

# Using Stata

For *Principles of Econometrics, Fifth Edition*



# Using Stata

*For Principles of Econometrics, Fifth Edition*

**LEE C. ADKINS**

*Oklahoma State University*

**R. CARTER HILL**

*Louisiana State University*



**JOHN WILEY & SONS, INC**

*New York / Chichester / Weinheim / Brisbane / Singapore / Toronto*

*Lee Adkins dedicates this work to his lovely and loving wife, Kathy*

*Carter Hill dedicates this work to Stan Johnson and George Judge*

ACQUISITIONS EDITOR	XXXXXX XXXXXXXX
MARKETING MANAGER	XXXXXX XXXXXXXX
PRODUCTION EDITOR	XXXXXX XXXXXXXX
PHOTO EDITOR	XXXXXX XXXXXXXX
ILLUSTRATION COORDINATOR	XXXXXX XXXXXXXX

This book was set in Times New Roman and printed and bound by .XXXXXXXXXXXXXXXXXXXX. The cover was printed by .XXXXXXXXXXXXXXXXXXXX.

This book is printed on acid-free paper. ∞

The paper in this book was manufactured by a mill whose forest management programs include sustained yield harvesting of its timberlands. Sustained yield harvesting principles ensure that the numbers of trees cut each year does not exceed the amount of new growth.

Copyright © John Wiley & Sons, Inc. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (508) 750-8400, fax (508) 750-4470. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158-0012, (212) 850-6011, fax (212) 850-6008, E-Mail: PERMREQ@WILEY.COM.

ISBN 0-471-xxxxx-x

Printed in the United States of America

**10 9 8 7 6 5 4 3 2 1**

# PREFACE

This book is a supplement to *Principles of Econometrics, 5<sup>th</sup> Edition* by R. Carter Hill, William E. Griffiths and Guay C. Lim (Wiley, 2018), hereinafter POE5. This book is not a substitute for the textbook, nor is it a stand alone computer manual. It is a companion to the textbook, showing how to perform the examples in the textbook using Stata Release 15. This book will be useful to students taking econometrics, as well as their instructors, and others who wish to use Stata for econometric analysis.

Stata is a very powerful program that is used in a wide variety of academic disciplines. The website is <http://www.stata.com>. There you will find a great deal of documentation. One great and visual resource is at UCLA: <https://stats.idre.ucla.edu/stata/>. We highly recommend this website.

In addition to this computer manual for Stata, there are similar manuals and support for the software packages EViews, Excel, Gretl and SAS. In addition, all the data for POE5 in various formats, including Stata, are available at <http://www.wiley.com/college/hill>.

Individual Stata data files, errata for this manual and the textbook can be found at <http://www.principlesofeconometrics.com/>.

The chapters in this book parallel the chapters in POE5. Thus, if you seek help for the examples in Chapter 11 of the textbook, check Chapter 11 in this book. However within a Chapter the sections numbers in POE5 do not necessarily correspond to the Stata manual sections. Data files and other resources for POE5 can be found at <http://www.stata.com/texts/s4poe5>.

We welcome comments on this book, and suggestions for improvement. We would like to acknowledge the help of the Stata Corporation, and in particular Bill Rising, for answering many of our questions and improving our prose and code.

Lee C. Adkins  
Department of Economics  
Oklahoma State University  
Stillwater, OK 74078  
[lee.adkins@okstate.edu](mailto:lee.adkins@okstate.edu)

R. Carter Hill  
Economics Department  
Louisiana State University  
Baton Rouge, LA 70803  
[eohill@lsu.edu](mailto:eohill@lsu.edu)

# **BRIEF CONTENTS**

1. Introducing Stata    **1**
2. Simple Linear Regression    **52**
3. Interval Estimation and Hypothesis Testing    **122**
4. Prediction, Goodness of Fit and Modeling Issues    **145**
5. Multiple Linear Regression    **201**
6. Further Inference in the Multiple Regression Model    **251**
7. Using Indicator Variables    **300**
8. Heteroskedasticity    **337**
9. Regression with Time-Series Data: Stationary Variables    **380**
10. Endogenous Regressors and Moment Based Estimation    **430**
11. Simultaneous Equations Models    **472**
12. Regression with Time-Series Data: Nonstationary Variables    **500**
13. Vector Error Correction and Vector Autoregressive Models    **538**
14. Time-Varying Volatility and ARCH Models    **562**
15. Panel Data Models    **585**
16. Qualitative and Limited Dependent Variable Models    **625**
  - A. Review of Math Essentials    **702**
  - B. Review of Probability Concepts    **715**
  - C. Review of Statistical Inference    **748**

# **CONTENTS**

## **Chapter 1 Introducing Stata 1**

### Key Terms 2

- 1.1 Starting Stata 2
- 1.2 The opening display 3
- 1.3 Exiting Stata 3
- 1.4 Stata data files for POE5 4
  - 1.4.1 A working directory 5
- 1.5 Opening Stata data files 6
  - 1.5.1 Using the toolbar 6
  - 1.5.2 The use command 7
  - 1.5.3 Using files on the internet 7
  - 1.5.4 Locating book files on the internet 7
- 1.6 The variables window 8
  - 1.6.1 Using the data utility for a single label 9
- 1.7 Describing data and obtaining summary statistics 9
- 1.8 The Stata help system 12
  - 1.8.1 Using keyword search 13
  - 1.8.2 Opening a dialog box 14
  - 1.8.3 Complete documentation in Stata manuals 14
  - 1.8.4 Advice 14
  - 1.8.5 Stata videos on YouTube 15
  - 1.8.6 Statalist 17
  - 1.8.7 Not elsewhere classified 17
- 1.9 Stata command syntax 17
  - 1.9.1 Syntax of summarize 18
  - 1.9.2 Learning syntax using the review window 19
- 1.10 Saving your work 22
  - 1.10.1 Copying and pasting 22
  - 1.10.2 Using a log file 23
- 1.11 Using the data browser 25
- 1.12 Using Stata graphics 25
  - 1.12.1 Histograms 25
  - 1.12.2 Scatter diagrams 30
- 1.13 Using Stata Do-files 32
- 1.14 Creating and managing variables 35
  - 1.14.1 Creating (generating) new variables 35
  - 1.14.2 Using the expression builder 36
  - 1.14.3 Dropping or keeping variables and observations 38
  - 1.14.4 Using arithmetic operators 40
  - 1.14.5 Using Stata math functions 41
- 1.15 Using Stata density functions 42
  - 1.15.1 Cumulative distribution functions 42
  - 1.15.2 Inverse cumulative distribution functions 43
- 1.16 Using and displaying scalars 43

- 1.16.1 Example of standard normal cdf 44
  - 1.16.2 Example of t-distribution tail-cdf 44
  - 1.16.3 Example computing percentile of the standard normal 45
  - 1.16.4 Example computing percentile of the *t*-distribution 45
  - 1.17 A scalar dialog box 45
  - 1.18 Using temporary scalars 47
- Chapter 1 Do-file 49

## **Chapter 2 Simple Linear Regression 52**

### Key Terms 52

- 2.1 The food expenditure data 53
    - 2.1.1 Starting a new problem 53
    - 2.1.2 Starting a log file 54
    - 2.1.3 Opening a Stata data file 55
    - 2.1.4 Browsing and listing the data 55
  - 2.2 Computing summary statistics 58
  - 2.3 Creating a scatter diagram 59
    - 2.3.1 Enhancing the plot 61
  - 2.4 Regression 64
    - 2.4.1 Fitted values and residuals 66
    - 2.4.2 Plotting the fitted regression line 68
  - 2.5 Using Stata to obtain predicted values 71
    - 2.5.1 Using saved coefficients 71
    - 2.5.2 Using lincom 73
    - 2.5.3 Using the margins command 73
    - 2.5.4 Using incomplete observations 76
    - 2.5.5 Computing an elasticity 78
  - 2.6 OLS estimator variances and covariance 83
    - 2.6.1 Estimating the variance of the error term 83
    - 2.6.2 Viewing estimated variances and covariance 84
    - 2.6.3 Saving the Stata data file 85
  - 2.7 Estimating nonlinear relationships 86
    - 2.7.1 A quadratic model 86
    - 2.7.2 A log-linear model 92
  - 2.8 Regression with indicator variables 99
  - Appendix 2A Average marginal effects 103
    - 2A.1 Elasticity in a linear relationship 103
    - 2A.2 Elasticity in a quadratic relationship 105
    - 2A.3 Slope in a log-linear model 106
  - Appendix 2B Simulation experiments 107
    - 2B.1 Fixed *x*'s 107
    - 2B.2 Random *x*'s 111
- Chapter 2 Do-file 114

## **Chapter 3 Interval Estimation and Hypothesis Testing 122**

### Key Terms 122

- 3.1 Interval estimates 123

3.1.1	Critical values from the $t$ -distribution	123
3.1.2	Creating an interval estimate	125
3.1.3	Creating an interval estimate using <code>lincom</code>	126
3.2	Hypothesis tests	127
3.2.1	Right-tail test of significance	127
3.2.2	Right-tail test of an economic hypothesis	128
3.2.3	Left-tail test of an economic hypothesis	129
3.2.4	Two-tail test of an economic hypothesis	129
3.2.5	Two-tail test of significance	130
3.3	$p$ -values	130
3.3.1	$p$ -value of a right-tail test	131
3.3.2	$p$ -value of a left-tail test	131
3.3.3	$p$ -value for a two-tail test	132
3.3.4	$p$ -values in Stata output	132
3.3.5	Testing and estimating linear combinations of parameters	133
Appendix 3A	Graphical tools	134
Appendix 3B	Monte Carlo simulation	136
3B.1	Fixed $x$ 's	136
3B.2	Random $x$ 's	139
Chapter 3	Do-file	140

## Chapter 4 Prediction, Goodness of Fit and Modeling Issues 145

Key Terms	145	
4.1	Least squares prediction	146
4.1.1	Editing the data	147
4.1.2	Estimate the regression and obtain postestimation results	147
4.1.3	Creating the prediction interval	148
4.1.4	Using margins to create the prediction Interval	149
4.2	Measuring goodness-of-fit	151
4.2.1	Correlations and $R^2$	152
4.3	The effects of scaling and transforming the Data	152
4.3.1	Reporting regression results	153
4.3.2	The linear-log functional form	157
4.3.3	Plotting the fitted linear-log model	159
4.3.4	Editing graphs	161
4.4	Analyzing the residuals	163
4.4.1	Residual plots	163
4.4.2	The Jarque-Bera test	166
4.4.3	Chi-square distribution critical values	168
4.4.4	Chi-square distribution $p$ -values	168
4.5	Polynomial models	173
4.5.1	Estimating and checking the linear	

	relationship	175
4.5.2	Estimating and checking a cubic equation	177
4.5.3	Estimating a log-linear yield growth model	179
4.6	Estimating a log-linear wage equation	181
4.6.1	The log-linear model	183
4.6.2	Calculating wage predictions	187
4.6.3	Constructing wage plots	188
4.6.4	Generalized $R^2$	190
4.6.5	Prediction intervals in the log-linear model	190
4.6.6	Prediction intervals in the log-linear model using margins	191
4.7	A log-log model	192
Chapter 4	Do-file	194

## Chapter 5 Multiple Linear Regression 201

Key Terms	201	
5.1	The Hamburger Chain Model	202
5.2	Least Squares Estimation	205
5.2.1	Least squares procedure	205
5.2.2	Least squares prediction	206
5.2.3	Rescaling the variables	207
5.2.4	Estimating the error variance	208
5.2.5	Measuring the goodness-of-fit	210
5.2.6	Frisch-Waugh-Lovell	211
5.3	Least Squares Precision	213
5.4	Confidence Intervals	216
5.4.1	Changing the confidence level	217
5.4.2	Linear combination of parameters	218
5.5	Hypothesis Tests	219
5.5.1	Two-sided $t$ -test	219
5.5.2	One-sided $t$ -test	220
5.5.3	Testing a linear combination of parameters	221
5.6	Interaction Variables	223
5.6.1	Polynomial regressors	223
5.6.2	Using factor variables for interactions	224
5.6.3	Interactions with other variables	226
5.6.4	Log-wages and quadratic interactions	228
5.6.5	Optimal level of advertising	230
5.6.6	Maximizing wages via experience	231
Appendix 5B.1	Nonlinear functions of a single parameter	232
Appendix 5B.2	Nonlinear functions of two parameters	234
Appendix 5C.1	Least squares estimation with chi-square errors	235
Appendix 5C.2	Monte Carlo simulation of the delta method	238
Appendix 5D	Bootstrapping	240



Chapter 5 Do-file 244

## **Chapter 6 Further Inference in the Multiple Regression Model 251**

Key Terms 251

- 6.1 Testing joint hypotheses: The  $F$ -test 251
    - 6.1.1 Testing the significance of the model 256
    - 6.1.2 Relationship between  $t$ - and  $F$ -tests 258
    - 6.1.3 More general  $F$ -tests 259
    - 6.1.4 Large sample tests 261
    - 6.1.5 Nonlinear hypothesis tests 262
  - 6.2 Stata programs 264
  - 6.3 Nonsample information 265
  - 6.4 Model specification 268
    - 6.4.1 Omitted variables 268
    - 6.4.2 Irrelevant variables 270
    - 6.4.3 Choosing the model 271
    - 6.4.4 RESET test for function form 275
    - 6.4.5 RESET program 276
    - 6.4.6 Control variables 277
    - 6.4.7 Prediction-forecast error variance 279
    - 6.4.8 Prediction-model selection and RMSE 281
  - 6.5 Poor data, collinearity, and insignificance 283
    - 6.5.1 Variance inflation factors 285
    - 6.5.2 Influential observations 286
  - 6.6 Nonlinear least squares 289
- Chapter 6 Do-file 293

## **Chapter 7 Using Indicator Variables 300**

Key Terms 300

- 7.1 Indicator variables 301
  - 7.1.1 Creating indicator variables 301
  - 7.1.2 Estimating an indicator variable regression 302
  - 7.1.3 Testing the significance of the indicator Variables 303
  - 7.1.4 Further calculations 304
  - 7.1.5 Computing average marginal effects 305
- 7.2 Applying indicator variables 306
  - 7.2.1 Interactions between qualitative factors 307
  - 7.2.2 Adding regional indicators 310
  - 7.2.3 Testing the equivalence of two regressions 311
  - 7.2.4 Estimating separate regressions 313
  - 7.2.5 Indicator variables in log-linear models 314

- 7.3 The linear probability model 317
  - 7.4 Treatment effects 319
  - 7.5 Differences-in-Differences estimation 326
- Chapter 7 Do-file 332

## **Chapter 8 Heteroskedasticity 337**

Key Terms 337

- 8.1 The nature of heteroskedasticity 338
  - 8.2 Heteroskedastic-consistent standard errors 341
  - 8.3 The generalized least squares estimator 342
    - 8.3.1 Feasible GLS—a more general case 347
    - 8.3.2 Feasible GLS with a heteroskedastic partition 350
  - 8.4 Detecting heteroskedasticity 353
    - 8.4.1 The Goldfeld-Quandt test using partitioned data 353
    - 8.4.2 The Goldfeld-Quandt test in the food expenditure model 354
    - 8.4.3 Lagrange multiplier tests 355
  - 8.5 Heteroskedasticity in the linear probability model 362
- Appendix 8D Alternative robust sandwich estimators 365
- Appendix 8E Monte Carlo evidence 367
- Chapter 8 Do-file 373

## **Chapter 9 Regression with Time-Series Data: Stationary Variables 380**

Key Terms 380

- 9.1 Introduction 381
    - 9.1.1 Defining time-series in Stata 381
    - 9.1.2 Time-series plots 384
    - 9.1.3 Stata's lag and difference operators 385
  - 9.2 Correlogram 388
  - 9.3 The AR(2) model 391
  - 9.4 Autoregressive distributed lag models 395
    - 9.4.1 Forecasts and forecast intervals 396
    - 9.4.2 Model selection 397
    - 9.4.3 Granger causality 402
  - 9.5 Serial correlation in residuals 403
    - 9.5.1 Detecting autocorrelation in residuals 403
    - 9.5.2 Okun's Law 407
    - 9.5.3 HAC standard errors 410
    - 9.5.4 Nonlinear least squares 413
    - 9.5.5 Feasible GLS 415
  - 9.6 The consumption function 417
  - 9.7 Multipliers for an IDL model 420
  - 9.8 Durbin-Watson Test 423
- Chapter 9 Do-file 424

## **Chapter 10 Endogenous Regressors and Moment Based Estimation 430**

### Key Terms 430

- 10.1 Least squares estimation of a wage equation 431
- 10.2 Two-stage least squares 432
- 10.3 IV estimation with surplus instruments 436
  - 10.3.1 Illustrating partial correlations 440
- 10.4 The Hausman test for endogeneity 442
- 10.5 Testing the validity of surplus instruments 445
- 10.6 Testing for weak instruments 446
- 10.7 Calculating the Cragg-Donald  $F$ -statistic 453
- 10.8 Illustrations using simulated data 455
- 10.9 A simulation experiment 459
- Chapter 10 Do-file 466

## **Chapter 11 Simultaneous Equations Models 472**

### Key Terms 472

- 11.1 Truffle supply and demand 472
- 11.2 Estimating the reduced form equations 474
- 11.3 2SLS estimates of truffle demand 475
- 11.4 2SLS estimates of truffle supply 480
- 11.5 Supply and demand of fish 481
- 11.6 Reduced forms for fish price and quantity 482
- 11.7 2SLS estimates of fish demand 484
- 11.8 2SLS alternatives 486
- 11.9 Monte Carlo simulation 489
- Chapter 11 Do-file 495

## **Chapter 12 Regression with Time-Series Data: Nonstationary Variables 500**

### Key Terms 500

- 12.1 Stationary and nonstationary data 501
  - 12.1.1 Review: generating dates in Stata 501
  - 12.1.2 Extracting dates 502
  - 12.1.3 Graphing the data 502
  - 12.1.4 Summary statistics using subsamples 505
  - 12.1.5 Correlogram 507
- 12.2 Deterministic trends 509
- 12.3 Spurious regressions 512
- 12.4 Unit root tests for stationarity 514
  - 12.4.1 Is GDP trend stationary? 522
  - 12.4.2 Is wheat yield stationary? 523
- 12.5 Integration and cointegration 524
  - 12.5.1 Order of integration 524

- 12.5.2 Engle-Granger test 526
- 12.5.3 The error correction model 527
- 12.5.4 Regression with no cointegration 528
- Chapter 12 Do-file 533

## **Chapter 13 Vector Error Correction and Vector Autoregressive Models 538**

### Key Terms 538

- 13.1 VEC and VAR models 538
- 13.2 Estimating a VEC model 539
- 13.3 Estimating a VAR 544
- 13.4 Impulse responses and variance decompositions 553
- Chapter 13 Do-file 559

## **Chapter 14 Time-Varying Volatility and ARCH Models 562**

### Key Terms 562

- 14.1 ARCH model and time-varying volatility 563
- 14.2 Simulating ARCH 565
- 14.3 Testing, estimating and forecasting 568
- 14.4 Extensions 575
  - 14.4.1 GARCH 576
  - 14.4.2 Threshold GARCH 577
  - 14.4.3 GARCH-in-mean 579
- Chapter 14 Do-file 582

## **Chapter 15 Panel Data Models 585**

### Key Terms 585

- 15.1 A microeconomic panel 586
- 15.2 The fixed-effects estimator 587
  - 15.2.1 The difference estimator:  $T = 2$  587
  - 15.2.2 The within estimator:  $T = 2$  591
  - 15.2.3 The within estimator:  $T = 3$  594
  - 15.2.4 The fixed-effects estimator: xtreg 595
  - 15.2.5 The least squares dummy variable estimator 597
  - 15.2.6 Testing for fixed effects 599
- 15.3 Panel data regression error assumptions 601
  - 15.3.1 OLS estimation with cluster-robust standard errors 601
  - 15.3.2 Fixed-effects estimation with cluster-robust standard errors 603
  - 15.3.3 Random-effects estimation of a production function 604
  - 15.3.4 Random-effects estimation of a wage equation 605
  - 15.3.5 Testing for random-effects 610

- 15.3.6 The Hausman contrast test for the production function 612
- 15.3.7 The Hausman contrast test for the wage equation 613
- 15.3.8 A regression based Hausman test for the production function 614
- 15.3.9 A regression based Hausman test for the wage equation 616
- 15.3.10 The Hausman-Taylor estimator 618
- Chapter 15 Do-file 620

## Chapter 16 Qualitative and Limited Dependent Variable Models 625

- Key Terms 625
- 16.1 Models with binary dependent variables 626
  - 16.1.1 The linear probability model 626
  - 16.1.2 Probit: a small example 628
  - 16.1.3 Probit: the transportation data 630
  - 16.1.4 Marginal effects 634
  - 16.1.5 Probit marginal effects: details 637
  - 16.1.6 Standard error of average marginal effect 616
- 16.2 The logit model for binary choice 641
  - 16.2.1 Wald tests 647
  - 16.2.2 Likelihood ratio tests 650
  - 16.2.3 Binary choice models with a continuous endogenous variable 653
- 16.3 Multinomial logit 656
- 16.4 Conditional logit 662
  - 16.4.1 Estimation using asclogit 666
- 16.5 Ordered choice models 669
- 16.6 Models for count data 673
- 16.7 Censored data models 679
- 16.8 Selection bias 684
- Appendix 16D Tobit Monte Carlo experiment 690
- Chapter 16 Do-file 694

## Appendix A Review of Math Essentials 702

- Key Terms 702
- A.1 Stata math and logical operators 702
- A.2 Math functions 703
- A.3 Extensions to generate 707
- A.4 The calculator 708
- A.5 Scientific notation 708
- A.6 Logarithms 709
- A.7 Numerical derivatives and integrals 710
- Appendix A Do-file 713

## Appendix B Review of Probability Concepts 715

- Key Terms 715
- B.1 Stata probability functions 716
- B.2 Binomial distribution 719
- B.3 Poisson distribution 721
- B.4 Normal distribution 723
  - B.4.1 Normal density plots 723
  - B.4.2 Normal probability calculations 724
- B.5 Chi-square distribution 725
  - B.5.1 Plotting the chi-square density 725
  - B.5.2 Chi-square probability calculations 726
  - B.5.3 The non-central chi-square pdf 727
- B.6 Student's *t*-distribution 728
  - B.6.1 Plot of standard normal and *t*(3) 728
  - B.6.2 *t*-distribution probabilities 728
  - B.6.3 Graphing tail probabilities 729
  - B.6.4 The non-central *t*-distribution 731
- B.7 *F*-distribution 732
  - B.7.1 Plotting the *F*-density 732
  - B.7.2 *F*-distribution probability calculations 732
  - B.7.3 The non-central *F*-distribution 733
- B.8 The log-normal distribution 734
- B.9 Random numbers 736
  - B.9.1 Using inversion method 737
  - B.9.2 Creating uniform random numbers 740
- Appendix B Do-file 743

## Appendix C Review of Statistical Inference 748

- Key Terms 748
- C.1 Examining the hip data 749
  - C.1.1 Constructing a histogram 749
  - C.1.2 Obtaining summary statistics 751
  - C.1.3 Estimating the population mean 752
- C.2 Using simulated data values 753
- C.3 The central limit theorem 757
- C.4 Estimating population moments 760
- C.5 Interval estimation 762
  - C.5.1 Computing confidence intervals 763
  - C.5.2 Using simulated data 763
  - C.5.3 Using the hip data 765
- C.6 Testing the mean of a normal population 767
  - C.6.1 Right-tail test 767
  - C.6.2 Two-tail test 769
- C.7 Testing the variance of a normal population 770
- C.8 Testing the equality of two normal population means 771
  - C.8.1 Population variances are equal 772

C.8.2 Population variances are unequal	772
C.9 Testing the equality of two normal population variances	774
C.10 Testing normality	775
C.11 Maximum likelihood estimation	776
C.11.1 Testing a population proportion	778
C.11.2 Likelihood ratio test	779
C.11.3 Wald test	780
C.11.4 Lagrange multiplier test	780
C.12 Least squares	781
C.13 Kernel density estimator	782
Appendix C Do-file	786