FUNDAMENTALS OF APPLIED ECONOMETRICS

by
RICHARD A. ASHLEY
Economics Department
Virginia Tech

WILEY
John Wiley and Sons, Inc.

Vice President & Executive Publisher	George Hoffman
Project Editor	Jennifer Manias
Assistant Editor	Emily McGee
Editorial Assistant	Erica Horowitz
Associate Director of Marketing	Amy Scholz
Marketing Manager	Jesse Cruz
Marketing Assistant	Courtney Luzzi
Executive Media Editor	Allison Morris
Media Editor	Greg Chaput
Senior Production Manager	Janis Soo
Associate Production Manager	Joyce Poh
Assistant Production Editor	Yee Lyn Song
Cover Designer	Jerel Seah
Cover Photo Credit	©AveryPhotography/iStockphoto

This book was set in 10/12 Times Roman by Thomson Digital and printed and bound by RR Donnelley. The cover was printed by RR Donnelly.

This book is printed on acid-free paper. ∞

Library of Congress Cataloging-in-Publication Data

Ashley, Richard A. (Richard Arthur), 1950-
Fundamentals of applied econometrics / by Richard Ashley. – 1st ed.
p. cm.
Includes index.
ISBN 978-0-470-59182-6 (hardback)
1. Econometrics. 2. Econometrics–Statistical methods. 3. Econometrics–Data processing. I. Title.
HB139.A84 2012
330.01'5195–dc23 2011041421

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

For Rosalind and Elisheba

BRIEF CONTENTS

What's Different about This Book		**xiii**
Working with Data in the "Active Learning Exercises"		**xxii**
Acknowledgments		**xxiii**
Notation		**xxiv**
Part I	**INTRODUCTION AND STATISTICS REVIEW**	**1**
Chapter 1	INTRODUCTION	3
Chapter 2	A REVIEW OF PROBABILITY THEORY	11
Chapter 3	ESTIMATING THE MEAN OF A NORMALLY DISTRIBUTED RANDOM VARIABLE	46
Chapter 4	STATISTICAL INFERENCE ON THE MEAN OF A NORMALLY DISTRIBUTED RANDOM VARIABLE	68
Part II	**REGRESSION ANALYSIS**	**97**
Chapter 5	THE BIVARIATE REGRESSION MODEL: INTRODUCTION, ASSUMPTIONS, AND PARAMETER ESTIMATES	99
Chapter 6	THE BIVARIATE LINEAR REGRESSION MODEL: SAMPLING DISTRIBUTIONS AND ESTIMATOR PROPERTIES	131
Chapter 7	THE BIVARIATE LINEAR REGRESSION MODEL: INFERENCE ON β	150
Chapter 8	THE BIVARIATE REGRESSION MODEL: R^2 AND PREDICTION	178
Chapter 9	THE MULTIPLE REGRESSION MODEL	191
Chapter 10	DIAGNOSTICALLY CHECKING AND RESPECIFYING THE MULTIPLE REGRESSION MODEL: DEALING WITH POTENTIAL OUTLIERS AND HETEROSCEDASTICITY IN THE CROSS-SECTIONAL DATA CASE	224
Chapter 11	STOCHASTIC REGRESSORS AND ENDOGENEITY	259
Chapter 12	INSTRUMENTAL VARIABLES ESTIMATION	303
Chapter 13	DIAGNOSTICALLY CHECKING AND RESPECIFYING THE MULTIPLE REGRESSION MODEL: THE TIME-SERIES DATA CASE (PART A)	342
Chapter 14	DIAGNOSTICALLY CHECKING AND RESPECIFYING THE MULTIPLE REGRESSION MODEL: THE TIME-SERIES DATA CASE (PART B)	389
Part III	**ADDITIONAL TOPICS IN REGRESSION ANALYSIS**	**455**
Chapter 15	REGRESSION MODELING WITH PANEL DATA (PART A)	459
Chapter 16	REGRESSION MODELING WITH PANEL DATA (PART B)	507
Chapter 17	A CONCISE INTRODUCTION TO TIME-SERIES ANALYSIS AND FORECASTING (PART A)	536
Chapter 18	A CONCISE INTRODUCTION TO TIME-SERIES ANALYSIS AND FORECASTING (PART B)	595
Chapter 19	PARAMETER ESTIMATION BEYOND CURVE-FITTING: MLE (WITH AN APPLICATION TO BINARY-CHOICE MODELS) AND GMM (WITH AN APPLICATION TO IV REGRESSION)	647
Chapter 20	CONCLUDING COMMENTS	681
Mathematics Review		**693**

TABLE OF CONTENTS

What's Different about This Book		**xiii**
Working with Data in the "Active Learning Exercises"		**xxii**
Acknowledgments		**xxiii**
Notation		**xxiv**
Part I	**INTRODUCTION AND STATISTICS REVIEW**	**1**
Chapter 1	INTRODUCTION	3
	1.1 Preliminaries	3
	1.2 Example: Is Growth Good for the Poor?	4
	1.3 What's to Come	7
	ALE 1a: An Econometrics "Time Capsule"	8
	ALE 1b: Investigating the Slope Graphically Using a Scatterplot	(Online)
	ALE 1c: Examining Some Disturbing Variations on Dollar & Kraay's Model	(Online)
	ALE 1d: The Pitfalls of Making Scatterplots with Trended Time-Series Data	(Online)
Chapter 2	A REVIEW OF PROBABILITY THEORY	11
	2.1 Introduction	11
	2.2 Random Variables	12
	2.3 Discrete Random Variables	13
	2.4 Continuous Random Variables	17
	2.5 Some Initial Results on Expectations	19
	2.6 Some Results on Variances	20
	2.7 A Pair of Random Variables	22
	2.8 The Linearity Property of Expectations	24
	2.9 Statistical Independence	26
	2.10 Normally Distributed Random Variables	29
	2.11 Three Special Properties of Normally Distributed Variables	31
	2.12 Distribution of a Linear Combination of Normally Distributed Random Variables	32

2.13 Conclusion 36
Exercises 37
ALE 2a: The Normal Distribution 42
ALE 2b: Central Limit Theorem Simulators on the Web (Online)
Appendix 2.1: The Conditional Mean of a Random Variable 44
Appendix 2.2: Proof of the Linearity Property for the Expectation of a Weighted Sum of Two Discretely Distributed Random Variables 45

Chapter 3 ESTIMATING THE MEAN OF A NORMALLY DISTRIBUTED RANDOM VARIABLE 46
3.1 Introduction 46
3.2 Estimating μ by Curve Fitting 48
3.3 The Sampling Distribution of $\bar{Y}$ 51
3.4 Consistency – A First Pass 54
3.5 Unbiasedness and the Optimal Estimator 55
3.6 The Squared Error Loss Function and the Optimal Estimator 56
3.7 The Feasible Optimality Properties: Efficiency and BLUness 58
3.8 Summary 61
3.9 Conclusions and Lead-in to Next Chapter 62
Exercises 62
ALE 3a: Investigating the Consistency of the Sample Mean and Sample Variance Using Computer-Generated Data 64
ALE 3b: Estimating Means and Variances Regarding the Standard & Poor's SP500 Stock Index (Online)

Chapter 4 STATISTICAL INFERENCE ON THE MEAN OF A NORMALLY DISTRIBUTED RANDOM VARIABLE 68
4.1 Introduction 68
4.2 Standardizing the distribution of $\bar{Y}$ 69
4.3 Confidence Intervals for μ When σ^2 Is Known 69
4.4 Hypothesis Testing when σ^2 Is Known 71
4.5 Using S^2 to Estimate σ^2 (and Introducing the Chi-Squared Distribution) 75
4.6 Inference Results on μ When σ^2 Is Unknown (and Introducing the Student's t Distribution) 78
4.7 Application: State-Level U.S. Unemployment Rates 82
4.8 Introduction to Diagnostic Checking: Testing the Constancy of μ across the Sample 84
4.9 Introduction to Diagnostic Checking: Testing the Constancy of σ^2 across the Sample 87
4.10 Some General Comments on Diagnostic Checking 89
4.11 Closing Comments 90
Exercises 91
ALE 4a: Investigating the Sensitivity of Hypothesis Test p-Values to Departures from the NIID(μ, σ^2) Assumption Using Computer-Generated Data 93
ALE 4b: Individual Income Data from the Panel Study on Income Dynamics (PSID) – Does Birth-Month Matter? (Online)

Part II REGRESSION ANALYSIS 97

Chapter 5 THE BIVARIATE REGRESSION MODEL: INTRODUCTION, ASSUMPTIONS, AND PARAMETER ESTIMATES 99

5.1 Introduction 99
5.2 The Transition from Mean Estimation to Regression: Analyzing the Variation of Per Capita Real Output across Countries 100
5.3 The Bivariate Regression Model – Its Form and the "Fixed in Repeated Samples" Causality Assumption 105
5.4 The Assumptions on the Model Error Term, U_i 106
5.5 Least Squares Estimation of α and β 109
5.6 Interpreting the Least Squares Estimates of α and β 118
5.7 Bivariate Regression with a Dummy Variable: Quantifying the Impact of College Graduation on Weekly Earnings 120
Exercises 127
ALE 5a: Exploring the Penn World Table Data 128
ALE 5b: Verifying $\hat{\alpha}^*_{ols}$ and $\hat{\beta}^*_{ols}$ over a Very Small Data Set (Online)
ALE 5c: Extracting and Downloading CPS Data from the Census Bureau Web Site (Online)
ALE 5d: Verifying That $\hat{\beta}^*_{ols}$ on a Dummy Variable Equals the Difference in the Sample Means (Online)
Appendix 5.1: $\hat{\beta}^*_{ols}$ When x_i Is a Dummy Variable 130

Chapter 6 THE BIVARIATE LINEAR REGRESSION MODEL: SAMPLING DISTRIBUTIONS AND ESTIMATOR PROPERTIES 131

6.1 Introduction 131
6.2 Estimates and Estimators 132
6.3 $\hat{\beta}$ as a Linear Estimator and the Least Squares Weights 132
6.4 The Sampling Distribution of $\hat{\beta}$ 134
6.5 Properties of $\hat{\beta}$: Consistency 140
6.6 Properties of $\hat{\beta}$: Best Linear Unbiasedness 140
6.7 Summary 143
Exercises 144
ALE 6a: Outliers and Other Perhaps Overly Influential Observations: Investigating the Sensitivity of $\hat{\beta}$ to an Outlier Using Computer-Generated Data 147
ALE 6b: Investigating the Consistency of $\hat{\beta}$ Using Computer-Generated Data (Online)

Chapter 7 THE BIVARIATE LINEAR REGRESSION MODEL: INFERENCE ON β 150

7.1 Introduction 150
7.2 A Statistic for β with a Known Distribution 152
7.3 A 95% Confidence Interval for β with σ^2 Given 152
7.4 Estimates versus Estimators and the Role of the Model Assumptions 154
7.5 Testing a Hypothesis about β with σ^2 Given 156
7.6 Estimating σ^2 158
7.7 Properties of S^2 159
7.8 A Statistic for β Not Involving σ^2 160

7.9 A 95% Confidence Interval for β with σ^2 Unknown 160
7.10 Testing a Hypothesis about β with σ^2 Unknown 162
7.11 Application: The Impact of College Graduation on Weekly Earnings (Inference Results) 164
7.12 Application: Is Growth Good for the Poor? 168
7.13 Summary 169
Exercises 169
ALE 7a: Investigating the Sensitivity of Slope Coefficient Inference to Departures from the $U_i \sim$ NIID(0, σ^2) Assumption Using Computer-Generated Data 172
ALE 7b: Distorted Inference in Time-Series Regressions with Serially Correlated Model Errors: An Investigation Using Computer-Generated Data (Online)
Appendix 7.1: Proof That S^2 Is Independent of $\hat{\beta}$ 177

Chapter 8 THE BIVARIATE REGRESSION MODEL: R^2 AND PREDICTION 178
8.1 Introduction 178
8.2 Quantifying How Well the Model Fits the Data 179
8.3 Prediction as a Tool for Model Validation 182
8.4 Predicting Y_{N+1} given x_{N+1} 184
Exercises 188
ALE 8a: On the Folly of Trying Too Hard: A Simple Example of "Data Mining" 189

Chapter 9 THE MULTIPLE REGRESSION MODEL 191
9.1 Introduction 191
9.2 The Multiple Regression Model 191
9.3 Why the Multiple Regression Model Is Necessary and Important 192
9.4 Multiple Regression Parameter Estimates via Least Squares Fitting 193
9.5 Properties and Sampling Distribution of $\hat{\beta}_{ols,\,1} \ldots \hat{\beta}_{ols,\,k}$ 195
9.6 Overelaborate Multiple Regression Models 202
9.7 Underelaborate Multiple Regression Models 205
9.8 Application: The Curious Relationship between Marriage and Death 206
9.9 Multicollinearity 208
9.10 Application: The Impact of College Graduation and Gender on Weekly Earnings 210
9.11 Application: Vote Fraud in Philadelphia Senatorial Elections 214
Exercises 218
ALE 9a: A Statistical Examination of the Florida Voting in the November 2000 Presidential Election – Did Mistaken Votes for Pat Buchanan Swing the Election from Gore to Bush? 220
ALE 9b: Observing and Interpreting the Symptoms of Multicollinearity (Online)
ALE 9c: The Market Value of a Bathroom in Georgia (Online)
Appendix 9.1: Prediction Using the Multiple Regression Model 222

Chapter 10 DIAGNOSTICALLY CHECKING AND RESPECIFYING THE MULTIPLE REGRESSION MODEL: DEALING WITH POTENTIAL OUTLIERS AND HETEROSCEDASTICITY IN THE CROSS-SECTIONAL DATA CASE 224
10.1 Introduction 224
10.2 The Fitting Errors as Large-Sample Estimates of the Model Errors, $U_1 \ldots U_N$ 227

10.3 Reasons for Checking the Normality of the Model Errors, $U_1, \ldots U_N$ 228
10.4 Heteroscedasticity and Its Consequences 237
10.5 Testing for Heteroscedasticity 239
10.6 Correcting for Heteroscedasticity of Known Form 243
10.7 Correcting for Heteroscedasticity of Unknown Form 248
10.8 Application: Is Growth Good for the Poor? Diagnostically Checking the Dollar/Kraay (2002) Model.[1] 252
Exercises 256
ALE 10a: The Fitting Errors as Approximations for the Model Errors 257
ALE 10b: Does Output Per Person Depend on Human Capital? (A Test of the Augmented Solow Model of Growth)[2] (Online)
ALE 10c: Is Trade Good or Bad for the Environment? (First Pass)[3] (Online)

Chapter 11 STOCHASTIC REGRESSORS AND ENDOGENEITY 259
11.1 Introduction 259
11.2 Unbiasedness of the OLS Slope Estimator with a Stochastic Regressor Independent of the Model Error 261
11.3 A Brief Introduction to Asymptotic Theory 264
11.4 Asymptotic Results for the OLS Slope Estimator with a Stochastic Regressor 269
11.5 Endogenous Regressors: Omitted Variables 272
11.6 Endogenous Regressors: Measurement Error 273
11.7 Endogenous Regressors: Joint Determination – Introduction to Simultaneous Equation Macroeconomic and Microeconomic Models 274
11.8 How Large a Sample Is "Large Enough"? The Simulation Alternative 278
11.9 An Example: Bootstrapping the Angrist-Krueger (1991) Model 282
Exercises 290
ALE 11a: Central Limit Theorem Convergence for $\hat{\beta}^{OLS}$ in the Bivariate Regression Model 293
ALE 11b: Bootstrap Analysis of the Convergence of the Asymptotic Sampling Distributions for Multiple Regression Model Parameter Estimators (Online)
Appendix 11.1: The Algebra of Probability Limits 298
Appendix 11.2: Derivation of the Asymptotic Sampling Distribution of the OLS Slope Estimator 299

Chapter 12 INSTRUMENTAL VARIABLES ESTIMATION 303
12.1 Introduction – Why It Is Challenging to Test for Endogeneity 303
12.2 Correlation versus Causation – Two Ways to Untie the Knot 305
12.3 The Instrumental Variables Slope Estimator (and Proof of Its Consistency) in the Bivariate Regression Model 311

[1] Uses data from Dollar, D., and A. Kraay (2002), "Growth Is Good for the Poor," *Journal of Economic Growth 7,* 195–225.

[2] Uses data from Mankiw, G. N., D. Romer, and D. N. Weil (1992), "A Contribution to the Empirics of Economic Growth," *The Quarterly Journal of Economics 107(2)*, 407–37. Mankiw et al. estimate and test a Solow growth model, augmenting it with a measure of human capital, quantified by the percentage of the population in secondary school.

[3] Uses data from Frankel, J. A., and A. K. Rose (2005), "Is Trade Good or Bad for the Environment? Sorting Out the Causality," *The Review of Economics and Statistics 87(1)*, 85–91. Frankel and Rose quantify and test the effect of trade openness $\{(X+M)/Y\}$ on three measures of environmental damage (SO_2, NO_2, and total suspended particulates). Since trade openness may well be endogenous, Frankel and Rose also obtain 2SLS estimates; these are examined in Active Learning Exercise 12b.

12.4 Inference Using the Instrumental Variables Slope Estimator 313
12.5 The Two-Stage Least Squares Estimator for the Overidentified Case 317
12.6 Application: The Relationship between Education and Wages (Angrist and Krueger, 1991) 321
Exercises 330
ALE 12a: The Role of Institutions "Rule of Law" in Economic Growth[4] 332
ALE 12b: Is Trade Good or Bad for the Environment? (Completion)[5] (Online)
ALE 12c: The Impact of Military Service on the Smoking Behavior of Veterans[6] (Online)
ALE 12d: The Effect of Measurement-Error Contamination on OLS Regression Estimates and the Durbin/Bartlett IV Estimators (Online)
Appendix 12.1: Derivation of the Asymptotic Sampling Distribution of the Instrumental Variables Slope Estimator 336
Appendix 12.2: Proof That the 2SLS Composite Instrument Is Asymptotically Uncorrelated with the Model Error Term 340

Chapter 13 DIAGNOSTICALLY CHECKING AND RESPECIFYING THE MULTIPLE REGRESSION MODEL: THE TIME-SERIES DATA CASE (PART A) 342
13.1 An Introduction to Time-Series Data, with a "Road Map" for This Chapter 342
13.2 The Bivariate Time-Series Regression Model with Fixed Regressors but Serially Correlated Model Errors, $U_1 \ldots U_T$ 348
13.3 Disastrous Parameter Inference with Correlated Model Errors: Two Cautionary Examples Based on U.S. Consumption Expenditures Data 353
13.4 The AR(1) Model for Serial Dependence in a Time-Series 363
13.5 The Consistency of $\hat{\varphi}_1^{OLS}$ as an Estimator of φ_1 in the AR(1) Model and Its Asymptotic Distribution 367
13.6 Application of the AR(1) Model to the Errors of the (Detrended) U.S. Consumption Function – and a Straightforward Test for Serially Correlated Regression Errors 370
13.7 Dynamic Model Respecification: An Effective Response to Serially Correlated Regression Model Errors, with an Application to the (Detrended) U.S. Consumption Function 374
Exercises 382
Appendix 13.1: Derivation of the Asymptotic Sampling Distribution of $\hat{\varphi}_1^{OLS}$ in the AR(1) Model 384

Chapter 14 DIAGNOSTICALLY CHECKING AND RESPECIFYING THE MULTIPLE REGRESSION MODEL: THE TIME-SERIES DATA CASE (PART B) 389
14.1 Introduction: Generalizing the Results to Multiple Time-Series 389
14.2 The Dynamic Multiple Regression Model 390

[4] Uses data from Acemoglu, D., S. Johnson, and J. A. Robinson (2001), "The Colonial Origins of Comparative Development," *The American Economic Review 91(5)*, 1369–1401. These authors argue that the European mortality rate in colonial times is a valid instrument for current institutional quality because Europeans settled (and imported their cultural institutions) only in colonies with climates they found healthy.

[5] See footnote for Active Learning Exercise 10c.

[6] Uses data from Bedard, K., and O. Deschênes (2006), "The Long-Term Impact of Military Service on Health: Evidence from World War II and Korean War Veterans." *The American Economic Review 96(1)*, 176–194. These authors quantify the impact of the provision of free and/or low-cost tobacco products to servicemen on smoking and (later) on mortality rates, using instrumental variable methods to control for the nonrandom selection into military service.

14.3 I(1) or "Random Walk" Time-Series 395
14.4 Capstone Example Part 1: Modeling Monthly U.S. Consumption Expenditures in Growth Rates 404
14.5 Capstone Example Part 2: Modeling Monthly U.S. Consumption Expenditures in Growth Rates and Levels (Cointegrated Model) 424
14.6 Capstone Example Part 3: Modeling the Level of Monthly U.S. Consumption Expenditures 431
14.7 Which Is Better: To Model in Levels or to Model in Changes? 447
Exercises 449
ALE 14a: Analyzing the Food Price Sub-Index of the Monthly U.S. Consumer Price Index 451
ALE 14b: Estimating Taylor Rules for How the U.S. Fed Sets Interest Rates (Online)

Part III ADDITIONAL TOPICS IN REGRESSION ANALYSIS 455

Chapter 15 REGRESSION MODELING WITH PANEL DATA (PART A) 459
15.1 Introduction: A Source of Large (but Likely Heterogeneous) Data Sets 459
15.2 Revisiting the Chapter 5 Illustrative Example Using Data from the Penn World Table 460
15.3 A Multivariate Empirical Example 462
15.4 The Fixed Effects and the Between Effects Models 469
15.5 The Random Effects Model 478
15.6 Diagnostic Checking of an Estimated Panel Data Model 490
Exercises 500
Appendix 15.1: Stata Code for the Generalized Hausman Test 503

Chapter 16 REGRESSION MODELING WITH PANEL DATA (PART B) 507
16.1 Relaxing Strict Exogeneity: Dynamics and Lagged Dependent Variables 507
16.2 Relaxing Strict Exogeneity: The First-Differences Model 515
16.3 Summary 528
Exercises 529
ALE 16a: Assessing the Impact of 4-H Participation on the Standardized Test Scores of Florida Schoolchildren 531
ALE16b: Using Panel Data Methods to Reanalyze Data from a Public Goods Experiment (Online)

Chapter 17 A CONCISE INTRODUCTION TO TIME-SERIES ANALYSIS AND FORECASTING (PART A) 536
17.1 Introduction: The Difference between Time-Series Analysis and Time-Series Econometrics 536
17.2 Optimal Forecasts: The Primacy of the Conditional-Mean Forecast and When It Is Better to Use a Biased Forecast 538
17.3 The Crucial Assumption (Stationarity) and the Fundamental Tools: The Time-Plot and the Sample Correlogram 543
17.4 A Polynomial in the Lag Operator and Its Inverse: The Key to Understanding and Manipulating Linear Time-Series Models 559
17.5 Identification/Estimation/Checking/Forecasting of an Invertible MA(q) Model 563

17.6 Identification/Estimation/Checking/Forecasting of a Stationary AR(p) Model 575
17.7 ARMA(p,q) Models and a Summary of the Box-Jenkins Modeling Algorithm 581
Exercises 586
ALE 17a: Conditional Forecasting Using a Large-Scale Macroeconometric Model 589
ALE 17b: Modeling U.S. GNP (Online)

Chapter 18 A CONCISE INTRODUCTION TO TIME-SERIES ANALYSIS AND FORECASTING (PART B) 595
18.1 Integrated – ARIMA(p,d,q) – Models and "Trend like" Behavior 595
18.2 A Univariate Application: Modeling the Monthly U.S. Treasury Bill Rate 604
18.3 Seasonal Time-Series Data and ARMA Deseasonalization of the U.S. Total Nonfarm Payroll Time-Series 611
18.4 Multivariate Time-Series Models 617
18.5 Post-Sample Model Forecast Evaluation and Testing for Granger-Causation 622
18.6 Modeling Nonlinear Serial Dependence in a Time-Series 623
18.7 Additional Topics in Forecasting 637
Exercises 645
ALE 18a: Modeling the South Korean Won – U.S. Dollar Exchange Rate 645
ALE 18b: Modeling the Daily Returns to Ford Motor Company Stock (Online)

Chapter 19 PARAMETER ESTIMATION BEYOND CURVE-FITTING: MLE (WITH AN APPLICATION TO BINARY-CHOICE MODELS) AND GMM (WITH AN APPLICATION TO IV REGRESSION) 647
19.1 Introduction 647
19.2 Maximum Likelihood Estimation of a Simple Bivariate Regression Model 648
19.3 Maximum Likelihood Estimation of Binary-Choice Regression Models 653
19.4 Generalized Method of Moments (GMM) Estimation 658
Exercises 671
ALE 19a: Probit Modeling of the Determinants of Labor Force Participation 674
Appendix 19.1: GMM Estimation of β in the Bivariate Regression Model (Optimal Penalty-Weights and Sampling Distribution) 678

Chapter 20 CONCLUDING COMMENTS 681
20.1 The Goals of This Book 681
20.2 Diagnostic Checking and Model Respecification 683
20.3 The Four "Big Mistakes" 685

Mathematics Review **693**
Index **699**

WHAT'S DIFFERENT ABOUT THIS BOOK

THE PURPOSE OF THE KIND OF ECONOMETRICS COURSE EMBODIED IN THIS BOOK

Econometrics is all about quantifying and testing economic relationships, using sample data which is most commonly not experimentally derived. Our most fundamental tool in this enterprise is simple multiple regression analysis, although we often need to transcend it, in the end, so as to deal with such real-world complications as endogeneity in the explanatory variables, binary-choice models, and the like.

Therefore, the econometrics course envisioned in the construction of this book focuses on helping a student to develop as clear and complete an understanding of the multiple regression model as is possible, given the structural constraints – discussed below – which most instructors face. The goals of this course are to teach the student how to

- Analyze actual economic data so as to produce a statistically adequate model
- Check the validity of the statistical assumptions underlying the model, using the sample data itself and revising the model specification as needed
- Use the model to obtain reasonably valid statistical tests of economic theory – i.e., of our understanding of the economic reality generating the sample data
- Use the model to obtain reasonably valid confidence intervals for the key coefficients, so that the estimates can be sensibly used for policy analysis
- Identify, estimate, and diagnostically check practical time-series forecasting models

The emphasis throughout this book is on empowering the student to thoroughly understand the most fundamental econometric ideas and tools, rather than simply accepting a collection of assumptions, results, and formulas on faith and then using computer software to estimate a lot of regression models. The intent of the book is to well serve both the student whose interest is in understanding how one can use sample data to illuminate/suggest/test economic theory *and* the student who wants and needs a solid intellectual foundation on which to build practical experiential expertise in econometric modeling and time-series forecasting.

REAL-WORLD CONSTRAINTS ON SUCH A COURSE

The goals described above are a very tall order in the actual academic settings of most basic econometrics courses. In addition to the limited time allotted to a typical such course – often just a single term – the reality is that the students enter our courses with highly heterogeneous (and often quite spotty) statistics backgrounds. A one-term introductory statistics course is almost always a course prerequisite, but the quality and focus of this statistics course is usually outside our control. This statistics course is also often just a distant memory by the time our students reach us. Moreover, even when the statistics prerequisite course is both recent and appropriately focused for the needs of our course, many students need a deeper understanding of basic statistical concepts than they were able to attain on their first exposure to these ideas.

In addition, of course, most undergraduate (and many graduate-level) econometrics courses must do without matrix algebra, since few students in their first econometrics course are sufficiently comfortable with this tool that its use clarifies matters rather than erecting an additional conceptual barrier. Even where students are entirely comfortable with linear algebra – as might well be the case in the first term of a high-quality Ph.D.-level econometrics sequence – a treatment which eschews the use of linear algebra can be extremely useful as complement to the kind of textbook typically assigned in such a course.

Therefore the design constraints on this book are threefold:

1. The probability and statistics concepts needed are all developed within the text itself: in Chapters 2 through 4 for the most fundamental part of the book (where the regression explanatory variables are fixed in repeated samples) and in Chapter 11 for the remainder of the book.
2. Linear algebra is not used at all – nary a matrix appears (outside of a very occasional footnote) until Appendix 19.1 at the very close of the book.[1]
3. Nevertheless, the focus is on teaching an understanding of the theory underlying modern econometric techniques – not just the mechanics of invoking them – so that the student can apply these techniques with both competence and confidence.

FINESSING THE CONSTRAINTS

This book deals with the linear algebra constraint by focusing primarily on a very thorough treatment of the bivariate regression model. This provides a strong foundation, from which multiple regression analysis can be introduced – without matrix algebra – in a less detailed way. Moreover, it turns out that the essential features of many advanced topics – e.g., instrumental variables estimation – can be brought out quite clearly in a bivariate formulation.[2]

The problem with the students' preparation in terms of basic probability theory and statistics is finessed in two ways. First, Chapter 2 provides a concise review of all the probability theory needed for analyzing regression models with fixed regressors, starting at the very beginning: with the definition of a random variable, its expectation, and its variance. The seamless integration of this material into the body of the text admits of a sufficiently complete presentation as to allow students with weak (or largely forgotten) preparation to catch up. It also provides textbook "backup" for an instructor, who can then pick and choose which topics to cover in class.

[1] The necessary elements of scalar algebra – i.e., the mechanics of dealing with summation notation – are summarized in a "Mathematics Review" section at the end of the book.

[2] This strategy does not eliminate the need for linear algebra in deriving the distribution of S^2, the usual estimator of the variance of the model error term. That problem is dealt with in Chapter 4 using a large-sample argument. Occasional references to particular matrices (e.g., the usual X matrix in the multiple regression model) or linear algebraic concepts (e.g., the rank of a matrix) necessarily occur, but are relegated to footnotes.

Second, the treatment here frames the linear regression model as an explicit parameterization of the conditional mean of the dependent variable – plus, of course, a model error term. From this point of view it is natural to initially focus (in Chapters 3 and 4) on what one might call the "univariate regression model":

$$Y_i = \alpha + U_i \quad U_i \sim \text{NIID}(0, \sigma^2)$$

The estimation of the parameters α and σ^2 in this model is essentially identical to the typical introductory-statistics-course topic of estimating the mean and variance of a normally distributed random variable. Consequently, using this "univariate regression model" to begin the coverage of the essential topics in regression analysis – the least squares estimator, its sampling distribution, its desirable properties, and the inference machinery based on it – provides a thorough and integrated review of the key topics which the students need to have understood (and retained) from their introductory statistics class. It also provides an extension, in the simplest possible setting, to key concepts – e.g., estimator properties – which are usually not covered in an introductory statistics course.

Bivariate and multiple regression analysis are then introduced in the middle part of the book (Chapters 5 through 10) as a relatively straightforward extension to this framework – directly exploiting the vocabulary, concepts, and techniques just covered in this initial analysis. The always-necessary statistics "review" is in this way gracefully integrated with the orderly development of the book's central topic.

The treatment of stochastic regressors requires the deeper understanding of asymptotic theory provided in Chapter 11; this material provides a springboard for the more advanced material which makes up the rest of the book. This portion of the book is ideal for the second term of an undergraduate econometrics sequence, a Master's degree level course, or as a companion (auxiliary) text in a first-term Ph.D. level course.[3]

A CHAPTER-BY-CHAPTER ROADMAP

After an introductory chapter, the concepts of basic probability theory needed for Chapters 3 through 10 are briefly reviewed in Chapter 2. As noted above, classroom coverage of much of this material can be skipped for relatively well prepared groups; it is essential, however, for students with weak (or half-forgotten) statistics backgrounds. The most fundamentally necessary tools are a clear understanding of what is meant by the probability distribution, expected value, and variance of a random variable. These concepts are developed in a highly accessible fashion in Chapter 2 by initially focusing on a discretely distributed random variable.

As noted above, Chapter 3 introduces the notion of a parameter estimator and its sampling distribution in the simple setting of the estimation of the mean of a normally distributed variate using a random sample. Both least squares estimation and estimator properties are introduced in this chapter. Chapter 4 then explains how one can obtain interval estimates and hypothesis tests regarding the population mean, again in this fundamental context.

Chapters 3 and 4 are the first point at which it becomes crucial to distinguish between an estimator as a random variable (characterized by its sampling distribution) and its sample realization – an ordinary number. One of the features of this book is that this distinction is explicitly incorporated in the notation used. This distinction is consistently maintained throughout – not just for estimators, but for all of the various kinds of random variables that come up in the development: dependent

[3] Thus, in using this book as the text for a one-term undergraduate course, an instructor might want to order copies of the book containing only Chapter 1 through 12 and Chapter 20. This can be easily done using the Wiley "Custom Select" facility at the customselect.wiley.com Web site.

variables, model error terms, and even model fitting errors. A summary of the notational conventions used for these various kinds of random variables (and their sample realizations) is given in the "Notation" section, immediately prior to Part I of the book. In helping beginners to keep track of which variables are random and which are not, this consistent notation is well worth the additional effort involved.

While Chapters 3 and 4 can be viewed as a carefully integrated "statistics review," most of the crucial concepts and techniques underlying the regression analysis covered in the subsequent chapters are first thoroughly developed here:

- What constitutes a "good" parameter estimator?
- How do the properties (unbiasedness, BLUness, etc.) embodying this "goodness" rest on the assumptions made?
- How can we obtain confidence intervals and hypothesis tests for the underlying parameters?
- How does the validity of this inference machinery rest on the assumptions made?

After this preparation, Part II of the book covers the basics of regression analysis. The analysis in Chapter 5 coherently segues – using an explicit empirical example – from the estimation of the mean of a random variable into the particular set of assumptions which is here called "The Bivariate Regression Model," where the (conditional) mean of a random variable is parameterized as a linear function of observed realizations of an explanatory variable. In particular, what starts out as a model for the mean of per capita real GDP (from the Penn World Table) becomes a regression model relating a country's output to its aggregate stock of capital. A microeconometric bivariate regression application later in Chapter 5 relates household weekly earnings (from the Census Bureau's Current Population Survey) to a college-graduation dummy variable. This early introduction to dummy variable regressors is useful on several grounds: it both echoes the close relationship between regression analysis and the estimation of the mean (in this case, the estimation of two means) and it also introduces the student early on to an exceedingly useful empirical tool.[4]

The detailed coverage of the Bivariate Regression Model then continues with the exposition (in Chapter 6) of how the model assumptions lead to least-squares parameter estimators with desirable properties and (in Chapter 7) to a careful derivation of how these assumptions yield confidence intervals and hypothesis tests. These results are all fairly straightforward extensions of the material just covered in Chapters 3 and 4. Indeed, that is the raison d'être for the coverage of this material in Chapters 3 and 4: it makes these two chapters on bivariate regression the *second* pass at this material. Topics related to goodness of fit (R^2) and simple prediction are covered in Chapter 8.

Chapter 9 develops these same results for what is here called "The Multiple Regression Model," as an extension of the analogous results obtained in detail for the Bivariate Regression Model. While the mathematical analysis of the Multiple Regression Model is necessarily limited here by the restriction to scalar algebra, the strategy is to leverage the thorough understanding of the Bivariate Regression Model gained in the previous chapters as much as is possible toward understanding the corresponding aspects of the Multiple Regression Model. A careful – albeit necessarily, at times, intuitive – discussion of several topics which could not be addressed in the exposition of the Bivariate Regression Model completes the exposition in Chapter 9. These topics include the issues arising from over-elaborate model specifications, underelaborate model specifications, and multicollinearity. This chapter closes with several worked applications and several directed applications ("Active Learning Exercises," discussed below) for the reader to pursue.

[4] Chapter 5 also makes the link – both numerically (in Active Learning Exercise 5d) and analytically (in Appendix 5.1) – between the estimated coefficient on a dummy variable regressor and sample mean estimates. This linkage is useful later on (in Chapter 15) when the fixed-effects model for panel data is discussed.

By this point in the book it is abundantly clear how the quality of the model parameter estimates and the validity of the statistical inference machinery both hinge on the model assumptions. Chapter 10 (and, later, Chapters 13 through 15) provide a coherent summary of how one can, with a reasonably large data set, in practice use the sample data to check these assumptions. Many of the usual methods aimed at testing and/or correcting for failures in these assumptions are in essence described in these chapters, but the emphasis is not on an encyclopedia-like coverage of all the specific tests and procedures in the literature. Rather, these chapters focus on a set of graphical methods (histograms and plots) and on a set of simple auxiliary regressions which together suggest revisions to the model specification that are likely to lead to a model which at least approximately satisfies the regression model assumptions.

In particular, Chapter 10 deals with the issues – gaussianity, homoscedasticity, and parameter stability – necessary in order to diagnostically check (and perhaps respecify) a regression model based on cross-sectional data. Robust (White) standard error estimates are obtained in a particularly transparent way, but the emphasis is on taking observed heteroscedasticity as a signal that the form of the dependent variable needs respecification, rather than on FGLS corrections or on simply replacing the usual standard error estimates by robust estimates. The material in this chapter suffices to allow the student to get started on a range of practical applications.[5]

The remaining portion of Part II – comprising Chapters 11 through 14 – abandons the rather artificial assumption that the explanatory variables are fixed in repeated samples. Stochastic regressors are, of course, necessary in order to deal with the essential real-world complications of endogeneity and dynamics, but the analysis of models with stochastic regressors requires a primer on asymptotic theory. Chapter 11 provides this primer and focuses on endogeneity; Chapter 12 focuses on instrumental variables estimation; and Chapters 13 and 14 focus on diagnostically checking the nonautocorrelation assumption and on modeling dynamics.

Each of these chapters is described in more detail below, but they all share a common approach in terms of the technical level of the exposition: The (scalar) algebra of probability limits is laid out – without proof – in Appendix 11.1; these results are then used in each of the chapters to rather easily examine the consistency (or otherwise) of the OLS slope estimator in the relevant bivariate regression models. Technical details are carefully considered, but relegated to footnotes. And the asymptotic sampling distributions of these slope estimators are fairly carefully derived, but these derivations are provided in chapter appendices. This approach facilitates the coverage of the basic econometric issues regarding endogeneity and dynamics in a straightforward way, while also allowing an instructor to easily fold in a more rigorous treatment, where the time available (and the students' preparation level) allows.

Chapter 11 examines how each of the three major sources of endogeneity – omitted variables, measurement error, and joint determination – induces a correlation between an explanatory variable and the model error. In particular, simultaneous equations are introduced at this point using the simplest possible economic example: a just-identified pair of supply and demand equations.[6] The chapter ends with a brief introduction to simulation methods (with special attention to the bootstrap and its implementation in Stata), in the context of answering the perennial question about asymptotic methods, "How large a sample is really necessary?"

Chapter 12 continues the discussion of endogeneity initiated in Chapter 11 – with particular emphasis on the "reverse causality" source of endogeneity and on the non-equivalence of

[5] In particular, see Active Learning Exercises 10b and 10c in the Table of Contents. Also, even though their primary focus is on 2SLS, students can begin working on the OLS-related portions of Active Learning Exercises 12a, 12b, and 12c at this point.

[6] Subsequently – in Chapter 12, where instrumental variables estimation is covered – 2SLS is heuristically derived and applied to either a just-identified or an over-identified equation from a system of simultaneous equations. The development here does not dwell on the order and rank conditions for model identification, however.

correlation and causality. Instrumental variables estimation is then developed as the solution to the problem of using a single (valid) instrument to obtain a consistent estimator of the slope coefficient in the Bivariate Regression Model with an endogenous regressor. The approach of restricting attention to this simple model minimizes the algebra needed and leverages the work done in Chapter 11. A derivation of the asymptotic distribution of the instrumental variables estimator is provided in Appendix 12.1, giving the instructor a graceful option to either cover this material or not. The two-stage least squares estimator is then heuristically introduced and applied to the classic Angrist-Krueger (1991) study of the impact of education on log-wages. Several other economic applications, whose sample sizes are more feasible for student-version software, are given as Active Learning Exercises at the end of the chapter.

Attention then shifts, in a pair of chapters – Chapters 13 and 14 – to time-series issues. Because Chapters 17 and 18 cover forecasting in some detail, Chapters 13 and 14 concentrate on the estimation and inference issues raised by time-series data.[7] The focus in Chapter 13 is on how to check the non-autocorrelation assumption on the regression model errors and deal with any violations. The emphasis here is not on named tests (in this case, for serially correlated errors) or on assorted versions of FGLS, but rather on how to sensibly respecify a model's dynamics so as to reduce or eliminate observed autocorrelation in the errors. Chapter 14 then deals with the implementation issues posed by integrated (and cointegrated) time-series, including the practical decision as to whether it is preferable to model the data in levels versus in differences. The "levels" versus "changes" issue is first addressed at this point, in part using insights gained from simulation work reported in Ashley and Verbrugge (2009). These results indicate that it is usually best to model in levels, but to generate inferential conclusions using a straightforward variation on the Lag-Augmented VAR approach of Toda and Yamamoto (1995).[8] On the other hand, the differenced data is easier to work with (because it is far less serially dependent) and it provides the opportunity (via the error-correction formulation) to dis-entangle the long-run and short-run dynamics. Thus, in the end, it is probably best to model the data both ways.[9] This synthesis of the material is carefully developed in the context of a detailed analysis of an illustrative empirical application: modeling monthly U.S. consumption expenditures data. This example also provides a capstone illustration of the diagnostic checking techniques described here.

The last portion of the book (Part III) consists of five chapters on advanced topics and a concluding chapter. These five "topics" chapters will be particularly useful for instructors who are able to move through Chapters 2 through 4 quickly because their students are well prepared; the "Concluding Comments" chapter – Chapter 20 – will be useful to all. Chapters 15 and 16 together provide a brief introduction to the analysis of panel data, and Chapters 17 and 18 together provide a concise introduction to the broad field of time-series analysis and forecasting. Chapter 19 introduces the two main alternatives to OLS for estimating parametric regression models: maximum likelihood estimation (MLE) and the generalized method of moments (GMM). Each of these chapters is described in a bit more detail below.

A great deal of micro-econometric analysis is nowadays based on panel data sets. Chapters 15 and 16 provide a straightforward, but comprehensive, treatment of panel data methods. The issues, and requisite panel-specific methods, for the basic situation – with strictly exogenous explanatory variables – are first carefully explained in Chapter 15, all in the context of an empirical example. This material

[7] Most of the usual (and most crucial) issues in using regression models for prediction are, in any case, covered much earlier – in Section 8.3.

[8] See Ashley, R., and R. Verbrugge (2009), "To Difference or Not to Difference: A Monte Carlo Investigation of Inference in Vector Autoregression Models." *International Journal of Data Analysis Techniques and Strategies1(3)*: 242–274 (ashley-mac.econ.vt.edu/working_papers/varsim.pdf) and Toda, H. Y., and T. Yamamoto (1995), "Statistical Inference in Vector Autoregressions with Possibly Integrated Processes," *J. Econometrics 66*, 225–250.

[9] The "difference" versus "detrend" issue comes up again in Section 18.1, where it is approached (and resolved) a bit differently, from a "time-series analysis" rather than a "time-series econometrics" perspective.

concentrates on the Fixed Effects and then on the Random Effects estimators. Then dynamics, in the form of lagged dependent variables, are added to the model in Chapter 16. (Many readers will be a bit surprised to find that the Random Effects estimator is still consistent in this context, so long as the model errors are homoscedastic and any failures in the strict exogeneity assumption are not empirically consequential.) Finally, the First-Differences model is introduced for dealing with endogeneity (as well as dynamics) via instrumental variables estimation. This IV treatment leads to an unsatisfactory 2SLS estimator, which motivates a detailed description of how to apply the Arellano-Bond estimator in working with such models. The description of the Arellano-Bond estimator does not go as deep (because GMM estimation is not covered until Chapter 19), but sufficient material is provided that the student can immediately begin working productively with panel data.

The primary focus of much applied economic work is on inferential issues – i.e., on the statistical significance of the estimated parameter on a particular explanatory variable whose inclusion in the model is prescribed by theory, or on a 95% confidence interval for a parameter whose value is policy-relevant. In other applied settings, however, forecasting is paramount. Chapters 17 and 18, which provide an introduction to the broad field of time-series analysis and forecasting, are particularly useful in the latter context. Chapter 17 begins with a careful treatment of forecasting theory, dealing with the fundamental issue of when (and to what extent) it is desirable to forecast with the conditional mean. The chapter then develops the basic tools – an understanding of the sample correlogram and the ability to invert a lag structure – needed in order to use Box-Jenkins (ARMA) methods to identify, estimate, and diagnostically check a univariate linear model for a time-series and to then obtain useful short-term conditional mean forecasts from it. These ideas and techniques are then extended – in Chapter 18 – to a variety of extensions of this framework into multivariate and nonlinear time-series modeling.

Up to this point in the book, regression analysis is basically framed in terms of least-squares estimation of parameterized models for the conditional mean of the variable whose sample fluctuations are to be "explained." As explicitly drawn out for the Bivariate Regression Model in Chapter 5, this is equivalent to fitting a straight line to a scatter diagram of the sample data.[10] Chapter 19 succinctly introduces the two most important parametric alternatives to this "curve-fitting" approach: maximum likelihood estimation and the generalized method of moments.

In the first part of Chapter 19 the maximum likelihood estimation framework is initially explained – as was least squares estimation in Part I of the book – in terms of the simple problem of estimating the mean and variance of a normally distributed variable. The primary advantage of the MLE approach is its ability to handle latent variable models, so a second application is then given to a very simple binary-choice regression model. In this way, the first sections of Chapter 19 provide a practical introduction to the entire field of "limited dependent variables" modeling.

The remainder of Chapter 19 provides an introduction to the Generalized Method of Moments (GMM) modeling framework. In the GMM approach, parameter identification and estimation are achieved through matching posited population moment conditions to analogous sample moments, where these sample moments depend on the coefficient estimates. The GMM framework thus directly involves neither least-squares curve-fitting nor estimation of the conditional mean. GMM is really the only graceful approach for estimating a rational expectations model via its implied Euler equation. Of more frequent relevance, it is currently the state-of-the-art approach for estimating IV regression models, especially where heteroscedastic model errors are an issue. Chapter 19 introduces GMM via a detailed description of the simplest non-trivial application to such an IV regression model: the one-parameter, two-instrument case. The practical application of GMM estimation is then illustrated using a

[10] The analogous point, using a horizontal straight line "fit" to a plot of the sample data versus observation number, is made in Chapter 3. And the (necessarily more abstract) extension to the fitting of a hyperplane to the sample data is described in Chapter 9. The corresponding relationship between the estimation of a parameterization of the conditional median of the dependent variable and estimation via least absolute deviations fitting is briefly explained in each of these cases also.

familiar full-scale empirical model, the well-known Angrist-Krueger (1991) model already introduced in Chapter 12: in this model there are 11 parameters to be estimated, using 40 moment conditions.

Even the simple one-parameter GMM estimation example, however, requires a linear-algebraic formulation of the estimator. This linear algebra (its only appearance in the book) is relegated to Appendix 19.1, where it is unpacked for this example. But this exigency marks a natural stopping-point for the exposition given here. Chapter 20 concludes the book with some sage – if, perhaps, opinionated – advice.

A great deal of important and useful econometrics was necessarily left out of the present treatment. Additional topics (such as nonparametric regression, quantile regression, Bayesian methods, and additional limited dependent variables models) could perhaps be covered in a subsequent edition.

WITH REGARD TO COMPUTER SOFTWARE

While sample computer commands and examples of the resulting output – mostly using Stata, and very occasionally using Eviews – are explicitly integrated into the text, this book is not designed to be a primer on any particular econometrics software package. There are too many different programs in widespread use for that to be useful. In any case, most students are rather good at learning the mechanics of software packages on their own. Instead, this book is more fundamentally designed, to help students develop a confident understanding of the part they often have great difficulty learning on their own: the underlying theory and practice of econometrics.

In fact, generally speaking, learning how to instruct the software to apply various econometric techniques to the data is not the tough part of this topic. Rather, the challenge is in in learning how to decide which techniques to apply and how to interpret the results. Consequently, the most important object here is to teach students how to become savvy, effective users of whatever software package comes their way. Via an appropriate amount of econometric theory (which is especially modest up through Chapter 10), a sequence of detailed examples, and exercises using actual economic data, this book can help an instructor equip students to tackle real-world econometric modeling using any software package.

In particular – while no knowledgeable person would choose Excel as an econometrics package – it is even possible to teach a good introductory econometrics course using Parts I and II of this book in conjunction with Excel. The main limitation in that case, actually, is that students would not themselves be able to compute the White-Eicker robust standard error estimates discussed in Chapter 10.[11]

An instructor using Stata, however, will find this book particularly easy to use, in that the appropriate implementing Stata commands are all noted, albeit sometimes (in Part I) using footnotes. It should not be at all difficult, however, to convert these into analogous commands for other packages, as the essential content here lies in explaining what one is asking the software to do – and why. Also, all data sets are supplied as comma-delimited (*.csv) files – as well as in Stata's proprietary format – so that any econometric software program can easily read them.

WITH REGARD TO STATISTICAL TABLES

Where a very brief table containing a few critical points is needed in order to illustrate a particular point, such a table is integrated right into the text. In Table 4-1 of Chapter 4, for example, a tabulation of a handful of critical points for the Student's t distribution exhibits the impact on the length of an estimated 95% confidence interval (for the mean of a normally distributed variate) of having to estimate its variance using a limited sample of data.

[11] And, of course, it is well known that Excel's implementation of multiple regression is not numerically well-behaved.

In general, however, tables of tail areas and critical points for the normal, χ^2, Student's t, and F distribution are functionally obsolete – as is the skill of reading values off of them. Ninety-nine times out of a hundred, the econometric software in use computes the necessary p-values for us: the valuable skill is in understanding the assumptions underlying their calculation and how to diagnostically check these assumptions. And, in the one-hundredth case, it is a matter of moments to load up a spreadsheet – e.g., Excel – and calculate the relevant tail area or critical point using a worksheet function.[12]

Consequently, this book does not included printed statistical tables.

SUPPLEMENTARY MATERIALS

A number of supplementary materials are posted on the companion Web site for this book, www.wiley.com/college/ashley. These include:

- Active Learning Exercises listed in the Table of Contents, including their accompanying data sets and any computer programs needed. Answer keys for these Exercises are posted also.
- Answer keys for all of the end-of-chapter exercises.
- Windows programs which compute tail areas for the normal, χ^2, t, and F distributions.
- PowerPoint slides for each chapter.
- Image Gallery – equations, tables, and figures – in JPEG format for each chapter. Sample presentation files based on these, in Adobe Acrobat PDF format, are also provided for each chapter.

HETEROGENEITY IN LEARNING STYLES

Some students learn best by reading a coherent description of the ideas, techniques, and applications in a textbook. Other students learn best by listening to an instructor work through a tough section and asking questions. Still other students learn best by working homework exercises, on their own or in groups, which deepen their understanding of the material. Most likely, every student needs all of these course components, in individually specific proportions.

In recognition of the fact that many students need to "do something" in order to really engage with the material, the text is peppered with what are here called "Active Learning Exercises." These are so important that the next section is devoted to describing them.

[12] The syntax for the relevant Excel spreadsheet function syntax is quoted in the text where these arise, as is a citation to a standard work quoting the computing approximations used in these worksheet functions. Stand-alone Windows programs implementing these approximations are posted at Web site www.wiley.com/college/ashley.

WORKING WITH DATA IN THE "ACTIVE LEARNING EXERCISES"

Most chapters of this textbook contain at least one "Active Learning Exercise" or "ALE." The titles of these Active Learning Exercises are given in the Table of Contents and listed on the inside covers of the book. Whereas the purpose of the end-of-chapter exercises is to help the student go deeper into the chapter material – and worked examples using economic data are integrated into the text – these Active Learning Exercises are designed to engage the student in structured, active exercises.

A typical Active Learning Exercise involves specific activities in which the student is either directed to download actual economic data from an academic/government Web site or is provided with data (real or simulated) from the companion Web site for this book, www.wiley.com/college/ashley. (This Web site will also provide access to the latest version of each Active Learning Exercise, as some of these exercises will need to be revised occasionally as Web addresses and content change.) These exercises will in some cases reproduce and/or expand on empirical results used as examples in the text; in other cases, the Active Learning Exercise will set the student working on new data. A number of the Active Learning Exercises involve replication of a portion of the empirical results of published articles from the economics literature.

The Active Learning Exercises are a more relaxed environment than the text itself, in that one of these exercises might, for example, involve a student in "doing" multiple regression in an informal way long before this topic is reached in the course of the careful development provided in the text. One could think of these exercises as highly structured "mini-projects." In this context, the Active Learning Exercises are also a great way to help students initiate their own term projects.

ACKNOWLEDGMENTS

My thanks to all of my students for their comments on various versions of the manuscript for this book; in particular, I would like to particularly express my appreciation to Bradley Shapiro and to James Boohaker for their invaluable help with the end-of-chapter exercises. Thanks are also due to Alfonso Flores-Lagunes, Chris Parmeter, Aris Spanos, and Byron Tsang for helpful discussions and/or access to data sets. Andrew Rose was particularly forthcoming in helping me to replicate his very interesting 2005 paper with Frankel in *The Review of Economics and Statistics* quantifying the impact of international trade on environmental air quality variables; this help was crucial to the construction of Active Learning Exercises 10c and 12b. I have benefited from the comments and suggestions from the following reviewers: Alfonso Flores-Lagunes, University of Florida, Gainesville; Scott Gilbert, Southern Illinois University, Carbondale; Denise Hare, Reed College; Alfred A. Haug, University of Otago, New Zealand; Paul A. Jargowsky, Rutgers-Camden; David Kimball, University of Missouri, St. Louis; Heather Tierney, College of Charleston; Margie Tieslau, University of North Texas; and several others who wish to remain anonymous. Thanks are also due to Lacey Vitteta, Jennifer Manias, Emily McGee, and Yee Lyn Song at Wiley for their editorial assistance. Finally, I would also like to thank Rosalind Ashley, Elizabeth Paule, Bill Beville, and George Lobell for their encouragement with regard to this project.

NOTATION

Logical and consistent notation is extremely helpful in keeping track of econometric concepts, particularly the distinction between random variables and realizations of random variables. This section summarizes the principles underlying the notation used below. This material can be skimmed on your first pass: this notational material is included here primarily for reference later on, after the relevant concepts to which the notational conventions apply are explained in the chapters to come.

Uppercase letters from the usual Latin-based alphabet – X, Y, Z, etc. – are used below to denote observable data. These will generally be treated as random variables, which will be discussed in Chapter 2. What is most important here is to note that an uppercase letter will be used to denote such a random variable; the corresponding lowercase letter will be used to denote a particular (fixed) realization of it – i.e., the numeric value actually observed. Thus, "X" is a random variable, whereas "x" is a realization of this random variable. Lowercase letters will *not* be used below to denote the deviation of a variable from its sample mean.

The fixed (but unknown) parameters in the econometric models considered below will usually be denoted by lowercase Greek letters – α, β, γ, δ, and so forth. As we shall see below, these parameters will be estimated using functions of the observable data – "estimators" – which are random variables. Because uppercase Greek letters are easily confused with letters from the Latin-based alphabet, however, such an estimator of a parameter – a random variable because it depends on the observable data, which are random variables – will typically be denoted by placing a hat ("^") over the corresponding lowercase Greek letter. Sample realizations of these parameter estimators will then be denoted by appending an asterisk. Thus, $\hat{\alpha}$ will typically be used to denote an estimator of the fixed parameter α and $\hat{\alpha}^*$ will be used to denote the (fixed) realization of this random variable, based on the particular values of the observable data which were actually observed. Where a second estimator of α needs to be considered, it will be denoted by $\tilde{\alpha}$ or the like. The only exceptions to these notational conventions which you will encounter later are that – so as to be consistent with the standard nomenclature – the usual convention of using $\bar{Y}$ and S^2 to denote the sample mean and variance will be used; sample realizations of these estimators will be denoted $\bar{y}$ and s^2, respectively.

The random error terms in the econometric models developed below will be denoted by uppercase letters from the Latin-based alphabet (typically, U, V, N, etc.) and fixed realizations of these error terms (which will come up very infrequently because model error terms are not, in practice, observable) will be denoted by the corresponding lowercase letter, just as with observable data.

When an econometric model is fit to sample data, however, one obtains observable "fitting errors." These can be usefully thought of as estimators of the model errors. These estimators – which will be random variables because they depend on the observable (random) observations – will be distinguished from the model errors themselves via a superscript "fit" on the corresponding letter for the model error. As with the model errors, the sample realizations of these fitting errors, based on particular realizations of the observable data, will be denoted by the corresponding lowercase letter.

The following table summarizes these notational rules and gives some examples:

	Random Variable	Realization
observable data (*i*th observation)	X_i, Y_i, Z_i	x_i, y_i, z_i
parameter estimator	$\hat{\alpha}, \hat{\beta}, \hat{\mu}, \bar{Y}, S^2$	$\hat{\alpha}^*, \hat{\beta}^*, \hat{\mu}^*, \bar{y}, s^2$
model error (*i*th observation)	U_i, V_i	u_i, v_i
model fitting error (*i*th observation)	$U_i^{\text{fit}}, V_i^{\text{fit}}$	$u_i^{\text{fit}}, v_i^{\text{fit}}$

Fundamentals of applied econometrics / by Richard Ashley. â€“ 1st ed. p. cm. Includes index. ISBN 978-0-470-59182-6 (hardback). 1. Econometrics. 2. Econometricsâ€“Statistical methods. 3. Econometricsâ€“Data processing. I. Title. HB139.A84 2012 330.0105195â€“dc23.Â Therefore, the econometrics course envisioned in the construction of this book focuses on helping a student to develop as clear and complete an understanding of the multiple regression model as is possible, given the structural constraints â€“ discussed below â€“ which most instructors face. The goals of this course are to teach the student how to Analyze actual economic data so as to produce a statistically adequate model Check the validity of the statistical assumptions underlying the model, using the sample data. Teaching: Econometrics and Basic Econometrics, Econometric Theory, Applied Econometrics, Micro-econometrics, and supervision of graduate and Ph.D. students.Â Doctor in economics mathematics and econometrics, with the aggregation of economics, Patrick Sevestre is Professor of Economics at the University Paris XII -Val de Marne (1994). He is also research fellow at l'ERUDITE, University Paris XII - Val de Marne, Consultant in the reserach center of Banque de France (since 1994) and teacher to the CEPE, Centre d'Etude des Programmes Economiques (1991-1994 and since1996). What s Different about Thi' Book xiii Working with Data in the "Active Learning Exercises" xxii Acknowledgments xxiii Notation xxiv Part I. Introduction and Statistics Review 1 Chapter 1. Introduction 3 Chapter 2. A Review of Probability Theory 11 Chapter 3. Estimating the Mean of a Normally Distributed Random Variable 46 Chapter 4. Statistical Inference on the Mean of a Normally Distributed Random Variable 68 Part II. Regression Analysis 97 Chapter 5. The Bivariate Regression Model: Introduction, Assumptions, and Parameter Estimates 99 Chapter 6. The Bivariate Linear Regression