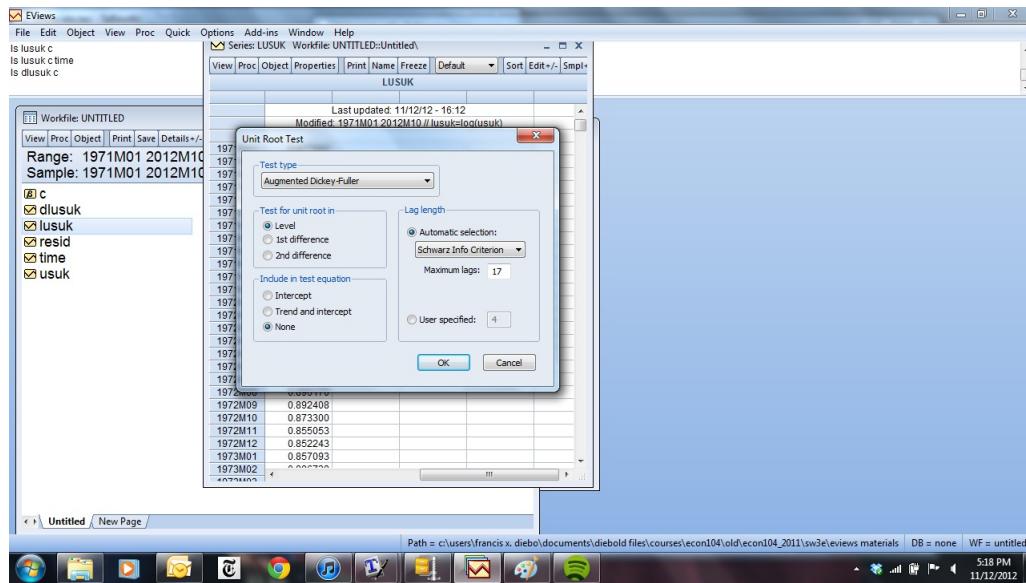
Regress a persistent variable on an *unrelated* persistent variable:

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

(Canonical case: y, x independent driftless random walks)

$$\frac{\hat{\beta}}{\sqrt{T}} \xrightarrow{d} RV \quad (\hat{\beta} \text{ diverges})$$

$$\frac{t}{\sqrt{T}} \xrightarrow{d} RV \quad (t \text{ diverges})$$



$$\begin{matrix} d \\ R^2 \rightarrow RV \text{ (not zero)} \end{matrix}$$

When are I(1) Levels Regressions *Not* Spurious?

Answer: When the variables are cointegrated.

Cointegration

Consider an N -dimensional variable x :

$$x \sim CI(d, b) \text{ if}$$

1. $x_i \sim I(d), i = 1, \dots, N$
2. $\exists 1$ or more linear combinations $z_t = \alpha' x_t$ s.t. $z_t \sim I(d - b), b > 0$

Leading Case

$$x \sim CI(1, 1) \text{ if}$$

$$(1) x_i \sim I(1), i = 1, \dots, N$$

$$(2) \exists 1 \text{ or more linear combinations}$$

$$z_t = \alpha' x_t \text{ s.t. } z_t \sim I(0)$$

Example

$$x_t = x_{t-1} + v_t, v_t \sim WN$$

$$y_t = x_{t-1} + \varepsilon_t, \varepsilon_t \sim WN, \varepsilon_t \perp v_{t-\tau}, \forall t, \tau$$

$$\Rightarrow (y_t - x_t) = \varepsilon_t - v_t = I(0)$$

Cointegration Motivation: Dynamic Factor Structure

Factor structure with $I(1)$ factors

$(N - R)$ $I(1)$ factors driving N variables

e.g., single-factor model:

$$\begin{pmatrix} y_{1t} \\ \vdots \\ y_{Nt} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} f_t + \begin{pmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{Nt} \end{pmatrix}$$

$$f_t = f_{t-1} + \eta_t$$

$R = (N - 1)$ cointegrating combs: $(y_{2t} - y_{1t}), \dots, (y_{Nt} - y_{1t})$

$(N - R) = N - (N - 1) = 1$ common trend

18.2 Exercises, Problems and Complements

1. (Using stochastic-trend unobserved-components models to implement smoothing techniques in a probabilistic framework)

In the text we noted that smoothing techniques, as typically implemented, are used as “black boxes” to produce point forecasts. There is no attempt to exploit stochastic structure to produce interval or density forecasts in addition to point forecasts. Recall, however, that the various smoothers produce optimal forecasts for specific data-generating processes specified as unobserved-components models.

- a. For what data-generating process is exponential smoothing optimal?
- b. For what data-generating process is Holt-Winters smoothing optimal?
- c. Under the assumption that the data-generating process for which exponential smoothing produces optimal forecasts is in fact the true

data-generating process, how might you estimate the unobserved-components model and use it to produce optimal interval and density forecasts? Hint: Browse through Harvey (1989).

- d. How would you interpret the interval and density forecasts produced by the method of part c, if we no longer assume a particular model for the true data-generating process?
- 2. (The Dickey-Fuller regression in the $AR(2)$ case)

Consider the $AR(2)$ process,

$$y_t + \phi_1 y_{t-1} + \phi_2 y_{t-2} = \varepsilon_t.$$

- a. Show that it can be written as

$$y_t = \rho_1 y_{t-1} + \rho_2 (y_{t-1} - y_{t-2}) + \varepsilon_t,$$

where

$$\rho_1 = -(\phi_1 + \phi_2)$$

$$\rho_2 = \phi_2.$$

- b. Show that it can also be written as a regression of Δy_t on y_{t-1} and Δy_{t-1} .
- c. Show that if $\rho_1 = 1$, the $AR(2)$ process is really an $AR(1)$ process in first differences; that is, the $AR(2)$ process has a unit root.
- 3. (Holt-Winters smoothing with multiplicative seasonality)

Consider a seasonal Holt-Winters smoother, written as

- (1) Initialize at $t = s$:

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

$$T_s = 0$$

$$F_j = \frac{y_j}{\left(\frac{1}{s} \sum_{t=1}^s y_t\right)}, j = 1, 2, \dots, s$$

(2) (2) Update:

$$\bar{y}_t = \alpha \left(\frac{y_t}{F_{t-s}} \right) + (1 - \alpha) (\bar{y}_{t-1} + T_{t-1}), 0 < \alpha < 1$$

$$T_t = \beta (\bar{y}_t - \bar{y}_{t-1}) + (1 - \beta) T_{t-1}, 0 < \beta < 1$$

$$F_t = \gamma \left(\frac{y_t}{\bar{y}_t} \right) + (1 - \gamma) F_{t-s}, 0 < \gamma < 1$$

t = s+1, ..., T.

(3) Forecast:

$$\hat{y}_{T+h,T} = (\bar{y}_T + hT_T) F_{T+h-s}, h = 1, 2, \dots, s,$$

$$\hat{y}_{T+h,T} = (\bar{y}_T + hT_T) F_{T+h-2s}, h = s+1, s+2, \dots, 2s,$$

etc.

- a. The Holt-Winters seasonal smoothing algorithm given in the text is more precisely called Holt-Winters seasonal smoothing with additive seasonality. The algorithm given above, in contrast, is called Holt-Winters seasonal smoothing with multiplicative seasonality. How does this algorithm differ from the one given in the text, and what, if anything, is the significance of the difference?
 - b. Assess the claim that Holt-Winters with multiplicative seasonality is appropriate when the seasonal pattern exhibits increasing variation.
 - c. How does Holt-Winters with multiplicative seasonality compare with the use of Holt-Winters with additive seasonality applied to logarithms of the original data?
4. **(Cointegration)**

Consider two series, x and y , both of which are $I(1)$. In general there is no way to form a weighted average of x and y to produce an $I(0)$ series, but in the very special case where such a weighting does exist, we say that x and y are cointegrated. Cointegration corresponds to situations in which variables tend to cling to one another, in the sense that the cointegrating combination is stationary, even though each variable is nonstationary. Such situations arise frequently in business, economics, and finance. To take a business example, it's often the case that both inventories and sales of a product appear $I(1)$, yet their ratio (or, when working in logs, their difference) appears $I(0)$, a natural byproduct of various schemes that adjust inventories to sales. Engle and Granger (1987) is the key early research paper on cointegration; Johansen (1995) surveys most of the more recent developments, with emphasis on maximum likelihood estimation.

- a. Consider the bivariate system,

$$x_t = x_{t-1} + v_t, v_{tWN}(0, \sigma^2)$$

$$y_t = x_t + \varepsilon_t, \varepsilon_{tWN}(0, \sigma^2).$$

Both x and y are $I(1)$. Why? Show, in addition, that x and y are cointegrated. What is the cointegrating combination?

- b. Engle and Yoo (1987) show that optimal long-run forecasts of cointegrated variables obey the cointegrating relationship exactly. Verify their result for the system at hand.

5. (Error-correction)

In an error-correction model, we take a long-run model relating $I(1)$ variables, and we augment it with short-run dynamics. Suppose, for example, that in long-run equilibrium y and x are related by $y = bx$.

Then the deviation from equilibrium is $z = y - bx$, and the deviation from equilibrium at any time may influence the future evolution of the variables, which we acknowledge by modeling Δx as a function of lagged values of itself, lagged values of Δy , *and the lagged value of z, the error-correction term.* For example, allowing for one lag of Δx and one lag of Δy on the right side, we write equation for x as

$$\Delta x_t = \alpha_x \Delta x_{t-1} + \beta_x \Delta y_{t-1} + \gamma_x z_{t-1} + \varepsilon_{xt}.$$

Similarly, the y equation is

$$\Delta y_t = \alpha_y \Delta x_{t-1} + \beta_y \Delta y_{t-1} + \gamma_y z_{t-1} + \varepsilon_{yt}.$$

So long as one or both of γ_x and γ_y are nonzero, the system is very different from a *VAR* in first differences; the key feature that distinguishes the error-correction system from a simple *VAR* in first differences is the inclusion of the error-correction term, so that the deviation from equilibrium affects the evolution of the system.

- a. Engle and Granger (1987) establish the key result that existence of cointegration in a *VAR* and existence of error-correction are equivalent – a *VAR* is cointegrated if and only if it has an error-correction representation. Try to sketch some intuition as to why the two should be linked. Why, in particular, might cointegration imply error correction?
- b. Why are cointegration and error correction of interest to forecasters in business, finance, economics and government?
- c. Evaluation of forecasts of cointegrated series poses special challenges, insofar as traditional accuracy measures don't value the preservation of cointegrating relationships, whereas presumably they *should*. For details and constructive suggestions, see Christoffersen and Diebold

(1998).

6. (Evaluating forecasts of integrated series) The unforecastability principle remains intact regardless of whether the series being forecast is stationary or integrated: the errors from optimal forecasts are not predictable on the basis of information available at the time the forecast was made. However, some additional implications of the unforecastability principle emerge in the case of forecasting $I(1)$ series, including:
 - a. If the series being forecast is $I(1)$, then so too is the optimal forecast.
 - b. An $I(1)$ series is always cointegrated with its optimal forecast, which means that there exists an $I(0)$ linear combination of the series and its optimal forecast, in spite of the fact that both the series and the forecast are $I(1)$.
 - c. The cointegrating combination is simply the difference of the actual and forecasted values – the forecast error. Thus the error corresponding to an optimal forecast of an $I(1)$ series is $I(0)$, in spite of the fact that the series is not.

Cheung and Chinn (1999) make good use of these results in a study of the information content of U.S. macroeconomic forecasts; try to sketch their intuition. (Hint: Suppose the error in forecasting an $I(1)$ series were *not* $I(0)$. What would that imply?)

7. (Spurious regression)

Consider two variables y and x , both of which are highly serially correlated, as are most series in business, finance and economics. Suppose in addition that y and x are completely unrelated, but that we don't know they're unrelated, and we regress y on x using ordinary least squares.

- a. If the usual regression diagnostics (e.g., R^2 , t -statistics, F -statistic) were reliable, we'd expect to see small values of all of them. Why?

- b. In fact the opposite occurs; we tend to see large R^2 , t -, and F -statistics, and *a very low Durbin-Watson statistic*. Why the low Durbin-Watson? Why, given the low Durbin-Watson, might you *expect* misleading R^2 , t -, and F -statistics?
- c. This situation, in which highly persistent series that are in fact unrelated nevertheless appear highly related, is called spurious regression. Study of the phenomenon dates to the early twentieth century, and a key study by Granger and Newbold (1974) drove home the prevalence and potential severity of the problem. How might you insure yourself against the spurious regression problem? (Hint: Consider allowing for lagged dependent variables, or dynamics in the regression disturbances, as we've advocated repeatedly.)

8. (Theil's U -statistic)

Sometimes it's informative to compare the accuracy of a forecast to that of a "naive" competitor. A simple and popular such comparison is achieved by the U statistic, which is the ratio of the 1-step-ahead MSE for a given forecast relative to that of a random walk forecast

$$y_{t+1,t} = y_t;$$

that is,

$$U = \frac{\sum_{t=1}^T (y_{t+1} - y_{t+1,t})^2}{\sum_{t=1}^T (y_{t+1} - y_t)^2}.$$

One must remember, of course, that the random walk is not necessarily a naive competitor, particularly for many economic and financial variables, so that values of U near one are not necessarily "bad."

The U -statistic is due to Theil (1966, p. 28), and is often called "Theil's U-statistic."

18.3 Notes

Chapter 19

Big Data: Selection, Shrinkage and Distillation

Often we have many possible RHS variables, and hence many models that could be entertained. How do we select? 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 econometrics, so we start with it. Then we move to shrinkage and distillation.

19.1 Good and Bad Approaches to Model Selection

19.1.1 MSE and R^2

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

Selection by MSE (or R^2) produces in-sample over-fitting

19.1.2 s^2 and \bar{R}^2

$$\begin{aligned}s^2 &= \frac{1}{T-K} \sum_{t=1}^T e_t^2 = \left(\frac{T}{T-K} \right) \frac{\sum_{t=1}^T e_t^2}{T} \\ \bar{R}^2 &= 1 - \frac{\frac{1}{T-K} \sum_{t=1}^T e_t^2}{\frac{1}{T-1} \sum_{t=1}^T (y_t - \bar{y})^2} = 1 - \frac{s^2}{\frac{1}{T-1} \sum_{t=1}^T (y_t - \bar{y})^2}\end{aligned}$$

Selection by s^2 (or \bar{R}^2) *still* produces in-sample over-fitting

Thus far we've mentioned only poor approaches to model selection. We now turn to better approaches, which are closely related, but nevertheless different in important ways.

19.2 All Subsets Selection

19.2.1 Information Criteria

$$SIC = -2 \ln(L) + K \ln(T)$$

In the Gaussian linear regression case SIC becomes:

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

“Oracle property”

No over-fitting (asymptotically)! Other procedures like “select the model with the highest R^2 have no such property and are not to be recommended!

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

In the Gaussian linear regression case AIC becomes:

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

“Efficiency”

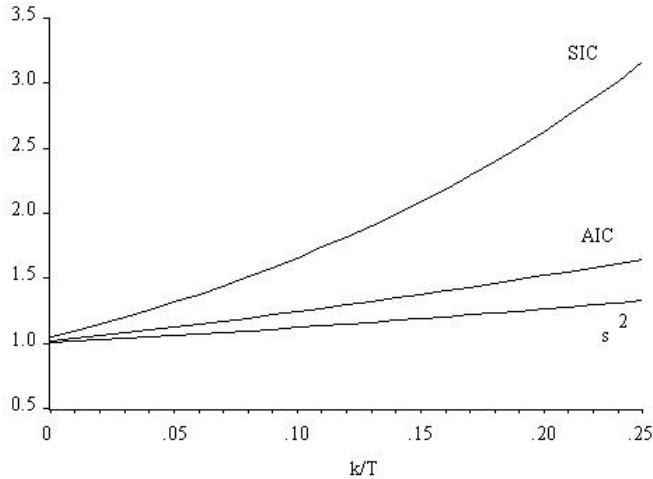


Figure 19.1: Degrees-of-Freedom Penalties

19.2.2 Cross Validation

Cross validation (CV) proceeds as follows. Consider selecting among J models. Start with model 1, estimate it using all data observations except the first, use it to predict the first observation, and compute the associated squared prediction error. Then estimate it using all observations except the second, use it to predict the second observation, and compute the associated squared error. Keep doing this – estimating the model with one observation deleted and then using the estimated model to predict the deleted observation – until each observation has been sequentially deleted, and average the squared errors in predicting each of the T sequentially deleted observations. Repeat the procedure for the other models, $j = 2, \dots, J$, and select the model with the smallest average squared prediction error.

Actually this is “ T -fold” CV, because we split the data into T parts (the T individual observations) and predict each of them. More generally we can split the data into M parts ($M < T$) and cross validate on them (“ M -fold” CV). As M falls, M -fold CV eventually becomes consistent. $M = 10$ often works well in practice. Generalizations to time-series contexts are available.

CV is more general than information criteria insofar as it can be used even when the model degrees of freedom is unclear. This will be important.

19.3 Stepwise Selection

All-subsets selection, whether by AIC, SIC or CV, quickly gets hard as there are 2^K subsets of K regressors.

19.3.1 Backward Stepwise Selection

Algorithm:

- Start with a regression that includes all K variables
- Move to a $K - 1$ variable model by dropping the variable with the largest p -value
 - Move to a $K - 2$ variable model by dropping the variable with the largest p -value
 - Continue until a stopping criterion is satisfied
 - Often people use information criteria or CV to select from the stepwise sequence of models.

This is a “greedy algorithm,” producing an decreasing sequence of candidate models. No optimality properties of the selected model.

19.3.2 Forward Stepwise Selection

Algorithm:

- Begin regressing only on an intercept
- Move to a one-regressor model by including that variable with the lowest p -value
 - Move to a two-regressor model by including that variable with the lowest p -value

- Move to a three-regressor model by including that variable with the lowest p -value

- Continue until a stopping criterion is satisfied

This is a “greedy algorithm,” producing an increasing sequence of candidate models. No 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.

- Note that forward selection can be done even if $K > T$, unlike backward selection.

19.4 Shrinkage

19.4.1 Bayesian Shrinkage

$$\hat{\beta}_{bayes} = \omega_1 \hat{\beta}_{MLE} + \omega_2 \beta_0,$$

where β_0 is the prior mean. The estimator is pulled, or “shrunken,” toward the prior mean. The weights depend on prior precision.

19.4.2 Ridge Shrinkage

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

λ can be chosen by CV.

19.5 Shrinkage and Selection

Consider the penalized estimator,

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

or equivalently

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

s.t.

$$\sum_{i=1}^K |\beta_i|^q \leq c.$$

Concave penalty functions non-differentiable at the origin produce selection. Smooth convex penalties produce shrinkage. Indeed one can show that taking $q \rightarrow 0$ produces subset selection, and taking $q = 2$ produces ridge regression. Hence penalized estimation nests those situations and includes an intermediate case ($q = 1$) that produces the lasso, to which we now turn.

19.5.1 The Lasso

The lasso solves the L1-penalized regression problem of finding

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

or equivalently

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

s.t.

$$\sum_{i=1}^K |\beta_i| \leq c.$$

Ridge shrinks, but the lasso shrinks *and* selects. It uses the smallest q for which the minimization problem is convex, which is highly valuable 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 q . 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).

19.5.2 Elastic Net

$$\hat{\beta}_{EN} = \operatorname{argmin}_{\beta} \left(\sum_{t=1}^T \left(y_t - \sum_i \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^K (\alpha |\beta_i| + (1-\alpha) \beta_i^2) \right)$$

- A mixture of Lasso and Ridge regression; that is, it combines L1 and L2 penalties.
- Unlike Lasso, it moves strongly correlated predictors in or out of the model together, hopefully producing improving prediction accuracy relative to Lasso.
- Unlike Lasso, there are two tuning parameters in the elastic net λ and α . For $\alpha = 1$ elastic net turns into a Lasso model, For $\alpha = 0$ it is equivalent to ridge regression.

19.5.3 Adaptive Lasso

$$\hat{\beta}_{ALASSO} = \operatorname{argmin}_{\beta} \left(\sum_{t=1}^T \left(y_t - \sum_i \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^K w_i |\beta_i| \right),$$

where $w_i = 1/\hat{\beta}_i^\nu$, $\hat{\beta}_i$ is the OLS estimate, and $\nu > 0$.

- Every parameter in the penalty function is weighted differently, in contrast to the “regular” Lasso.
- The weights are calculated by OLS.
- Oracle property.

19.5.4 Adaptive Elastic Net

$$\hat{\beta}_{AEN} = \operatorname{argmin}_{\beta} \left(\sum_{t=1}^T \left(y_t - \sum_i \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^K (\alpha w_i |\beta_i| + (1-\alpha) \beta_i^2) \right),$$

where $w_i = 1/\hat{\beta}_i^\nu$, $\hat{\beta}_i$ is the OLS estimate, and $\nu > 0$.

- A combination of elastic net and adaptive Lasso.
- Oracle property.

19.6 Distillation: Principal Components

19.6.1 Distilling “ X Variables” into Principal Components

Data Summarization. Think of a giant (wide) X matrix and how to “distill” it.

$X'X$ eigen-decomposition:

$$X'X = VD^2V'$$

The j^{th} column of V , v_j , is the j^{th} eigenvector of $X'X$

Diagonal matrix D^2 contains the descending eigenvalues of $X'X$

First principal component (PC):

$$z_1 = Xv_1$$

$$\operatorname{var}(z_1) = d_1^2/T$$

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

In general:

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

$$\operatorname{var}(z_j) \leq d_j^2/T$$

19.6.2 Principal Components Regression

“Principal components regression” (PCR), or “Factor-Augmented Regression” (FAVAR)

Ridge 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).

19.7 Exercises, Problems and Complements

1. “Data-rich” environments.

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

19.8 Notes

Part I

Appendices

Appendix A

Elements of Probability and Statistics

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 $y = 0$, $f(y) = 0.50$ 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-

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.

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.

Often we assess dispersion using the square root of the variance, which is

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.

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,

$$\text{cov}(y, x) = E[(y_t - \mu_y)(x_t - \mu_x)].$$

A positive covariance means that y and x are positively related; that is, when y is above its mean x tends to be above its mean, and when y is below its mean x tends to be below its mean. Conversely, a negative covariance means that y and x are inversely related; that is, when y is below its mean x tends to be above its mean, and vice versa. The covariance can take any value in the real numbers.

Frequently we convert the covariance to a **correlation** by standardizing

⁵We could of course consider more than two variables, but for pedagogical reasons we presently limit ourselves to two.

by the product of σ_y and σ_x ,

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

The correlation takes values in $[-1, 1]$. Note that covariance depends on units of measurement (e.g., dollars, cents, billions of dollars), but correlation does not. Hence correlation is more immediately interpretable, which is the reason for its popularity.

Note also that covariance and correlation measure only *linear* dependence; in particular, a zero covariance or correlation between y and x does not necessarily imply that y and x are independent. That is, they may be *non-linearly* related. If, however, two random variables are jointly *normally* distributed with zero covariance, then they are independent.

Our multivariate discussion has focused on the joint distribution $f(y, x)$. In various chapters we will also make heavy use of the **conditional distribution** $f(y|x)$, that is, the distribution of the random variable Y *conditional* upon $X = x$. **Conditional moments** are similarly important. In particular, the **conditional mean** and **conditional variance** play key roles in econometrics, in which attention often centers on the mean or variance of a series conditional upon the past.

A.2 Samples: Sample Moments

A.2.1 Univariate

Thus far we've reviewed aspects of known distributions of random variables, in **population**. Often, however, we have a **sample** of data drawn from an unknown population distribution f ,

$$\{y_i\}_{i=1}^N \sim f(y),$$

and we want to learn from the sample about various aspects of f , such as its moments. To do so we use various **estimators**.⁶ We can obtain estimators by replacing population expectations with sample averages, because the arithmetic average is the sample analog of the population expectation. Such “analog estimators” turn out to have good properties quite generally. The **sample mean** is simply the arithmetic average,

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

It provides an empirical measure of the location of y .

The **sample variance** is the average squared deviation from the sample mean,

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

It provides an empirical measure of the dispersion of y around its mean.

We commonly use a slightly different version of $\hat{\sigma}^2$, which corrects for the one degree of freedom used in the estimation of \bar{y} , thereby producing an unbiased estimator of 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 .

⁶An estimator is an example of a **statistic**, or **sample statistic**, which is simply a function of the sample observations.

The **sample skewness** is

$$\hat{S} = \frac{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^3}{\hat{\sigma}^3}.$$

It provides an empirical measure of the amount of asymmetry in the distribution of y .

The **sample kurtosis** is

$$\hat{K} = \frac{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^4}{\hat{\sigma}^4}.$$

It provides an empirical measure of the fatness of the tails of the distribution of y relative to a normal distribution.

Many of the most famous and important statistical sampling distributions arise in the context of sample moments, and the normal distribution is the father of them all. In particular, the celebrated central limit theorem establishes that under quite general conditions the sample mean \bar{y} will have a normal distribution as the sample size gets large. The 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**,

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

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 indeed the minimum variance unbiased (MVUE) estimator. We write

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

or equivalently

$$\sqrt{N}(\bar{y} - \mu) \sim N(0, \sigma^2).$$

We construct exact finite-sample confidence intervals for μ as

$$\mu \in \left[\bar{y} \pm t_{1-\frac{\alpha}{2}}(N-1) \frac{s}{\sqrt{N}} \right] \text{ w.p. } \alpha,$$

where $t_{1-\frac{\alpha}{2}}(N-1)$ is the $1 - \frac{\alpha}{2}$ percentile of a t distribution with $N-1$ degrees of freedom. Similarly, we construct exact finite-sample (likelihood ratio) hypothesis tests of $H_0 : \mu = \mu_0$ against the two-sided alternative $H_0 : \mu \neq \mu_0$ using

$$\frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{N}}} \sim t_{1-\frac{\alpha}{2}}(N-1).$$

A.3.2 Approximate Asymptotic Results (Under Weaker Assumptions)

Much of statistical inference is linked to large-sample considerations, such as the law of large numbers and the central limit theorem, which you also studied earlier. Here we again refresh your memory.

Consider again a simple random sample, but without the normality assumption,

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

Despite our dropping the normality assumption we still have that \bar{y} is unbiased, consistent, **asymptotically** normally distributed with variance σ^2/N , and best linear unbiased (BLUE). We write,

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

More precisely, as $T \rightarrow \infty$,

$$\sqrt{N}(\bar{y} - \mu) \rightarrow_d N(0, \sigma^2).$$

This result forms the basis for asymptotic inference. It is a Gaussian central limit theorem, and it also has a law of large numbers ($\bar{y} \rightarrow_p \mu$) imbedded within it.

We construct asymptotically-valid confidence intervals for μ as

$$\mu \in \left[\bar{y} \pm z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{N}} \right] \text{ w.p. } \alpha,$$

where $z_{1-\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ percentile of a $N(0, 1)$ distribution. Similarly, we construct asymptotically-valid hypothesis tests of $H_0 : \mu = \mu_0$ against the two-sided alternative $H_0 : \mu \neq \mu_0$ using

$$\frac{\bar{y} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{N}}} \sim N(0, 1).$$

A.4 Exercises, Problems and Complements

1. (Interpreting distributions and densities)

The Sharpe Pencil Company has a strict quality control monitoring program. As part of that program, it has determined that the distribution of the amount of graphite in each batch of one hundred pencil leads produced is continuous and uniform between one and two grams. That is, $f(y) = 1$ for y in $[1, 2]$, and zero otherwise, where y is the graphite content per batch of one hundred leads.

- a. Is y a discrete or continuous random variable?
- b. Is $f(y)$ a probability distribution or a density?
- c. What is the probability that y is between 1 and 2? Between 1 and 1.3? Exactly equal to 1.67?
- d. For high-quality pencils, the desired graphite content per batch is 1.8 grams, with low variation across batches. With that in mind, discuss the nature of the density $f(y)$.

2. (Covariance and correlation)

Suppose that the annual revenues of world's two top oil producers have a covariance of 1,735,492.

- a. Based on the covariance, the claim is made that the revenues are "very strongly positively related." Evaluate the claim.
- b. Suppose instead that, again based on the covariance, the claim is made that the revenues are "positively related." Evaluate the claim.
- c. Suppose you learn that the revenues have a *correlation* of 0.93. In light of that new information, re-evaluate the claims in parts a and b above.

3. (Simulation)

You will often need to simulate data from various models. The simplest model is the $iidN(\mu, \sigma^2)$ (Gaussian simple random sampling) model.

- a. Using a random number generator, simulate a sample of size 30 for y , where $y \sim iidN(0, 1)$.
- b. What is the sample mean? Sample standard deviation? Sample skewness? Sample kurtosis? Discuss.
- c. Form an appropriate 95 percent confidence interval for $E(y)$.
- d. Perform a t test of the hypothesis that $E(y) = 0$.
- e. Perform a t test of the hypothesis that $E(y) = 1$.

4. (Sample moments of the wage data)

Use the 1995 wage dataset.

- a. Calculate the sample mean wage and test the hypothesis that it equals \$9/hour.
- b. Calculate sample skewness.

- c. Calculate and discuss the sample correlation between wage and years of education.
5. Notation.

We have used standard cross-section notation: $i = 1, \dots, N$. The standard time-series notation is $t = 1, \dots, T$. Much of our discussion will be valid in *both* cross-section and time-series environments, but still we have to pick a notation. Without loss of generality, henceforth we will typically use $t = 1, \dots, T$.

A.5 Notes

Numerous good introductory probability and statistics books exist. [Wonnacott and Wonnacott \(1990\)](#) remains a time-honored classic, which you may wish to consult to refresh your memory on statistical distributions, estimation and hypothesis testing. [Anderson et al. \(2008\)](#) is a well-written recent text.

Appendix B

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 12 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

- Anderson, D.R., D.J. Sweeney, and T.A. Williams (2008), *Statistics for Business and Economics*, South-Western.
- Angrist, J.D. and J.-S. Pischke (2009), *Mostly Harmless Econometrics*, Princeton University Press.
- Gladwell, M. (2000), *The Tipping Point*, Little, Brown and Company.
- Granger, C.W.J. (1969), “Investigating Causal Relations by Econometric Models and Cross-Spectral Methods,” *Econometrica*, 37, 424–438.
- Harvey, A.C. (1991), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press.
- Jarque, C.M. and A.K. Bera (1987), “A Test for Normality of Observations and Regression Residuals,” *International Statistical Review*, 55, 163–172.
- Koenker, R. (2005), *Quantile Regression*, Econometric Society Monograph Series, Cambridge University Press, 2005.
- 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.

Wonnacott, T.H. and R.J. Wonnacott (1990), *Introductory Statistics*. New York: John Wiley and Sons, Fifth Edition.

Index

- F distribution, 34
 F -statistic, 54
 R -squared, 55
 s -squared, 55
 t distribution, 34
 t -statistic, 51
 χ^2 distribution, 34
Fitted values, 44
Holiday variation, 77
Seasonal dummy variables, 75
Adjusted R -squared, 56
Akaike information criterion, 56
Analog principle, 161
Analysis of variance, 80
AR(p) process, 171
ARCH(p) process, 289
Aspect ratio, 19
Asymmetric response, 295
Asymmetry, 31
Asymptotic, 35
Autocorrelation function, 156
Autocovariance function, 154
Autoregressions, 156
Autoregressive (AR) model, 166
Banking to 45 degrees, 19
Binary data, 5
binomial logit, 336
Box-Cox transformation, 109
Box-Pierce Q-statistic, 163
Breusch-Godfrey test, 178
Calendar effects, 76
Central tendency, 30
Chartjunk, 19
Cointegration, 280
Common scales, 19
Conditional distribution, 32
Conditional expectation, 47
Conditional mean, 32
Conditional mean and variance, 160
Conditional mean function, 97
Conditional moment, 32
Conditional variance, 32
Constant term, 51
Continuous data, 5
Continuous random variable, 30
Correlation, 32
Correlogram, 162
Correlogram analysis, 164

- Covariance, 31
- Covariance stationary, 154
- Cross correlation function, 242
- Cross sectional data, 5
- Cross sections, 6
- Cross-variable dynamics, 235
- CUSUM, 137
- CUSUM plot, 137
- Cycles, 153
- Data mining, 57
- Data-generating process (DGP), 48
- De-trending, 81
- Deterministic seasonality, 74
- Deterministic trend, 72, 249
- Dickey-Fuller distribution, 255
- Discrete probability distribution, 29
- Discrete random variable, 29
- Dispersion, 30
- Distributed lag, 165
- Distributed lag model, 222
- Distributed lag regression model with lagged dependent variables, 233
- Distributed-lag regression model with *AR* disturbances, 233
- Disturbance, 47
- Dummy left-hand-side variable, 331
- Dummy right-hand-side variable, 331
- Dummy variable, 69
- Durbin's *h* test, 191
- Durbin-Watson statistic, 57
- Econometric modeling, 3
- Error-correction, 281
- Estimator, 33
- Ex post smoothing, 119
- Expected value, 30
- Exploratory data analysis, 23
- Exponential GARCH, 296
- Exponential smoothing, 261
- Exponential trend, 115
- Exponentially weighted moving average, 261
- Feedback, 243
- Financial econometrics, 286
- First-order serial correlation, 178
- Fourier series expansions, 123
- Functional form, 107
- GARCH(p,q) process, 290
- Gaussian distribution, 31
- Gaussian white noise, 158
- Generalized linear model, 109, 337
- GLM, 109, 337
- Golden ratio, 23
- Goodness of fit, 56
- Heteroskedasticity, 285
- Histogram, 15
- Hodrick-Prescott filtering, 119
- Holt-Winters Smoothing, 263

- Holt-Winters Smoothing with Seasonality, 264
- Impulse-response function, 238
- In-sample overfitting, 57
- Independent white noise, 158
- Indicator variable, 69, 331
- Innovation outliers, 190
- Instrumental variables, 349
- Integrated, 248
- Interaction effects, 111, 122
- Intercept, 72
- Intercept dummies, 70
- Interval data, 9
- Intrinsically non-linear models, 110
- Jarque-Bera test, 100
- Joint distribution, 31
- Kurtosis, 31
- Lag operator, 165
- Least absolute deviations, 103
- Least squares, 41
- Leptokurtosis, 31
- Likelihood function, 53
- Limited dependent variable, 331
- Linear probability model, 332
- Linear projection, 97, 343
- Linear trend, 72
- Link function, 109, 338
- Ljung-Box Q-statistic, 163
- Location, 30
- Log-lin regression, 108, 109
- Log-linear trend, 115
- Log-log regression, 107
- Logistic function, 332
- Logistic model, 110
- Logistic trend, 130
- Logit model, 332
- Marginal distribution, 31
- Markov-switching model, 140
- Maximum likelihood estimation, 53
- Mean, 30
- Measurement outliers, 190
- Model selection, 319
- Moments, 30, 160
- Multinomial logit, 338
- Multiple comparisons, 18
- Multiple linear regression, 44
- Multivariate, 13
- Multivariate GARCH, 317
- Multiway scatterplot, 16
- Neural networks, 126
- Nominal data, 9
- Non-data ink, 19
- non-linear least squares (NLS), 110
- Non-linearity, 97
- Non-normality, 97
- Normal distribution, 31
- Normal white noise, 158

- Odds, 336
- Off-line smoothing, 119
- On-line smoothing, 119
- One-sided moving average, 119
- One-sided weighted moving average, 119
- Ordered logit, 334
- Ordered outcomes, 333
- Ordinal data, 9
- Outliers, 97
- Panel data, 5
- Panels, 6
- Parameter instability, 136
- Parameters, 47
- Partial autocorrelation function, 156
- Partial correlation, 21
- Polynomial distributed lag, 222
- Polynomial in the lag operator, 165
- Polynomial trend, 117
- Population, 32
- Population model, 47
- Population regression, 156
- Positive serial correlation, 57
- Predictive causality, 237
- $\text{Prob}(F\text{-statistic})$, 54
- Probability density function, 30
- Probability value, 52
- Probit model, 337
- Proportional odds, 334
- QQ plots, 99
- Quadratic trend, 117
- Random number generator, 189
- Random walk, 137, 248
- Random walk with drift, 249
- Ratio data, 9
- Real-time smoothing, 119
- Realization, 154
- Recursive residuals, 136
- Regime switching, 139
- Regression function, 47
- Regression intercept, 47
- Regression on seasonal dummies, 75
- Regression slope coefficients, 47
- Relational graphics, 15
- RESET test, 112
- Residual plot, 59
- Residual scatter, 58
- Residuals, 44
- Robustness iteration, 151
- Sample, 32
- Sample autocorrelation function, 162
- Sample correlation, 34
- Sample covariance, 34
- Sample kurtosis, 33
- Sample mean, 33, 162
- Sample mean of the dependent variable, 53
- Sample partial autocorrelation, 164

- Sample path, 154
Sample skewness, 33
Sample standard deviation, 33
Sample standard deviation of the dependent variable, 53
Sample statistic, 33
Sample variance, 33
Scale, 30
Scatterplot matrix, 16
Schwarz information criterion, 57
Seasonal adjustment, 81
Seasonality, 72, 74
Second-order stationarity, 155
Serial correlation, 57
Serially uncorrelated, 158
Simple correlation, 21
Simple exponential smoothing, 261
Simple random sampling, 34
Simulating time series processes, 189
Single exponential smoothing, 261
Skewness, 31
Slope, 72
Slope dummies, 78
Smoothing, 118
Spurious regression, 282
Standard deviation, 30
Standard error of the regression, 55
Standard errors, 51
Standardized recursive residuals, 137
Statistic, 33
Stochastic processes, 166
Stochastic seasonality, 74
Stochastic trend, 72, 249
Strong white noise, 158
Student's-t GARCH, 317
Sum of squared residuals, 53
Superconsistency, 253
Taylor series expansions, 121
Threshold GARCH, 295
Threshold model, 139
Time dummy, 72
Time series, 6, 154
Time series data, 5
Time series of cross sections, 5
Time series plot, 13
Time series process, 157
Time-varying volatility, 285
Tobit model, 338
Trading-day variation, 77
Trend, 72
Two-sided moving average, 119
Unconditional mean and variance, 160
Unit autoregressive root, 247
Unit root, 247
Univariate, 13
Variance, 30
Vector autoregression of order p , 235

Volatility clustering, 289

Volatility dynamics , 289

Weak stationarity, 155

Weak white noise, 158

White noise, 158

Yule-Walker equation, 169

Zero-mean white noise, 158