Contents

Preface 1

1. Introduction **3**
1.1. Getting Started 3
1.1.1. Software 3
1.1.2. R Scripts 4
1.1.3. Packages 7
1.1.4. File names and the Working Directory 8
1.1.5. Errors and Warnings 9
1.1.6. Other Resources 9
1.2. Objects in *R* 10
1.2.1. Basic Calculations and Objects 10
1.2.2. Vectors 12
1.2.3. Special Types of Vectors . . 14
1.2.4. Naming and Indexing Vectors 15
1.2.5. Matrices 16
1.2.6. Lists 19
1.3. Data Frames and Data Files 20
1.3.1. Data Frames 20
1.3.2. Subsets of Data 21
1.3.3. *R* Data Files 22
1.3.4. Basic Information on a Data Set 22
1.3.5. Import and Export of Text Files 23
1.3.6. Import and Export of Other Data Formats 24
1.3.7. Data Sets in the Examples . 25
1.4. Graphics 26
1.4.1. Basic Graphs 26
1.4.2. Customizing Graphs with Options 28
1.4.3. Overlaying Several Plots . . 29
1.4.4. Legends 30
1.4.5. Exporting to a File 32
1.4.6. Advanced Graphs 33
1.5. Descriptive Statistics 34
1.5.1. Discrete Distributions: Frequencies and Contingency Tables 34
1.5.2. Continuous Distributions: Histogram and Density . . 37
1.5.3. Empirical Cumulative Distribution Function (ECDF) . 38
1.5.4. Fundamental Statistics . . . 40
1.6. Probability Distributions 41
1.6.1. Discrete Distributions . . . 41
1.6.2. Continuous Distributions . 43
1.6.3. Cumulative Distribution Function (CDF) 44
1.6.4. Random Draws from Probability Distributions 45
1.7. Confidence Intervals and Statistical Inference 47
1.7.1. Confidence Intervals 47
1.7.2. *t* Tests 50
1.7.3. *p* Values 51
1.7.4. Automatic calculations . . . 52
1.8. Advanced *R* 56
1.8.1. Conditional Execution . . . 56
1.8.2. Loops 56
1.8.3. Functions 57
1.8.4. Outlook 57
1.9. Monte Carlo Simulation 58
1.9.1. Finite Sample Properties of Estimators 58
1.9.2. Asymptotic Properties of Estimators 61
1.9.3. Simulation of Confidence Intervals and *t* Tests 64

I. Regression Analysis with Cross-Sectional Data **67**

2. The Simple Regression Model **69**
2.1. Simple OLS Regression 69
2.2. Coefficients, Fitted Values, and Residuals 74
2.3. Goodness of Fit 77
2.4. Nonlinearities 79
2.5. Regression through the Origin and Regression on a Constant 80
2.6. Expected Values, Variances, and Standard Errors 82
2.7. Monte Carlo Simulations 84
2.7.1. One sample 84
2.7.2. Many Samples 86

2.7.3. Violation of SLR.4 89
2.7.4. Violation of SLR.5 89

3. Multiple Regression Analysis: Estimation 91
3.1. Multiple Regression in Practice . . 91
3.2. OLS in Matrix Form 95
3.3. Ceteris Paribus Interpretation and Omitted Variable Bias 97
3.4. Standard Errors, Multicollinearity, and VIF 99

4. Multiple Regression Analysis: Inference 103
4.1. The t Test 103
4.1.1. General Setup 103
4.1.2. Standard case 104
4.1.3. Other hypotheses 106
4.2. Confidence Intervals 108
4.3. Linear Restrictions: F-Tests 109
4.4. Reporting Regression Results . . . 113

5. Multiple Regression Analysis: OLS Asymptotics 115
5.1. Simulation Exercises 115
5.1.1. Normally Distributed Error Terms 115
5.1.2. Non-Normal Error Terms . 116
5.1.3. (Not) Conditioning on the Regressors 119
5.2. LM Test 121

6. Multiple Regression Analysis: Further Issues 123
6.1. Model Formulae 123
6.1.1. Data Scaling: Arithmetic Operations Within a Formula 123
6.1.2. Standardization: Beta Coefficients 125
6.1.3. Logarithms 126
6.1.4. Quadratics and Polynomials 126
6.1.5. Interaction Terms 128
6.2. Prediction 130
6.2.1. Confidence Intervals for Predictions 130
6.2.2. Prediction Intervals 132
6.2.3. Effect Plots for Nonlinear Specifications 133

7. Multiple Regression Analysis with Qualitative Regressors 135
7.1. Linear Regression with Dummy Variables as Regressors 135
7.2. Logical Variables 137
7.3. Factor variables 138
7.4. Breaking a Numeric Variable Into Categories 139
7.5. Interactions and Differences in Regression Functions Across Groups 141

8. Heteroscedasticity 143
8.1. Heteroscedasticity-Robust Inference 143
8.2. Heteroscedasticity Tests 147
8.3. Weighted Least Squares 150

9. More on Specification and Data Issues 155
9.1. Functional Form Misspecification . 155
9.2. Measurement Error 157
9.3. Missing Data and Nonrandom Samples 160
9.4. Outlying Observations 163
9.5. Least Absolute Deviations (LAD) Estimation 164

II. Regression Analysis with Time Series Data 165

10. Basic Regression Analysis with Time Series Data 167
10.1. Static Time Series Models 167
10.2. Time Series Data Types in R 168
10.2.1. Equispaced Time Series in R 168
10.2.2. Irregular Time Series in R . 170
10.3. Other Time Series Models 173
10.3.1. The `dynlm` Package 173
10.3.2. Finite Distributed Lag Models 173
10.3.3. Trends 176
10.3.4. Seasonality 177

11. Further Issues In Using OLS with Time Series Data 179
11.1. Asymptotics with Time Series . . . 179
11.2. The Nature of Highly Persistent Time Series 182
11.3. Differences of Highly Persistent Time Series 185

11.4. Regression with First Differences . 186

12. Serial Correlation and Heteroscedasticity in Time Series Regressions 187
12.1. Testing for Serial Correlation of the Error Term 187
12.2. FGLS Estimation 191
12.3. Serial Correlation-Robust Inference with OLS 192
12.4. Autoregressive Conditional Heteroscedasticity 193

III. Advanced Topics 195

13. Pooling Cross-Sections Across Time: Simple Panel Data Methods 197
13.1. Pooled Cross-Sections 197
13.2. Difference-in-Differences 198
13.3. Organizing Panel Data 201
13.4. Panel-specific computations 202
13.5. First Differenced Estimator 204

14. Advanced Panel Data Methods 207
14.1. Fixed Effects Estimation 207
14.2. Random Effects Models 209
14.3. Dummy Variable Regression and Correlated Random Effects 212
14.4. Robust (Clustered) Standard Errors 216

15. Instrumental Variables Estimation and Two Stage Least Squares 219
15.1. Instrumental Variables in Simple Regression Models 219
15.2. More Exogenous Regressors 221
15.3. Two Stage Least Squares 222
15.4. Testing for Exogeneity of the Regressors 224
15.5. Testing Overidentifying Restrictions 225
15.6. Instrumental Variables with Panel Data 225

16. Simultaneous Equations Models 229
16.1. Setup and Notation 229
16.2. Estimation by 2SLS 230
16.3. Joint Estimation of System 231
16.4. Outlook: Estimation by 3SLS . . . 233

17. Limited Dependent Variable Models and Sample Selection Corrections 235
17.1. Binary Responses 235
17.1.1. Linear Probability Models . 235
17.1.2. Logit and Probit Models: Estimation 237
17.1.3. Inference 240
17.1.4. Predictions 241
17.1.5. Partial Effects 242
17.2. Count Data: The Poisson Regression Model 245
17.3. Corner Solution Responses: The Tobit Model 248
17.4. Censored and Truncated Regression Models 251
17.5. Sample Selection Corrections . . . 253

18. Advanced Time Series Topics 255
18.1. Infinite Distributed Lag Models . . 255
18.2. Testing for Unit Roots 257
18.3. Spurious Regression 260
18.4. Cointegration and Error Correction Models 262
18.5. Forecasting 263

19. Carrying Out an Empirical Project 267
19.1. Working with *R* Scripts 267
19.2. Logging Output in Text Files . . . 269
19.3. Formatted Documents and Reports with *R* Markdown 269
19.3.1. Basics 269
19.3.2. Advanced Features 270
19.3.3. Bottom Line 272
19.4. Combining *R* with LaTeX 274
19.4.1. Automatic Document Generation using `Sweave` and `knitr` 274
19.4.2. Separating *R* and LaTeX code 278

IV. Appendices 281

R Scripts 283
1. Scripts Used in Chapter 01 283
2. Scripts Used in Chapter 02 293
3. Scripts Used in Chapter 03 299
4. Scripts Used in Chapter 04 301
5. Scripts Used in Chapter 05 303
6. Scripts Used in Chapter 06 305

7. Scripts Used in Chapter 07 307
8. Scripts Used in Chapter 08 309
9. Scripts Used in Chapter 09 311
10. Scripts Used in Chapter 10 314
11. Scripts Used in Chapter 11 316
12. Scripts Used in Chapter 12 318
13. Scripts Used in Chapter 13 320
14. Scripts Used in Chapter 14 322
15. Scripts Used in Chapter 15 324
16. Scripts Used in Chapter 16 326
17. Scripts Used in Chapter 17 326
18. Scripts Used in Chapter 18 330
19. Scripts Used in Chapter 19 333

Bibliography **335**

List of Wooldridge (2016) examples **337**

Index **339**