Microeconometrics Using Stata

Revised Edition

A. COLIN CAMERON
Department of Economics
University of California
Davis, CA

PRAVIN K. TRIVEDI
Department of Economics
Indiana University
Bloomington, IN

A Stata Press Publication
StataCorp LP
College Station, Texas

Published by Stata Press, 4905 Lakeway Drive, College Station, Texas 77845
Typeset in LaTeX 2ε
Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

ISBN-10: 1-59718-073-4
ISBN-13: 978-1-59718-073-3

Contents

List of tables **xxxv**

List of figures **xxxvii**

Preface to the Revised Edition **xxxix**

Preface to the First Edition **xli**

1 Stata basics **1**

1.1 Interactive use 1

1.2 Documentation 2

1.2.1 Stata manuals 2

1.2.2 Additional Stata resources 3

1.2.3 The help command 3

1.2.4 The search, findit, and hsearch commands 4

1.3 Command syntax and operators 5

1.3.1 Basic command syntax 5

1.3.2 Example: The summarize command 6

1.3.3 Example: The regress command 7

1.3.4 Factor variables 9

1.3.5 Abbreviations, case sensitivity, and wildcards 11

1.3.6 Arithmetic, relational, and logical operators 12

1.3.7 Error messages 12

1.4 Do-files and log files 13

1.4.1 Writing a do-file 13

1.4.2 Running do-files 14

1.4.3 Log files 14

1.4.4 A three-step process 15

1.4.5 Comments and long lines . . . 16
1.4.6 Different implementations of Stata . . . 17
1.5 Scalars and matrices . . . 17
1.5.1 Scalars . . . 17
1.5.2 Matrices . . . 18
1.6 Using results from Stata commands . . . 18
1.6.1 Using results from the r-class command summarize . . . 18
1.6.2 Using results from the e-class command regress . . . 19
1.7 Global and local macros . . . 21
1.7.1 Global macros . . . 21
1.7.2 Local macros . . . 22
1.7.3 Scalar or macro? . . . 23
1.8 Looping commands . . . 24
1.8.1 The foreach loop . . . 25
1.8.2 The forvalues loop . . . 26
1.8.3 The while loop . . . 26
1.8.4 The continue command . . . 27
1.9 Some useful commands . . . 27
1.10 Template do-file . . . 27
1.11 User-written commands . . . 28
1.12 Stata resources . . . 29
1.13 Exercises . . . 29
2 Data management and graphics 31
2.1 Introduction . . . 31
2.2 Types of data . . . 31
2.2.1 Text or ASCII data . . . 32
2.2.2 Internal numeric data . . . 32
2.2.3 String data . . . 33
2.2.4 Formats for displaying numeric data . . . 33

2.3 Inputting data 34
2.3.1 General principles 34
2.3.2 Inputting data already in Stata format 35
2.3.3 Inputting data from the keyboard 36
2.3.4 Inputting nontext data 36
2.3.5 Inputting text data from a spreadsheet 37
2.3.6 Inputting text data in free format 38
2.3.7 Inputting text data in fixed format 38
2.3.8 Dictionary files 39
2.3.9 Common pitfalls 39
2.4 Data management 40
2.4.1 PSID example 40
2.4.2 Naming and labeling variables 43
2.4.3 Viewing data 44
2.4.4 Using original documentation 45
2.4.5 Missing values 45
2.4.6 Imputing missing data 47
2.4.7 Transforming data (generate, replace, egen, recode) 48
The generate and replace commands 48
The egen command 49
The recode command 49
The by prefix 49
Indicator variables 50
Set of indicator variables 50
Interactions 51
Demeaning 52
2.4.8 Saving data 52
2.4.9 Selecting the sample 53
2.5 Manipulating datasets 54
2.5.1 Ordering observations and variables 55

2.5.2 Preserving and restoring a dataset . . . 55
2.5.3 Wide and long forms for a dataset . . . 55
2.5.4 Merging datasets . . . 56
2.5.5 Appending datasets . . . 58
2.6 Graphical display of data . . . 58
2.6.1 Stata graph commands . . . 59
Example graph commands . . . 59
Saving and exporting graphs . . . 60
Learning how to use graph commands . . . 61
2.6.2 Box-and-whisker plot . . . 61
2.6.3 Histogram . . . 63
2.6.4 Kernel density plot . . . 63
2.6.5 Twoway scatterplots and fitted lines . . . 66
2.6.6 Lowess, kernel, local linear, and nearest-neighbor regression 67
2.6.7 Multiple scatterplots . . . 69
2.7 Stata resources . . . 70
2.8 Exercises . . . 70
3 Linear regression basics 73
3.1 Introduction . . . 73
3.2 Data and data summary . . . 73
3.2.1 Data description . . . 73
3.2.2 Variable description . . . 74
3.2.3 Summary statistics . . . 75
3.2.4 More-detailed summary statistics . . . 76
3.2.5 Tables for data . . . 77
3.2.6 Statistical tests . . . 80
3.2.7 Data plots . . . 80
3.3 Regression in levels and logs . . . 81
3.3.1 Basic regression theory . . . 81
3.3.2 OLS regression and matrix algebra . . . 82

3.3.3 Properties of the OLS estimator . . . 83
3.3.4 Heteroskedasticity-robust standard errors . . . 84
3.3.5 Cluster–robust standard errors . . . 84
3.3.6 Regression in logs . . . 85
3.4 Basic regression analysis . . . 86
3.4.1 Correlations . . . 86
3.4.2 The regress command . . . 87
3.4.3 Hypothesis tests . . . 88
3.4.4 Tables of output from several regressions . . . 89
3.4.5 Even better tables of regression output . . . 90
3.4.6 Factor variables for categorical variables and interactions . . . 92
3.5 Specification analysis . . . 94
3.5.1 Specification tests and model diagnostics . . . 94
3.5.2 Residual diagnostic plots . . . 95
3.5.3 Influential observations . . . 96
3.5.4 Specification tests . . . 97
Test of omitted variables . . . 98
Test of the Box–Cox model . . . 98
Test of the functional form of the conditional mean . . . 99
Heteroskedasticity test . . . 100
Omnibus test . . . 102
3.5.5 Tests have power in more than one direction . . . 102
3.6 Prediction . . . 104
3.6.1 In-sample prediction . . . 104
3.6.2 MEs and elasticities . . . 106
3.6.3 Prediction in logs: The retransformation problem . . . 108
3.6.4 Prediction exercise . . . 109
3.7 Sampling weights . . . 111
3.7.1 Weights . . . 111
3.7.2 Weighted mean . . . 112

3.7.3 Weighted regression . . . 113
3.7.4 Weighted prediction and MEs . . . 114
3.8 OLS using Mata . . . 115
3.9 Stata resources . . . 117
3.10 Exercises . . . 117
4 Simulation **119**
4.1 Introduction . . . 119
4.2 Pseudorandom-number generators: Introduction . . . 120
4.2.1 Uniform random-number generation . . . 120
4.2.2 Draws from normal . . . 122
4.2.3 Draws from t, chi-squared, F, gamma, and beta . . . 123
4.2.4 Draws from binomial, Poisson, and negative binomial . . . 124
Independent (but not identically distributed) draws from binomial . . . 124
Independent (but not identically distributed) draws from Poisson . . . 125
Histograms and density plots . . . 126
4.3 Distribution of the sample mean . . . 127
4.3.1 Stata program . . . 128
4.3.2 The simulate command . . . 129
4.3.3 Central limit theorem simulation . . . 129
4.3.4 The postfile command . . . 130
4.3.5 Alternative central limit theorem simulation . . . 131
4.4 Pseudorandom-number generators: Further details . . . 131
4.4.1 Inverse-probability transformation . . . 132
4.4.2 Direct transformation . . . 133
4.4.3 Other methods . . . 133
4.4.4 Draws from truncated normal . . . 134
4.4.5 Draws from multivariate normal . . . 135
Direct draws from multivariate normal . . . 135
Transformation using Cholesky decomposition . . . 136

4.4.6 Draws using Markov chain Monte Carlo method 136
4.5 Computing integrals 138
4.5.1 Quadrature 139
4.5.2 Monte Carlo integration 139
4.5.3 Monte Carlo integration using different S 140
4.6 Simulation for regression: Introduction 141
4.6.1 Simulation example: OLS with χ^2 errors 141
4.6.2 Interpreting simulation output 144
Unbiasedness of estimator 144
Standard errors 144
t statistic 144
Test size 145
Number of simulations 146
4.6.3 Variations 146
Different sample size and number of simulations 146
Test power 146
Different error distributions 147
4.6.4 Estimator inconsistency 147
4.6.5 Simulation with endogenous regressors 148
4.7 Stata resources 150
4.8 Exercises 150
5 GLS regression **153**
5.1 Introduction 153
5.2 GLS and FGLS regression 153
5.2.1 GLS for heteroskedastic errors 153
5.2.2 GLS and FGLS 154
5.2.3 Weighted least squares and robust standard errors 155
5.2.4 Leading examples 155
5.3 Modeling heteroskedastic data 156
5.3.1 Simulated dataset 156

5.3.2 OLS estimation . . . 157
5.3.3 Detecting heteroskedasticity . . . 158
5.3.4 FGLS estimation . . . 160
5.3.5 WLS estimation . . . 162
5.4 System of linear regressions . . . 162
5.4.1 SUR model . . . 162
5.4.2 The sureg command . . . 163
5.4.3 Application to two categories of expenditures . . . 164
5.4.4 Robust standard errors . . . 166
5.4.5 Testing cross-equation constraints . . . 167
5.4.6 Imposing cross-equation constraints . . . 168
5.5 Survey data: Weighting, clustering, and stratification . . . 169
5.5.1 Survey design . . . 170
5.5.2 Survey mean estimation . . . 173
5.5.3 Survey linear regression . . . 173
5.6 Stata resources . . . 175
5.7 Exercises . . . 175
6 Linear instrumental-variables regression 177
6.1 Introduction . . . 177
6.2 IV estimation . . . 177
6.2.1 Basic IV theory . . . 177
6.2.2 Model setup . . . 179
6.2.3 IV estimators: IV, 2SLS, and GMM . . . 180
6.2.4 Instrument validity and relevance . . . 181
6.2.5 Robust standard-error estimates . . . 182
6.3 IV example . . . 183
6.3.1 The ivregress command . . . 183
6.3.2 Medical expenditures with one endogenous regressor . . . 184
6.3.3 Available instruments . . . 185
6.3.4 IV estimation of an exactly identified model . . . 186

6.3.5 IV estimation of an overidentified model 187
6.3.6 Testing for regressor endogeneity 188
6.3.7 Tests of overidentifying restrictions 191
6.3.8 IV estimation with a binary endogenous regressor 192
6.4 Weak instruments 194
6.4.1 Finite-sample properties of IV estimators 194
6.4.2 Weak instruments 195
Diagnostics for weak instruments 195
Formal tests for weak instruments 196
6.4.3 The estat firststage command 197
6.4.4 Just-identified model 197
6.4.5 Overidentified model 199
6.4.6 More than one endogenous regressor 200
6.4.7 Sensitivity to choice of instruments 200
6.5 Better inference with weak instruments 202
6.5.1 Conditional tests and confidence intervals 202
6.5.2 LIML estimator 204
6.5.3 Jackknife IV estimator 204
6.5.4 Comparison of 2SLS, LIML, JIVE, and GMM 205
6.6 3SLS systems estimation 206
6.7 Stata resources 208
6.8 Exercises 208
7 Quantile regression **211**
7.1 Introduction 211
7.2 QR 211
7.2.1 Conditional quantiles 212
7.2.2 Computation of QR estimates and standard errors 213
7.2.3 The qreg, bsqreg, and sqreg commands 213
7.3 QR for medical expenditures data 214
7.3.1 Data summary 214

7.3.2 QR estimates . . . 215
7.3.3 Interpretation of conditional quantile coefficients . . . 216
7.3.4 Retransformation . . . 217
7.3.5 Comparison of estimates at different quantiles . . . 218
7.3.6 Heteroskedasticity test . . . 219
7.3.7 Hypothesis tests . . . 220
7.3.8 Graphical display of coefficients over quantiles . . . 221
7.4 QR for generated heteroskedastic data . . . 222
7.4.1 Simulated dataset . . . 222
7.4.2 QR estimates . . . 225
7.5 QR for count data . . . 226
7.5.1 Quantile count regression . . . 227
7.5.2 The qcount command . . . 228
7.5.3 Summary of doctor visits data . . . 228
7.5.4 Results from QCR . . . 230
7.6 Stata resources . . . 232
7.7 Exercises . . . 232
8 Linear panel-data models: Basics 235
8.1 Introduction . . . 235
8.2 Panel-data methods overview . . . 235
8.2.1 Some basic considerations . . . 236
8.2.2 Some basic panel models . . . 237
Individual-effects model . . . 237
Fixed-effects model . . . 237
Random-effects model . . . 238
Pooled model or population-averaged model . . . 238
Two-way–effects model . . . 238
Mixed linear models . . . 239
8.2.3 Cluster–robust inference . . . 239
8.2.4 The xtreg command . . . 239

8.2.5 Stata linear panel-data commands 240
8.3 Panel-data summary 240
8.3.1 Data description and summary statistics 240
8.3.2 Panel-data organization 242
8.3.3 Panel-data description 243
8.3.4 Within and between variation 244
8.3.5 Time-series plots for each individual 247
8.3.6 Overall scatterplot 248
8.3.7 Within scatterplot 249
8.3.8 Pooled OLS regression with cluster–robust standard errors . 250
8.3.9 Time-series autocorrelations for panel data 251
8.3.10 Error correlation in the RE model 253
8.4 Pooled or population-averaged estimators 254
8.4.1 Pooled OLS estimator 254
8.4.2 Pooled FGLS estimator or population-averaged estimator . 254
8.4.3 The xtreg, pa command 255
8.4.4 Application of the xtreg, pa command 256
8.5 Within estimator 257
8.5.1 Within estimator 257
8.5.2 The xtreg, fe command 257
8.5.3 Application of the xtreg, fe command 258
8.5.4 Least-squares dummy-variables regression 259
8.6 Between estimator 260
8.6.1 Between estimator 260
8.6.2 Application of the xtreg, be command 261
8.7 RE estimator 261
8.7.1 RE estimator 262
8.7.2 The xtreg, re command 262
8.7.3 Application of the xtreg, re command 263

8.8 Comparison of estimators . . . 264
8.8.1 Estimates of variance components . . . 264
8.8.2 Within and between R-squared . . . 264
8.8.3 Estimator comparison . . . 265
8.8.4 Fixed effects versus random effects . . . 266
8.8.5 Hausman test for fixed effects . . . 266
The hausman command . . . 267
Robust Hausman test . . . 267
8.8.6 Prediction . . . 268
8.9 First-difference estimator . . . 269
8.9.1 First-difference estimator . . . 270
8.9.2 Strict and weak exogeneity . . . 271
8.10 Long panels . . . 271
8.10.1 Long-panel dataset . . . 271
8.10.2 Pooled OLS and PFGLS . . . 273
8.10.3 The xtpcse and xtgls commands . . . 273
8.10.4 Application of the xtgls, xtpcse, and xtscc commands . . . 274
8.10.5 Separate regressions . . . 276
8.10.6 FE and RE models . . . 277
8.10.7 Unit roots and cointegration . . . 278
8.11 Panel-data management . . . 280
8.11.1 Wide-form data . . . 280
8.11.2 Convert wide form to long form . . . 280
8.11.3 Convert long form to wide form . . . 281
8.11.4 An alternative wide-form data . . . 282
8.12 Stata resources . . . 284
8.13 Exercises . . . 284
9 Linear panel-data models: Extensions 287
9.1 Introduction . . . 287
9.2 Panel IV estimation . . . 287

9.2.1 Panel IV . . . 287
9.2.2 The xtivreg command . . . 288
9.2.3 Application of the xtivreg command . . . 288
9.2.4 Panel IV extensions . . . 290
9.3 Hausman–Taylor estimator . . . 290
9.3.1 Hausman–Taylor estimator . . . 290
9.3.2 The xthtaylor command . . . 291
9.3.3 Application of the xthtaylor command . . . 291
9.4 Arellano–Bond estimator . . . 293
9.4.1 Dynamic model . . . 293
9.4.2 IV estimation in the FD model . . . 294
9.4.3 The xtabond command . . . 295
9.4.4 Arellano–Bond estimator: Pure time series . . . 296
9.4.5 Arellano–Bond estimator: Additional regressors . . . 298
9.4.6 Specification tests . . . 300
9.4.7 The xtdpdsys command . . . 301
9.4.8 The xtdpd command . . . 303
9.5 Mixed linear models . . . 305
9.5.1 Mixed linear model . . . 305
9.5.2 The xtmixed command . . . 306
9.5.3 Random-intercept model . . . 306
9.5.4 Cluster–robust standard errors . . . 307
9.5.5 Random-slopes model . . . 308
9.5.6 Random-coefficients model . . . 310
9.5.7 Two-way random-effects model . . . 311
9.6 Clustered data . . . 312
9.6.1 Clustered dataset . . . 312
9.6.2 Clustered data using nonpanel commands . . . 313
9.6.3 Clustered data using panel commands . . . 314
9.6.4 Hierarchical linear models . . . 316

9.7 Stata resources . . . 317
9.8 Exercises . . . 318
10 Nonlinear regression methods 319
10.1 Introduction . . . 319
10.2 Nonlinear example: Doctor visits . . . 320
10.2.1 Data description . . . 320
10.2.2 Poisson model description . . . 321
10.3 Nonlinear regression methods . . . 322
10.3.1 MLE . . . 322
10.3.2 The poisson command . . . 323
10.3.3 Postestimation commands . . . 324
10.3.4 NLS . . . 325
10.3.5 The nl command . . . 325
10.3.6 GLM . . . 327
10.3.7 The glm command . . . 327
10.3.8 The gmm command . . . 328
10.3.9 Other estimators . . . 330
10.4 Different estimates of the VCE . . . 330
10.4.1 General framework . . . 330
10.4.2 The vce() option . . . 331
10.4.3 Application of the vce() option . . . 332
10.4.4 Default estimate of the VCE . . . 333
10.4.5 Robust estimate of the VCE . . . 334
10.4.6 Cluster–robust estimate of the VCE . . . 335
10.4.7 Heteroskedasticity- and autocorrelation-consistent estimate of the VCE . . . 335
10.4.8 Bootstrap standard errors . . . 336
10.4.9 Statistical inference . . . 336
10.5 Prediction . . . 336
10.5.1 The predict and predictnl commands . . . 337

10.5.2 Application of predict and predictnl . . . 337
10.5.3 Out-of-sample prediction . . . 338
10.5.4 Prediction at a specified value of one of the regressors . . . 339
10.5.5 Prediction at a specified value of all the regressors . . . 340
10.5.6 Prediction of other quantities . . . 341
10.5.7 The margins command for prediction . . . 341
10.6 Marginal effects . . . 343
10.6.1 Calculus and finite-difference methods . . . 343
10.6.2 MEs estimates AME, MEM, and MER . . . 344
10.6.3 Elasticities and semielasticities . . . 344
10.6.4 Simple interpretations of coefficients in single-index models 345
10.6.5 The margins command for marginal effects . . . 346
10.6.6 MEM: Marginal effect at mean . . . 347
Comparison of calculus and finite-difference methods . . . 348
10.6.7 MER: Marginal effect at representative value . . . 348
10.6.8 AME: Average marginal effect . . . 349
10.6.9 Elasticities and semielasticities . . . 351
10.6.10 AME computed manually . . . 352
10.6.11 Polynomial regressors . . . 354
10.6.12 Interacted regressors . . . 355
10.6.13 Complex interactions and nonlinearities . . . 356
10.7 Model diagnostics . . . 357
10.7.1 Goodness-of-fit measures . . . 357
10.7.2 Information criteria for model comparison . . . 359
10.7.3 Residuals . . . 359
10.7.4 Model-specification tests . . . 361
10.8 Stata resources . . . 361
10.9 Exercises . . . 361
11 Nonlinear optimization methods 363
11.1 Introduction . . . 363

11.2 Newton–Raphson method 363
11.2.1 NR method 363
11.2.2 NR method for Poisson 364
11.2.3 Poisson NR example using Mata 365
Core Mata code for Poisson NR iterations 365
Complete Stata and Mata code for Poisson NR iterations . 365
11.3 Gradient methods 367
11.3.1 Maximization options 367
11.3.2 Gradient methods 368
11.3.3 Messages during iterations 369
11.3.4 Stopping criteria 369
11.3.5 Multiple maximums 369
11.3.6 Numerical derivatives 370
11.4 The ml command: lf method 371
11.4.1 The ml command 372
11.4.2 The lf method 372
11.4.3 Poisson example: Single-index model 373
11.4.4 Negative binomial example: Two-index model 375
11.4.5 NLS example: Nonlikelihood model 376
11.5 Checking the program 376
11.5.1 Program debugging using ml check and ml trace 377
11.5.2 Getting the program to run 378
11.5.3 Checking the data 379
11.5.4 Multicollinearity and near collinearity 379
11.5.5 Multiple optimums 380
11.5.6 Checking parameter estimation 381
11.5.7 Checking standard-error estimation 382
11.6 The ml command: d0, d1, d2, lf0, lf1, and lf2 methods 383
11.6.1 Evaluator functions 383
11.6.2 The d0 method 385

11.6.3 The d1 method . . . 386
11.6.4 The lf1 method with the robust estimate of the VCE . . . 387
11.6.5 The d2 and lf2 methods . . . 388
11.7 The Mata optimize() function . . . 389
11.7.1 Type d and gf evaluators . . . 389
11.7.2 Optimize functions . . . 390
11.7.3 Poisson example . . . 390
Evaluator program for Poisson MLE . . . 390
The optimize() function for Poisson MLE . . . 391
11.8 Generalized method of moments . . . 392
11.8.1 Definition . . . 393
11.8.2 Nonlinear IV example . . . 393
11.8.3 GMM using the Mata optimize() function . . . 394
11.9 Stata resources . . . 396
11.10 Exercises . . . 396
12 Testing methods **399**
12.1 Introduction . . . 399
12.2 Critical values and p-values . . . 399
12.2.1 Standard normal compared with Student's t . . . 400
12.2.2 Chi-squared compared with F . . . 400
12.2.3 Plotting densities . . . 400
12.2.4 Computing p-values and critical values . . . 402
12.2.5 Which distributions does Stata use? . . . 403
12.3 Wald tests and confidence intervals . . . 403
12.3.1 Wald test of linear hypotheses . . . 403
12.3.2 The test command . . . 405
Test single coefficient . . . 406
Test several hypotheses . . . 406
Test of overall significance . . . 407
Test calculated from retrieved coefficients and VCE . . . 407

12.3.3 One-sided Wald tests . . . 408
12.3.4 Wald test of nonlinear hypotheses (delta method) . . . 409
12.3.5 The testnl command . . . 409
12.3.6 Wald confidence intervals . . . 410
12.3.7 The lincom command . . . 410
12.3.8 The nlcom command (delta method) . . . 411
12.3.9 Asymmetric confidence intervals . . . 412
12.4 Likelihood-ratio tests . . . 413
12.4.1 Likelihood-ratio tests . . . 413
12.4.2 The lrtest command . . . 415
12.4.3 Direct computation of LR tests . . . 415
12.5 Lagrange multiplier test (or score test) . . . 416
12.5.1 LM tests . . . 416
12.5.2 The estat command . . . 417
12.5.3 LM test by auxiliary regression . . . 417
12.6 Test size and power . . . 419
12.6.1 Simulation DGP: OLS with chi-squared errors . . . 419
12.6.2 Test size . . . 420
12.6.3 Test power . . . 421
12.6.4 Asymptotic test power . . . 424
12.7 Specification tests . . . 425
12.7.1 Moment-based tests . . . 425
12.7.2 Information matrix test . . . 425
12.7.3 Chi-squared goodness-of-fit test . . . 426
12.7.4 Overidentifying restrictions test . . . 426
12.7.5 Hausman test . . . 426
12.7.6 Other tests . . . 427
12.8 Stata resources . . . 427
12.9 Exercises . . . 427

13 Bootstrap methods 429

13.1 Introduction . . . 429

13.2 Bootstrap methods . . . 429

13.2.1 Bootstrap estimate of standard error . . . 429

13.2.2 Bootstrap methods . . . 430

13.2.3 Asymptotic refinement . . . 430

13.2.4 Use the bootstrap with caution . . . 430

13.3 Bootstrap pairs using the vce(bootstrap) option . . . 431

13.3.1 Bootstrap-pairs method to estimate VCE . . . 431

13.3.2 The vce(bootstrap) option . . . 432

13.3.3 Bootstrap standard-errors example . . . 432

13.3.4 How many bootstraps? . . . 433

13.3.5 Clustered bootstraps . . . 434

13.3.6 Bootstrap confidence intervals . . . 435

13.3.7 The postestimation estat bootstrap command . . . 436

13.3.8 Bootstrap confidence-intervals example . . . 437

13.3.9 Bootstrap estimate of bias . . . 437

13.4 Bootstrap pairs using the bootstrap command . . . 438

13.4.1 The bootstrap command . . . 438

13.4.2 Bootstrap parameter estimate from a Stata estimation command . . . 439

13.4.3 Bootstrap standard error from a Stata estimation command 440

13.4.4 Bootstrap standard error from a user-written estimation command . . . 440

13.4.5 Bootstrap two-step estimator . . . 441

13.4.6 Bootstrap Hausman test . . . 443

13.4.7 Bootstrap standard error of the coefficient of variation . . . 444

13.5 Bootstraps with asymptotic refinement . . . 445

13.5.1 Percentile-t method . . . 445

13.5.2 Percentile-t Wald test . . . 446

13.5.3 Percentile-t Wald confidence interval . . . 447

13.6 Bootstrap pairs using bsample and simulate . . . 448
13.6.1 The bsample command . . . 448
13.6.2 The bsample command with simulate . . . 448
13.6.3 Bootstrap Monte Carlo exercise . . . 450
13.7 Alternative resampling schemes . . . 450
13.7.1 Bootstrap pairs . . . 451
13.7.2 Parametric bootstrap . . . 451
13.7.3 Residual bootstrap . . . 453
13.7.4 Wild bootstrap . . . 454
13.7.5 Subsampling . . . 455
13.8 The jackknife . . . 455
13.8.1 Jackknife method . . . 455
13.8.2 The vce(jackknife) option and the jackknife command . . . 456
13.9 Stata resources . . . 456
13.10 Exercises . . . 456
14 Binary outcome models 459
14.1 Introduction . . . 459
14.2 Some parametric models . . . 459
14.2.1 Basic model . . . 459
14.2.2 Logit, probit, linear probability, and clog-log models . . . 460
14.3 Estimation . . . 460
14.3.1 Latent-variable interpretation and identification . . . 461
14.3.2 ML estimation . . . 461
14.3.3 The logit and probit commands . . . 462
14.3.4 Robust estimate of the VCE . . . 462
14.3.5 OLS estimation of LPM . . . 462
14.4 Example . . . 463
14.4.1 Data description . . . 463
14.4.2 Logit regression . . . 464
14.4.3 Comparison of binary models and parameter estimates . . . 465

14.5 Hypothesis and specification tests 466
14.5.1 Wald tests 467
14.5.2 Likelihood-ratio tests 467
14.5.3 Additional model-specification tests 468
Lagrange multiplier test of generalized logit 468
Heteroskedastic probit regression 469
14.5.4 Model comparison 470
14.6 Goodness of fit and prediction 471
14.6.1 Pseudo-R^2 measure 471
14.6.2 Comparing predicted probabilities with sample frequencies . 471
14.6.3 Comparing predicted outcomes with actual outcomes 473
14.6.4 The predict command for fitted probabilities 474
14.6.5 The prvalue command for fitted probabilities 475
14.7 Marginal effects 476
14.7.1 Marginal effect at a representative value (MER) 476
14.7.2 Marginal effect at the mean (MEM) 477
14.7.3 Average marginal effect (AME) 478
14.7.4 The prchange command 479
14.8 Endogenous regressors 479
14.8.1 Example 480
14.8.2 Model assumptions 481
14.8.3 Structural-model approach 481
The ivprobit command 482
Maximum likelihood estimates 482
Two-step sequential estimates 483
14.8.4 IVs approach 485
14.9 Grouped data 486
14.9.1 Estimation with aggregate data 487
14.9.2 Grouped-data application 487
14.10 Stata resources 489

14.11 Exercises 489

15 Multinomial models 491

15.1 Introduction 491

15.2 Multinomial models overview 491

15.2.1 Probabilities and MEs 491

15.2.2 Maximum likelihood estimation 492

15.2.3 Case-specific and alternative-specific regressors 493

15.2.4 Additive random-utility model 493

15.2.5 Stata multinomial model commands 494

15.3 Multinomial example: Choice of fishing mode 494

15.3.1 Data description 494

15.3.2 Case-specific regressors 497

15.3.3 Alternative-specific regressors 497

15.4 Multinomial logit model 498

15.4.1 The mlogit command 498

15.4.2 Application of the mlogit command 499

15.4.3 Coefficient interpretation 500

15.4.4 Predicted probabilities 501

15.4.5 MEs 502

15.5 Conditional logit model 503

15.5.1 Creating long-form data from wide-form data 503

15.5.2 The asclogit command 505

15.5.3 The clogit command 506

15.5.4 Application of the asclogit command 506

15.5.5 Relationship to multinomial logit model 507

15.5.6 Coefficient interpretation 507

15.5.7 Predicted probabilities 508

15.5.8 MEs 509

15.6 Nested logit model 511
15.6.1 Relaxing the independence of irrelevant alternatives assumption 511
15.6.2 NL model 511
15.6.3 The nlogit command 512
15.6.4 Model estimates 513
15.6.5 Predicted probabilities 516
15.6.6 MEs 516
15.6.7 Comparison of logit models 517
15.7 Multinomial probit model 517
15.7.1 MNP 517
15.7.2 The mprobit command 518
15.7.3 Maximum simulated likelihood 519
15.7.4 The asmprobit command 519
15.7.5 Application of the asmprobit command 520
15.7.6 Predicted probabilities and MEs 522
15.8 Random-parameters logit 522
15.8.1 Random-parameters logit 523
15.8.2 The mixlogit command 523
15.8.3 Data preparation for mixlogit 524
15.8.4 Application of the mixlogit command 524
15.9 Ordered outcome models 525
15.9.1 Data summary 525
15.9.2 Ordered outcomes 526
15.9.3 Application of the ologit command 527
15.9.4 Predicted probabilities 528
15.9.5 MEs 528
15.9.6 Other ordered models 529
15.10 Multivariate outcomes 529
15.10.1 Bivariate probit 529

15.10.2 Nonlinear SUR 532
15.11 Stata resources 532
15.12 Exercises 533
16 Tobit and selection models **535**
16.1 Introduction 535
16.2 Tobit model 535
16.2.1 Regression with censored data 535
16.2.2 Tobit model setup 536
16.2.3 Unknown censoring point 537
16.2.4 Tobit estimation 537
16.2.5 ML estimation in Stata 538
16.3 Tobit model example 538
16.3.1 Data summary 538
16.3.2 Tobit analysis 539
16.3.3 Prediction after tobit 540
16.3.4 Marginal effects 541
Left-truncated, left-censored, and right-truncated examples 541
Left-censored case computed directly 542
Marginal impact on probabilities 543
16.3.5 The ivtobit command 544
16.3.6 Additional commands for censored regression 545
16.4 Tobit for lognormal data 545
16.4.1 Data example 546
16.4.2 Setting the censoring point for data in logs 546
16.4.3 Results 547
16.4.4 Two-limit tobit 548
16.4.5 Model diagnostics 549
16.4.6 Tests of normality and homoskedasticity 550
Generalized residuals and scores 550
Test of normality 551

Test of homoskedasticity 552
16.4.7 Next step? 553
16.5 Two-part model in logs 553
16.5.1 Model structure 553
16.5.2 Part 1 specification 554
16.5.3 Part 2 of the two-part model 555
16.6 Selection model 556
16.6.1 Model structure and assumptions 556
16.6.2 ML estimation of the sample-selection model 558
16.6.3 Estimation without exclusion restrictions 558
16.6.4 Two-step estimation 560
16.6.5 Estimation with exclusion restrictions 561
16.7 Prediction from models with outcome in logs 562
16.7.1 Predictions from tobit 563
16.7.2 Predictions from two-part model 564
16.7.3 Predictions from selection model 565
16.8 Stata resources 565
16.9 Exercises 566
17 Count-data models **567**
17.1 Introduction 567
17.2 Features of count data 567
17.2.1 Generated Poisson data 568
17.2.2 Overdispersion and negative binomial data 569
17.2.3 Modeling strategies 570
17.2.4 Estimation methods 571
17.3 Empirical example 1 571
17.3.1 Data summary 571
17.3.2 Poisson model 572
Poisson model results 573
Robust estimate of VCE for Poisson MLE 574

Test of overdispersion 575
Coefficient interpretation and marginal effects 576
17.3.3 NB2 model 577
NB2 model results 577
Fitted probabilities for Poisson and NB2 models 579
The countfit command 579
The prvalue command 581
Discussion 581
Generalized NB model 581
17.3.4 Nonlinear least-squares estimation 582
17.3.5 Hurdle model 583
Variants of the hurdle model 585
Application of the hurdle model 585
17.3.6 Finite-mixture models 589
FMM specification 589
Simulated FMM sample with comparisons 589
ML estimation of the FMM 591
The fmm command 592
Application: Poisson finite-mixture model 592
Interpretation 593
Comparing marginal effects 594
Application: NB finite-mixture model 596
Model selection 598
Cautionary note 599
17.4 Empirical example 2 599
17.4.1 Zero-inflated data 599
17.4.2 Models for zero-inflated data 600
17.4.3 Results for the NB2 model 601
The prcounts command 602
17.4.4 Results for ZINB 603

17.4.5 Model comparison 604
The countfit command 604
Model comparison using countfit 604
17.5 Models with endogenous regressors 605
17.5.1 Structural-model approach 606
Model and assumptions 606
Two-step estimation 607
Application 607
17.5.2 Nonlinear IV method 610
17.6 Stata resources 611
17.7 Exercises 612
18 Nonlinear panel models **615**
18.1 Introduction 615
18.2 Nonlinear panel-data overview 615
18.2.1 Some basic nonlinear panel models 615
FE models 616
RE models 616
Pooled models or population-averaged models 616
Comparison of models 617
18.2.2 Dynamic models 617
18.2.3 Stata nonlinear panel commands 617
18.3 Nonlinear panel-data example 618
18.3.1 Data description and summary statistics 618
18.3.2 Panel-data organization 620
18.3.3 Within and between variation 620
18.3.4 FE or RE model for these data? 621
18.4 Binary outcome models 621
18.4.1 Panel summary of the dependent variable 621
18.4.2 Pooled logit estimator 622
18.4.3 The xtlogit command 623

18.4.4 The xtgee command . . . 624
18.4.5 PA logit estimator . . . 624
18.4.6 RE logit estimator . . . 625
18.4.7 FE logit estimator . . . 627
18.4.8 Panel logit estimator comparison . . . 629
18.4.9 Prediction and marginal effects . . . 630
18.4.10 Mixed-effects logit estimator . . . 630
18.5 Tobit model . . . 631
18.5.1 Panel summary of the dependent variable . . . 631
18.5.2 RE tobit model . . . 631
18.5.3 Generalized tobit models . . . 632
18.5.4 Parametric nonlinear panel models . . . 633
18.6 Count-data models . . . 633
18.6.1 The xtpoisson command . . . 633
18.6.2 Panel summary of the dependent variable . . . 634
18.6.3 Pooled Poisson estimator . . . 634
18.6.4 PA Poisson estimator . . . 635
18.6.5 RE Poisson estimators . . . 636
18.6.6 FE Poisson estimator . . . 638
18.6.7 Panel Poisson estimators comparison . . . 640
18.6.8 Negative binomial estimators . . . 641
18.7 Stata resources . . . 642
18.8 Exercises . . . 643
A Programming in Stata 645
A.1 Stata matrix commands . . . 645
A.1.1 Stata matrix overview . . . 645
A.1.2 Stata matrix input and output . . . 645
Matrix input by hand . . . 645
Matrix input from Stata estimation results . . . 646
A.1.3 Stata matrix subscripts and combining matrices . . . 647

A.1.4 Matrix operators 648
A.1.5 Matrix functions 648
A.1.6 Matrix accumulation commands 649
A.1.7 OLS using Stata matrix commands 650
A.2 Programs 651
A.2.1 Simple programs (no arguments or access to results) 651
A.2.2 Modifying a program 652
A.2.3 Programs with positional arguments 652
A.2.4 Temporary variables 653
A.2.5 Programs with named positional arguments 653
A.2.6 Storing and retrieving program results 654
A.2.7 Programs with arguments using standard Stata syntax . . . 655
A.2.8 Ado-files 656
A.3 Program debugging 657
A.3.1 Some simple tips 658
A.3.2 Error messages and return code 658
A.3.3 Trace 659
B Mata **661**
B.1 How to run Mata 661
B.1.1 Mata commands in Mata 661
B.1.2 Mata commands in Stata 662
B.1.3 Stata commands in Mata 662
B.1.4 Interactive versus batch use 662
B.1.5 Mata help 662
B.2 Mata matrix commands 663
B.2.1 Mata matrix input 663
Matrix input by hand 663
Identity matrices, unit vectors, and matrices of constants . . 664
Matrix input from Stata data 665
Matrix input from Stata matrix 665

Stata interface functions 666
B.2.2 Mata matrix operators 666
Element-by-element operators 666
B.2.3 Mata functions 667
Scalar and matrix functions 667
Matrix inversion 668
B.2.4 Mata cross products 669
B.2.5 Mata matrix subscripts and combining matrices 669
B.2.6 Transferring Mata data and matrices to Stata 671
Creating Stata matrices from Mata matrices 671
Creating Stata data from a Mata vector 671
B.3 Programming in Mata 672
B.3.1 Declarations 672
B.3.2 Mata program 672
B.3.3 Mata program with results output to Stata 673
B.3.4 Stata program that calls a Mata program 673
B.3.5 Using Mata in ado-files 674
Glossary of abbreviations **675**
References **679**
Author index **687**
Subject index **691**

(Pages omitted)

Preface to the Revised Edition

Microeconometrics Using Stata, published in December 2008, was written for Stata 10.1. The book incorporated version 10.1 additions to Stata 10.0, most notably, the new random-number generators.

In this revised edition, we present other additions to Stata 10 that appear for the first time in Stata 11. With few exceptions, we present these additions in a way that reproduces the results given in the first edition.

First, we introduce the new construct of factor variables. These provide a simple way to specify models with sets of indicator variables formed from a categorical variable and to specify models with interactions. Factor variables replace the `xi` prefix command. See especially section 1.3.4 and the end of section 2.4.7.

Second, we describe the new `margins` command for prediction and for computation of marginal effects in regression models. The `margins` command with options including the `dydx()` option replaces the Stata `mfx` command and the user-written `margeff` command. Additionally, the `margins` command when used in conjunction with factor variables can simplify computation of marginal effects in models with interactions. See sections 10.5 and 10.6, especially subsections 10.5.7 and 10.6.5. Throughout this revised edition, notably, in chapters 14–17, we replace `mfx` and `margeff` with the `margins` command.

In the first edition, we most often calculated the marginal effect at the mean (MEM), rather than the average marginal effect (AME), because the `mfx` command did not compute the AME. The new `margins` command can compute both the MEM and the AME. In this revised edition, we have endeavored to replicate the results given in the first edition. For that reason, we continue to most frequently calculate the MEM, though in practice, the AME is usually preferred.

Third, we describe the new `gmm` command for generalized method of moments and nonlinear instrumental-variables estimation. See sections 10.3.8 and 17.5.2.

Fourth, we present some minor changes that need to be made to the existing `ml` command when the `d1` and `d2` methods are used. These changes arise because the `ml` command is now a front-end to the new Mata `moptimize()` function. We also present the new `lf0`, `lf1`, and `lf2` methods. See section 11.6. The Mata `optimize()` v evaluator has been renamed to gf evaluator; see section 11.7.

We thank the Stata staff, especially Patricia Branton, David Drukker, Lisa Gilmore, Deirdre Patterson, and Brian Poi, for their assistance in preparing this revised edition.

Davis, CA
Bloomington, IN
January 2010

A. Colin Cameron
Pravin K. Trivedi

Preface to the First Edition

This book explains how an econometrics computer package, Stata, can be used to perform regression analysis of cross-section and panel data. The term microeconometrics is used in the book title because the applications are to economics-related data and because the coverage includes methods such as instrumental-variables regression that are emphasized more in economics than in some other areas of applied statistics. However, many issues, models, and methodologies discussed in this book are also relevant to other social sciences.

The main audience is graduate students and researchers. For them, this book can be used as an adjunct to our own *Microeconometrics: Methods and Applications* (Cameron and Trivedi 2005), as well as to other graduate-level texts such as Greene (2008) and Wooldridge (2002). By comparison to these books, we present little theory and instead emphasize practical aspects of implementation using Stata. More advanced topics we cover include quantile regression, weak instruments, nonlinear optimization, bootstrap methods, nonlinear panel-data methods, and Stata's matrix programming language, Mata.

At the same time, the book provides introductions to topics such as ordinary least-squares regression, instrumental-variables estimation, and logit and probit models so that it is suitable for use in an undergraduate econometrics class, as a complement to an appropriate undergraduate-level text. The following table suggests sections of the book for an introductory class, with the caveat that in places formulas are provided using matrix algebra.

Stata basics	Chapter 1.1–1.4
Data management	Chapter 2.1–2.4, 2.6
OLS	Chapter 3.1–3.6
Simulation	Chapter 4.6–4.7
GLS (heteroskedasticity)	Chapter 5.3
Instrumental variables	Chapter 6.2–6.3
Linear panel data	Chapter 8
Logit and probit models	Chapter 14.1–14.4
Tobit model	Chapter 16.1–16.3

Although we provide considerable detail on Stata, the treatment is by no means complete. In particular, we introduce various Stata commands but avoid detailed listing and description of commands as they are already well documented in the Stata manuals

and online help. Typically, we provide a pointer and a brief discussion and often an example.

As much as possible, we provide template code that can be adapted to other problems. Keep in mind that to shorten output for this book, our examples use many fewer regressors than necessary for serious research. Our code often suppresses intermediate output that is important in actual research, because of extensive use of command `quietly` and options `nolog`, `nodots`, and `noheader`. And we minimize the use of graphs compared with typical use in exploratory data analysis.

We have used Stata 10, including Stata updates.[1] Instructions on how to obtain the datasets and the do-files used in this book are available on the Stata Press web site at http://www.stata-press.com/data/mus.html. Any corrections to the book will be documented at http://www.stata-press.com/books/mus.html.

We have learned a lot of econometrics, in addition to learning Stata, during this project. Indeed, we feel strongly that an effective learning tool for econometrics is hands-on learning by opening a Stata dataset and seeing the effect of using different methods and variations on the methods, such as using robust standard errors rather than default standard errors. This method is beneficial at all levels of ability in econometrics. Indeed, an efficient way of familiarizing yourself with Stata's leading features might be to execute the commands in a relevant chapter on your own dataset.

We thank the many people who have assisted us in preparing this book. The project grew out of our 2005 book, and we thank Scott Parris for his expert handling of that book. Juan Du, Qian Li, and Abhijit Ramalingam carefully read many of the book chapters. Discussions with John Daniels, Oscar Jorda, Guido Kuersteiner, and Doug Miller were particularly helpful. We thank Deirdre Patterson for her excellent editing and Lisa Gilmore for managing the LaTeX formatting and production of this book. Most especially, we thank David Drukker for his extensive input and encouragement at all stages of this project, including a thorough reading and critique of the final draft, which led to many improvements in both the econometrics and Stata components of this book. Finally, we thank our respective families for making the inevitable sacrifices as we worked to bring this multiyear project to completion.

Davis, CA
Bloomington, IN
October 2008

A. Colin Cameron
Pravin K. Trivedi

1. To see whether you have the latest update, type `update query`. For those with earlier versions of Stata, some key changes are the following: Stata 9 introduced the matrix programming language, Mata. The syntax for Stata 10 uses the `vce(robust)` option rather than the `robust` option to obtain robust standard errors. A mid-2008 update of version 10 introduced new random-number functions, such as `runiform()` and `rnormal()`.

(Pages omitted)

3 Linear regression basics

3.1 Introduction

Linear regression analysis is often the starting point of an empirical investigation. Because of its relative simplicity, it is useful for illustrating the different steps of a typical modeling cycle that involves an initial specification of the model followed by estimation, diagnostic checks, and model respecification. The purpose of such a linear regression analysis may be to summarize the data, generate conditional predictions, or test and evaluate the role of specific regressors. We will illustrate these aspects using a specific data example.

This chapter is limited to basic regression analysis on cross-section data of a continuous dependent variable. The setup is for a single equation and exogenous regressors. Some standard complications of linear regression, such as misspecification of the conditional mean and model errors that are heteroskedastic, will be considered. In particular, we model the natural logarithm of medical expenditures instead of the level. We will ignore other various aspects of the data that can lead to more sophisticated nonlinear models presented in later chapters.

3.2 Data and data summary

The first step is to decide what dataset will be used. In turn, this decision depends on the population of interest and the research question itself. We discussed how to convert a raw dataset to a form amenable to regression analysis in chapter 2. In this section, we present ways to summarize and gain some understanding of the data, a necessary step before any regression analysis.

3.2.1 Data description

We analyze medical expenditures of individuals 65 years and older who qualify for health care under the U.S. Medicare program. The original data source is the Medical Expenditure Panel Survey (MEPS).

Medicare does not cover all medical expenses. For example, copayments for medical services and expenses of prescribed pharmaceutical drugs were not covered for the time period studied here. About half of eligible individuals therefore purchase supplementary insurance in the private market that provides insurance coverage against various out-of-pocket expenses.

In this chapter, we consider the impact of this supplementary insurance on total annual medical expenditures of an individual, measured in dollars. A formal investigation must control for the influence of other factors that also determine individual medical expenditure, notably, sociodemographic factors such as age, gender, education and income, geographical location, and health-status measures such as self-assessed health and presence of chronic or limiting conditions. In this chapter, as in other chapters, we instead deliberately use a short list of regressors. This permits shorter output and simpler discussion of the results, an advantage because our intention is to simply explain the methods and tools available in Stata.

3.2.2 Variable description

Given the Stata dataset for analysis, we begin by using the `describe` command to list various features of the variables to be used in the linear regression. The command without a variable list describes all the variables in the dataset. Here we restrict attention to the variables used in this chapter.

```
. * Variable description for medical expenditure dataset
. use mus03data.dta
. describe totexp ltotexp posexp suppins phylim actlim totchr age female income
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
totexp          double %12.0g                 Total medical expenditure
ltotexp         float  %9.0g                  ln(totexp) if totexp > 0
posexp          float  %9.0g                  =1 if total expenditure > 0
suppins         float  %9.0g                  =1 if has supp priv insurance
phylim          double %12.0g                 =1 if has functional limitation
actlim          double %12.0g                 =1 if has activity limitation
totchr          double %12.0g                 # of chronic problems
age             double %12.0g                 Age
female          double %12.0g                 =1 if female
income          double %12.0g                 annual household income/1000
```

The variable types and format columns indicate that all the data are numeric. In this case, some variables are stored in single precision (`float`) and some in double precision (`double`). From the variable labels, we expect `totexp` to be nonnegative; `ltotexp` to be missing if `totexp` equals zero; `posexp`, `suppins`, `phylim`, `actlim`, and `female` to be 0 or 1; `totchr` to be a nonnegative integer; `age` to be positive; and `income` to be negative or positive. Note that the integer variables could have been stored much more compactly as `integer` or `byte`. The variable labels provide a short description that is helpful but may not fully describe the variable. For example, the key regressor `suppins` was created by aggregating across several types of private supplementary insurance. No labels for the values taken by the categorical variables have been provided.

3.2.3 Summary statistics

It is essential in any data analysis to first check the data by using the `summarize` command.

```
. * Summary statistics for medical expenditure dataset
. summarize totexp ltotexp posexp suppins phylim actlim totchr age female income
```

Variable	Obs	Mean	Std. Dev.	Min	Max
totexp	3064	7030.889	11852.75	0	125610
ltotexp	2955	8.059866	1.367592	1.098612	11.74094
posexp	3064	.9644256	.1852568	0	1
suppins	3064	.5812663	.4934321	0	1
phylim	3064	.4255875	.4945125	0	1
actlim	3064	.2836162	.4508263	0	1
totchr	3064	1.754243	1.307197	0	7
age	3064	74.17167	6.372938	65	90
female	3064	.5796345	.4936982	0	1
income	3064	22.47472	22.53491	-1	312.46

On average, 96% of individuals incur medical expenditures during a year; 58% have supplementary insurance; 43% have functional limitations; 28% have activity limitations; and 58% are female, as the elderly population is disproportionately female because of the greater longevity of women. The only variable to have missing data is `ltotexp`, the natural logarithm of `totexp`, which is missing for the $(3064 - 2955) = 109$ observations with `totexp` $= 0$.

All variables have the expected range, except that income is negative. To see how many observations on `income` are negative, we use the `tabulate` command, restricting attention to nonpositive observations to limit output.

```
. * Tabulate variable
. tabulate income if income <= 0
```

annual household income/1000	Freq.	Percent	Cum.
-1	1	1.14	1.14
0	87	98.86	100.00
Total	88	100.00	

Only one observation is negative, and negative income is possible for income from self-employment or investment. We include the observation in the analysis here, though checking the original data source may be warranted.

Much of the subsequent regression analysis will drop the 109 observations with zero medical expenditures, so in a research paper, it would be best to report summary statistics without these observations.

3.2.4 More-detailed summary statistics

Additional descriptive analysis of key variables, especially the dependent variable, is useful. For `totexp`, the level of medical expenditures, `summarize, detail` yields

```
. * Detailed summary statistics of a single variable
. summarize totexp, detail
                    Total medical expenditure
-------------------------------------------------------------
      Percentiles      Smallest
 1%           0              0
 5%         112              0
10%         393              0       Obs                3064
25%        1271              0       Sum of Wgt.        3064
50%      3134.5                      Mean           7030.889
                        Largest      Std. Dev.      11852.75
75%        7151         104823
90%       17050         108256       Variance       1.40e+08
95%       27367         123611       Skewness       4.165058
99%       62346         125610       Kurtosis       26.26796
```

Medical expenditures vary greatly across individuals, with a standard deviation of 11,853, which is almost twice the mean. The median of 3,134 is much smaller than the mean of 7,031, reflecting the skewness of the data. For variable x, the skewness statistic is a scale-free measure of skewness that estimates $E\{(x-\mu)^3\}/\sigma^{3/2}$, the third central moment standardized by the second central moment. The skewness is zero for symmetrically distributed data. The value here of 4.16 indicates considerable right skewness. The kurtosis statistic is an estimate of $E\{(x-\mu)^4\}/\sigma^4$, the fourth central moment standardized by the second central moment. The reference value is 3, the value for normally distributed data. The much higher value here of 26.26 indicates that the tails are much thicker than those of a normal distribution. You can obtain additional summary statistics by using the `centile` command to obtain other percentiles and by using the `table` command, which is explained in section 3.2.5.

We conclude that the distribution of the dependent variable is considerably skewed and has thick tails. These complications often arise for commonly studied individual-level economic variables such as expenditures, income, earnings, wages, and house prices. It is possible that including regressors will eliminate the skewness, but in practice, much of the variation in the data will be left unexplained ($R^2 < 0.3$ is common for individual-level data) and skewness and excess kurtosis will remain.

Such skewed, thick-tailed data suggest a model with multiplicative errors instead of additive errors. A standard solution is to transform the dependent variable by taking the natural logarithm. Here this is complicated by the presence of 109 zero-valued observations. We take the expedient approach of dropping the zero observations from analysis in either logs or levels. This should make little difference here because only 3.6% of the sample is then dropped. A better approach, using two-part or selection models, is covered in chapter 16.

The output for `tabstat` in section 3.2.5 reveals that taking the natural logarithm for these data essentially eliminates the skewness and excess kurtosis.

The user-written `fsum` command (Wolfe 2002) is an enhancement of `summarize` that enables formatting the output and including additional information such as percentiles and variable labels. The user-written `outsum` command (Papps 2006) produces a text file of means and standard deviations for one or more subsets of the data, e.g., one column for the full sample, one for a male subsample, and one for a female subsample.

3.2.5 Tables for data

One-way tables can be created by using the `table` command, which produces just frequencies, or the `tabulate` command, which additionally produces percentages and cumulative percentages; an example was given in section 3.2.3.

Two-way tables can also be created by using these commands. For frequencies, only `table` produces clean output. For example,

```
. * Two-way table of frequencies
. table female totchr
```

=1 if female	# of chronic problems							
	0	1	2	3	4	5	6	7
0	239	415	323	201	82	23	4	1
1	313	466	493	305	140	46	11	2

provides frequencies for a two-way tabulation of gender against the number of chronic conditions. The `tabulate` command is much richer. For example,

```
. * Two-way table with row and column percentages and Pearson chi-squared
. tabulate female suppins, row col chi2
```

Key
frequency *row percentage* *column percentage*

=1 if female	=1 if has supp priv insurance		Total
	0	1	
0	488	800	1,288
	37.89	62.11	100.00
	38.04	44.92	42.04
1	795	981	1,776
	44.76	55.24	100.00
	61.96	55.08	57.96
Total	1,283	1,781	3,064
	41.87	58.13	100.00
	100.00	100.00	100.00

Pearson chi2(1) = 14.4991 Pr = 0.000

Comparing the row percentages for this sample, we see that while a woman is more likely to have supplemental insurance than not, the probability that a woman in this sample has purchased supplemental insurance is lower than the probability that a man in this sample has purchased supplemental insurance. Although we do not have the information to draw these inferences for the population, the results for Pearson's chi-squared test soundly reject the null hypothesis that these variables are independent. Other tests of association are available. The related command `tab2` will produce all possible two-way tables that can be obtained from a list of several variables.

For multiway tables, it is best to use `table`. For the example at hand, we have

```
. * Three-way table of frequencies
. table female totchr suppins
```

	=1 if has supp priv insurance and # of chronic problems							
=1 if	0							
female	0	1	2	3	4	5	6	7
0	102	165	121	68	25	6	1	
1	135	212	233	134	56	22	1	2

	=1 if has supp priv insurance and # of chronic problems							
=1 if	1							
female	0	1	2	3	4	5	6	7
0	137	250	202	133	57	17	3	1
1	178	254	260	171	84	24	10	

An alternative is to use `tabulate` with the `by` prefix, but the results are not as neat as those from `table`.

The preceding tabulations will produce voluminous output if one of the variables being tabulated takes on many values. Then it is much better to use `table` with the `contents()` option to present tables that give key summary statistics for that variable, such as the mean and standard deviation. Such tabulations can be useful even when variables take on few values. For example, when summarizing the number of chronic problems by gender, `table` yields

```
. * One-way table of summary statistics
. table female, contents(N totchr mean totchr sd totchr p50 totchr)
```

=1 if female	N(totchr)	mean(totchr)	sd(totchr)	med(totchr)
0	1,288	1.659937888	1.261175	1
1	1,776	1.822635135	1.335776	2

Women on average have more chronic problems (1.82 versus 1.66 for men). The option `contents()` can produce many other statistics, including the minimum, maximum, and key percentiles.

The `table` command with the `contents()` option can additionally produce two-way and multiway tables of summary statistics. As an example,

```
. * Two-way table of summary statistics
. table female suppins, contents(N totchr mean totchr)
```

=1 if female	=1 if has supp priv insurance	
	0	1
0	488	800
	1.530737705	1.73875
1	795	981
	1.803773585	1.837920489

shows that those with supplementary insurance on average have more chronic problems. This is especially so for males (1.74 versus 1.53).

The `tabulate, summarize()` command can be used to produce one-way and two-way tables with means, standard deviations, and frequencies. This is a small subset of the statistics that can be produced using `table`, so we might as well use `table`.

The `tabstat` command provides a table of summary statistics that permits more flexibility than `summarize`. The following output presents summary statistics on medical expenditures and the natural logarithm of expenditures that are useful in determining skewness and kurtosis.

```
. * Summary statistics obtained using command tabstat
. tabstat totexp ltotexp, stat (count mean p50 sd skew kurt) col(stat)
```

variable	N	mean	p50	sd	skewness	kurtosis
totexp	3064	7030.889	3134.5	11852.75	4.165058	26.26796
ltotexp	2955	8.059866	8.111928	1.367592	-.3857887	3.842263

This reproduces information given in section 3.2.4 and shows that taking the natural logarithm eliminates most skewness and kurtosis. The `col(stat)` option presents the results with summary statistics given in the columns and each variable being given in a separate row. Without this option, we would have summary statistics in rows and variables in the columns. A two-way table of summary statistics can be obtained by using the `by()` option.

(*Continued on next page*)

3.2.6 Statistical tests

The `ttest` command can be used to test hypotheses about the population mean of a single variable (H_0: $\mu = \mu^*$ for specified value μ^*) and to test the equality of means ($H_0 : \mu_1 = \mu_2$). For more general analysis of variance and analysis of covariance, the `oneway` and `anova` commands can be used, and several other tests exist for more specialized examples such as testing the equality of proportions. These commands are rarely used in microeconometrics because they can be recast as a special case of regression with an intercept and appropriate indicator variables. Furthermore, regression has the advantage of reliance on less restrictive distributional assumptions, provided samples are large enough for asymptotic theory to provide a good approximation.

For example, consider testing the equality of mean medical expenditures for those with and without supplementary health insurance. The `ttest totexp, by(suppins) unequal` command performs the test but makes the restrictive assumption of a common variance for all those with `suppins=0` and a (possibly different) common variance for all those with `suppins=1`. An alternative method is to perform ordinary least-squares (OLS) regression of `totexp` on an intercept and `suppins` and then test whether `suppins` has coefficient zero. Using this latter method, we can permit all observations to have a different variance by using the `vce(robust)` option for `regress` to obtain heteroskedastic-consistent standard errors; see section 3.3.4.

3.2.7 Data plots

It is useful to plot a histogram or a density estimate of the dependent variable. Here we use the `kdensity` command, which provides a kernel estimate of the density.

The data are highly skewed, with a 97th percentile of approximately $40,000 and a maximum of $1,000,000. The `kdensity totexp` command will therefore bunch 97% of the density in the first 4% of the x axis. One possibility is to type `kdensity totexp if totexp < 40000`, but this produces a kernel density estimate assuming the data are truncated at $40,000. Instead, we use command `kdensity totexp`, we save the evaluation points in `kx1` and the kernel density estimates in `kd1`, and then we line-plot `kd1` against `kx1`.

We do this for both the level and the natural logarithm of medical expenditures, and we use `graph combine` to produce a figure that includes both density graphs (shown in figure 3.1). We have

```
. * Kernel density plots with adjustment for highly skewed data
. kdensity totexp if posexp==1, generate (kx1 kd1) n(500)
. graph twoway (line kd1 kx1) if kx1 < 40000, name(levels)
. kdensity ltotexp if posexp==1, generate (kx2 kd2) n(500)
. graph twoway (line kd2 kx2) if kx2 < ln(40000), name(logs)
. graph combine levels logs, iscale(1.0)
```

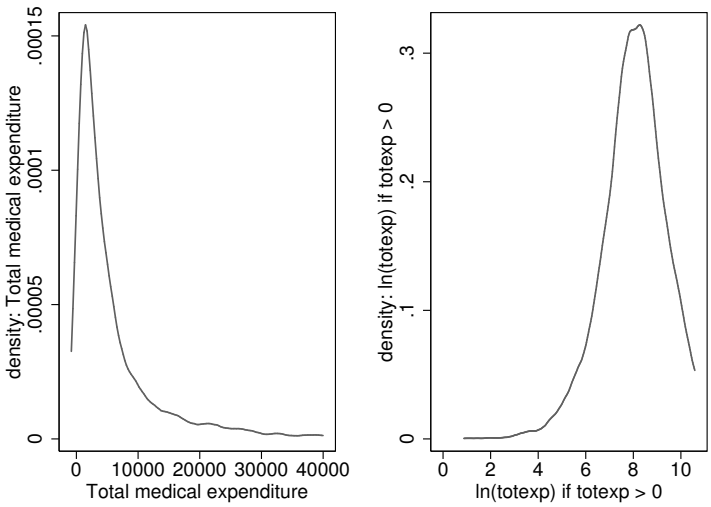

Figure 3.1. Comparison of densities of level and natural logarithm of medical expenditures

Only positive expenditures are considered, and for graph readability, the very long right tail of `totexp` has been truncated at $40,000. In figure 3.1, the distribution of `totexp` is very right-skewed, whereas that of `ltotexp` is fairly symmetric.

3.3 Regression in levels and logs

We present the linear regression model, first in levels and then for a transformed dependent variable, here in logs.

3.3.1 Basic regression theory

We begin by introducing terminology used throughout the rest of this book. Let $\boldsymbol{\theta}$ denote the vector of parameters to be estimated, and let $\widehat{\boldsymbol{\theta}}$ denote an estimator of $\boldsymbol{\theta}$. Ideally, the distribution of $\widehat{\boldsymbol{\theta}}$ is centered on $\boldsymbol{\theta}$ with small variance, for precision, and a known distribution, to permit statistical inference. We restrict analysis to estimators that are consistent for $\boldsymbol{\theta}$, meaning that in infinitely large samples, $\widehat{\boldsymbol{\theta}}$ equals $\boldsymbol{\theta}$ aside from negligible random variation. This is denoted by $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$ or more formally by $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ denotes the unknown "true" parameter value. A necessary condition for consistency is correct model specification or, in some leading cases, correct specification of key components of the model, most notably the conditional mean.

Under additional assumptions, the estimators considered in this book are asymptotically normally distributed, meaning that their distribution is well approximated by the multivariate normal in large samples. This is denoted by

$$\widehat{\boldsymbol{\theta}} \overset{a}{\sim} N\{\boldsymbol{\theta},\ \text{Var}(\widehat{\boldsymbol{\theta}})\}$$

where $\text{Var}(\widehat{\boldsymbol{\theta}})$ denotes the (asymptotic) variance–covariance matrix of the estimator (VCE). More efficient estimators have smaller VCEs. The VCE depends on unknown parameters, so we use an estimate of the VCE, denoted by $\widehat{V}(\widehat{\boldsymbol{\theta}})$. Standard errors of the parameter estimates are obtained as the square root of diagonal entries in $\widehat{V}(\widehat{\boldsymbol{\theta}})$. Different assumptions about the data-generating process (DGP), such as heteroskedasticity, can lead to different estimates of the VCE.

Test statistics based on asymptotic normal results lead to the use of the standard normal distribution and chi-squared distribution to compute critical values and p-values. For some estimators, notably, the OLS estimator, tests are instead based on the t distribution and the F distribution. This makes essentially no difference in large samples with, say, degrees of freedom greater than 100, but it may provide a better approximation in smaller samples.

3.3.2 OLS regression and matrix algebra

The goal of linear regression is to estimate the parameters of the linear conditional mean

$$E(y|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K \tag{3.1}$$

where usually an intercept is included so that $x_1 = 1$. Here $\mathbf{x}$ is a $K \times 1$ column vector with the jth entry—the jth regressor x_j—and $\boldsymbol{\beta}$ is a $K \times 1$ column vector with the jth entry β_j.

Sometimes $E(y|\mathbf{x})$ is of direct interest for prediction. More often, however, econometrics studies are interested in one or more of the associated marginal effects (MEs),

$$\frac{\partial E(y|\mathbf{x})}{\partial x_j} = \beta_j$$

for the jth regressor. For example, we are interested in the marginal effect of supplementary private health insurance on medical expenditures. An attraction of the linear model is that estimated MEs are given directly by estimates of the slope coefficients.

The linear regression model specifies an additive error so that, for the typical ith observation,

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \ u_i, \quad i = 1, \ldots, N$$

The OLS estimator minimizes the sum of squared errors, $\sum_{i=1}^{N}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$.

Matrix notation provides a compact way to represent the estimator and variance matrix formulas that involve sums of products and cross products. We define the $N \times 1$

column vector $\mathbf{y}$ to have the ith entry y_i, and we define the $N \times K$ regressor matrix $\mathbf{X}$ to have the ith row $\mathbf{x}_i'$. Then the OLS estimator can be written in several ways, with

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= \left(\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i y_i \\
&= \begin{bmatrix} \sum_{i=1}^{N} x_{1i}^2 & \sum_{i=1}^{N} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{N} x_{1i}x_{Ki} \\ \sum_{i=1}^{N} x_{2i}x_{1i} & \sum_{i=1}^{N} x_{2i}^2 & & \vdots \\ & & \ddots & \\ \sum_{i=1}^{N} x_{Ki}x_{1i} & & \cdots & \sum_{i=1}^{N} x_{Ki}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{N} x_{1i}y_i \\ \sum_{i=1}^{N} x_{2i}y_i \\ \vdots \\ \sum_{i=1}^{N} x_{Ki}y_i \end{bmatrix}
\end{aligned}$$

We define all vectors as column vectors, with a transpose if row vectors are desired. By contrast, Stata commands and Mata commands define vectors as row vectors, so in parts of Stata and Mata code, we need to take a transpose to conform to the notation in the book.

3.3.3 Properties of the OLS estimator

The properties of any estimator vary with the assumptions made about the DGP. For the linear regression model, this reduces to assumptions about the regression error u_i.

The starting point for analysis is to assume that u_i satisfies the following classical conditions:

1. $E(u_i|\mathbf{x}_i) = \mathbf{0}$ (exogeneity of regressors)
2. $E(u_i^2|\mathbf{x}_i) = \sigma^2$ (conditional homoskedasticity)
3. $E(u_i u_j|\mathbf{x}_i, \mathbf{x}_j) = \mathbf{0},\ i \neq j$, (conditionally uncorrelated observations)

Assumption 1 is essential for consistent estimation of $\boldsymbol{\beta}$ and implies that the conditional mean given in (3.1) is correctly specified. This means that the conditional mean is linear and that all relevant variables have been included in the regression. Assumption 1 is relaxed in chapter 6.

Assumptions 2 and 3 determine the form of the VCE of $\widehat{\boldsymbol{\beta}}$. Assumptions 1–3 lead to $\widehat{\boldsymbol{\beta}}$ being asymptotically normally distributed with the default estimator of the VCE

$$\widehat{V}_{\text{default}}(\widehat{\boldsymbol{\beta}}) = s^2(\mathbf{X}'\mathbf{X})^{-1}$$

where

$$s^2 = (N-k)^{-1}\sum_i \widehat{u}_i^2 \tag{3.2}$$

and $\widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$. Under assumptions 1–3, the OLS estimator is fully efficient. If, additionally, u_i is normally distributed, then "t statistics" are exactly t distributed. This

fourth assumption is not made, but it is common to continue to use the t distribution in the hope that it provides a better approximation than the standard normal in finite samples.

When assumptions 2 and 3 are relaxed, OLS is no longer fully efficient. In chapter 5, we present examples of more-efficient feasible generalized least-squares (FGLS) estimation. In the current chapter, we continue to use the OLS estimator, as is often done in practice, but we use alternative estimates of the VCE that are valid when assumption 2, assumption 3, or both are relaxed.

3.3.4 Heteroskedasticity-robust standard errors

Given assumptions 1 and 3, but not 2, we have heteroskedastic uncorrelated errors. Then a robust estimator, or more precisely a heteroskedasticity-robust estimator, of the VCE of the OLS estimator is

$$\widehat{V}_{\text{robust}}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\left(\frac{N}{N-k}\sum\nolimits_i \widehat{u}_i^2\mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{X}'\mathbf{X})^{-1} \tag{3.3}$$

For cross-section data that are independent, this estimator, introduced by White (1980), has supplanted the default variance matrix estimate in most applied work because heteroskedasticity is the norm, and in that case, the default estimate of the VCE is incorrect.

In Stata, a robust estimate of the VCE is obtained by using the `vce(robust)` option of the `regress` command, as illustrated in section 3.4.2. Related options are `vce(hc2)` and `vce(hc3)`, which may provide better heteroskedasticity-robust estimates of the VCE when the sample size is small; see [R] **regress**. The robust estimator of the VCE has been extended to other estimators and models, and a feature of Stata is the `vce(robust)` option, which is applicable for many estimation commands. Some user-written commands use `robust` in place of `vce(robust)`.

3.3.5 Cluster–robust standard errors

When errors for different observations are correlated, assumption 3 is violated. Then both default and robust estimates of the VCE are invalid. For time-series data, this is the case if errors are serially correlated, and the `newey` command should be used. For cross-section data, this can arise when errors are clustered.

Clustered or grouped errors are errors that are correlated within a cluster or group and are uncorrelated across clusters. A simple example of clustering arises when sampling is of independent units but errors for individuals within the unit are correlated. For example, 100 independent villages may be sampled, with several people from each village surveyed. Then, if a regression model overpredicts y for one village member, it is likely to overpredict for other members of the same village, indicating positive correlation. Similar comments apply when sampling is of households with several individuals in each household. Another leading example is panel data with independence over individuals but with correlation over time for a given individual.

Given assumption 1, but not 2 or 3, a cluster–robust estimator of the VCE of the OLS estimator is

$$\widehat{V}_{\text{cluster}}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\left(\frac{G}{G-1}\frac{N-1}{N-k}\sum\nolimits_g \mathbf{X}_g\widehat{\mathbf{u}}_g\widehat{\mathbf{u}}_g'\mathbf{X}_g'\right)(\mathbf{X}'\mathbf{X})^{-1}$$

where $g = 1, \ldots, G$ denotes the cluster (such as village), $\widehat{\mathbf{u}}_g$ is the vector of residuals for the observations in the gth cluster, and $\mathbf{X}_g$ is a matrix of the regressors for the observations in the gth cluster. The key assumptions made are error independence across clusters and that the number of clusters $G \to \infty$.

Cluster–robust standard errors can be computed by using the `vce(cluster` *clustvar*`)` option in Stata, where clusters are defined by the different values taken by the *clustvar* variable. The estimate of the VCE is in fact heteroskedasticity-robust and cluster–robust, because there is no restriction on $\text{Cov}(u_{gi}, u_{gj})$. The cluster VCE estimate can be applied to many estimators and models; see section 9.6.

Cluster–robust standard errors must be used when data are clustered. For a scalar regressor x, a rule of thumb is that cluster–robust standard errors are $\sqrt{1 + \rho_x\rho_u(M-1)}$ times the incorrect default standard errors, where ρ_x is the within-cluster correlation coefficient of the regressor, ρ_u is the within-cluster correlation coefficient of the error, and M is the average cluster size.

It can be necessary to use cluster–robust standard errors even where it is not immediately obvious. This is particularly the case when a regressor is an aggregated or macro variable, because then $\rho_x = 1$. For example, suppose we use data from the U.S. Current Population Survey and regress individual earnings on individual characteristics and a state-level regressor that does not vary within a state. Then, if there are many individuals in each state so M is large, even slight error correlation for individuals in the same state can lead to great downward bias in default standard errors and in heteroskedasticity-robust standard errors. Clustering can also be induced by the design of sample surveys. This topic is pursued in section 5.5.

3.3.6 Regression in logs

The medical expenditure data are very right-skewed. Then a linear model in levels can provide very poor predictions because it restricts the effects of regressors to be additive. For example, aging 10 years is assumed to increase medical expenditures by the same amount regardless of observed health status. Instead, it is more reasonable to assume that aging 10 years has a multiplicative effect. For example, it may increase medical expenditures by 20%.

We begin with an exponential mean model for positive expenditures, with error that is also multiplicative, so $y_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})\varepsilon_i$. Defining $\varepsilon_i = \exp(u_i)$, we have $y_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + u_i)$, and taking the natural logarithm, we fit the log-linear model

$$\ln y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

by OLS regression of $\ln y$ on $\mathbf{x}$. The conditional mean of $\ln y$ is being modeled, rather than the conditional mean of y. In particular,

$$E(\ln y|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$$

assuming u_i is independent with conditional mean zero.

Parameter interpretation requires care. For regression of $\ln y$ on $\mathbf{x}$, the coefficient β_j measures the effect of a change in regressor x_j on $E(\ln y|\mathbf{x})$, but ultimate interest lies instead on the effect on $E(y|\mathbf{x})$. Some algebra shows that β_j measures the proportionate change in $E(y|\mathbf{x})$ as x_j changes, called a semielasticity, rather than the level of change in $E(y|\mathbf{x})$. For example, if $\beta_j = 0.02$, then a one-unit change in x_j is associated with a proportionate increase of 0.02, or 2%, in $E(y|\mathbf{x})$.

Prediction of $E(y|\mathbf{x})$ is substantially more difficult because it can be shown that $E(\ln y|\mathbf{x}) \neq \exp(\mathbf{x}'\boldsymbol{\beta})$. This is pursued in section 3.6.3.

3.4 Basic regression analysis

We use `regress` to run an OLS regression of the natural logarithm of medical expenditures, `ltotexp`, on `suppins` and several demographic and health-status measures. Using $\ln y$ rather than y as the dependent variable leads to no change in the implementation of OLS but, as already noted, will change the interpretation of coefficients and predictions.

Many of the details we provide in this section are applicable to all Stata estimation commands, not just to `regress`.

3.4.1 Correlations

Before regression, it can be useful to investigate pairwise correlations of the dependent variables and key regressor variables by using `correlate`. We have

```
. * Pairwise correlations for dependent variable and regressor variables
. correlate ltotexp suppins phylim actlim totchr age female income
(obs=2955)
```

	ltotexp	suppins	phylim	actlim	totchr	age
ltotexp	1.0000					
suppins	0.0941	1.0000				
phylim	0.2924	-0.0243	1.0000			
actlim	0.2888	-0.0675	0.5904	1.0000		
totchr	0.4283	0.0124	0.3334	0.3260	1.0000	
age	0.0858	-0.1226	0.2538	0.2394	0.0904	1.0000
female	-0.0058	-0.0796	0.0943	0.0499	0.0557	0.0774
income	0.0023	0.1943	-0.1142	-0.1483	-0.0816	-0.1542

	female	income
female	1.0000	
income	-0.1312	1.0000

Medical expenditures are most highly correlated with the health-status measures `phylim`, `actlim`, and `totchr`. The regressors are only weakly correlated with each other, aside from the health-status measures. Note that `correlate` restricts analysis to the 2,955 observations where data are available for all variables in the variable list. The related command `pwcorr`, not demonstrated, with the `sig` option gives the statistical significance of the correlations.

3.4.2 The regress command

The `regress` command performs OLS regression and yields an analysis-of-variance table, goodness-of-fit statistics, coefficient estimates, standard errors, t statistics, p-values, and confidence intervals. The syntax of the command is

`regress` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [, *options*]

Other Stata estimation commands have similar syntaxes. The output from `regress` is similar to that from many linear regression packages.

For independent cross-section data, the standard approach is to use the `vce(robust)` option, which gives standard errors that are valid even if model errors are heteroskedastic; see section 3.3.4. In that case, the analysis-of-variance table, based on the assumption of homoskedasticity, is dropped from the output. We obtain

```
. * OLS regression with heteroskedasticity-robust standard errors
. regress ltotexp suppins phylim actlim totchr age female income, vce(robust)
Linear regression                                      Number of obs =    2955
                                                       F(  7,  2947) =  126.97
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.2289
                                                       Root MSE      =  1.2023

------------------------------------------------------------------------------
             |               Robust
    ltotexp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
    suppins |   .2556428   .0465982     5.49   0.000     .1642744    .3470112
     phylim |   .3020598    .057705     5.23   0.000     .1889136     .415206
     actlim |   .3560054   .0634066     5.61   0.000     .2316797    .4803311
     totchr |   .3758201   .0187185    20.08   0.000     .3391175    .4125228
        age |   .0038016   .0037028     1.03   0.305    -.0034587     .011062
     female |  -.0843275    .045654    -1.85   0.065    -.1738444    .0051894
     income |   .0025498   .0010468     2.44   0.015     .0004973    .0046023
      _cons |   6.703737   .2825751    23.72   0.000     6.149673    7.257802
------------------------------------------------------------------------------
```

The regressors are jointly statistically significant, because the overall F statistic of 126.97 has a p-value of 0.000. At the same time, much of the variation is unexplained with $R^2 = 0.2289$. The root MSE statistic reports s, the standard error of the regression, defined in (3.2). By using a two-sided test at level 0.05, all regressors are individually statistically significant because $p < 0.05$, aside from `age` and `female`. The strong statistical insignificance of age may be due to sample restriction to elderly people and the inclusion of several health-status measures that capture well the health effect of age.

Statistical significance of coefficients is easily established. More important is the economic significance of coefficients, meaning the measured impact of regressors on medical expenditures. This is straightforward for regression in levels, because we can directly use the estimated coefficients. But here the regression is in logs. From section 3.3.6, in the log-linear model, parameters need to be interpreted as semielasticities. For example, the coefficient on `suppins` is 0.256. This means that private supplementary insurance is associated with a 0.256 proportionate rise, or a 25.6% rise, in medical expenditures. Similarly, large effects are obtained for the health-status measures, whereas health expenditures for women are 8.4% lower than those for men after controlling for other characteristics. The `income` coefficient of 0.0025 suggests a very small effect, but this is misleading. The standard deviation of `income` is 22, so a 1–standard deviation in `income` leads to a 0.055 proportionate rise, or 5.5% rise, in medical expenditures.

MEs in nonlinear models are discussed in more detail in section 10.6. The preceding interpretations are based on calculus methods that consider very small changes in the regressor. For larger changes in the regressor, the finite-difference method is more appropriate. Then the interpretation in the log-linear model is similar to that for the exponential conditional mean model; see section 10.6.4. For example, the estimated effect of going from no supplementary insurance (`suppins=0`) to having supplementary insurance (`suppins=1`) is more precisely a $100 \times (e^{0.256} - 1)$, or 29.2%, rise.

The `regress` command provides additional results that are not listed. In particular, the estimate of the VCE is stored in the matrix `e(V)`. Ways to access this and other stored results from regression have been given in section 1.6. Various postestimation commands enable prediction, computation of residuals, hypothesis testing, and model specification tests. Many of these are illustrated in subsequent sections. Two useful commands are

```
. * Display stored results and list available postestimation commands
. ereturn list
  (output omitted)
. help regress postestimation
  (output omitted)
```

3.4.3 Hypothesis tests

The `test` command performs hypothesis tests using the Wald test procedure that uses the estimated model coefficients and VCE. We present some leading examples here, with a more extensive discussion deferred to section 12.3. The F statistic version of the Wald test is used after `regress`, whereas for many other estimators the chi-squared version is instead used.

A common test is one of equality of coefficients. For example, consider testing that having a functional limitation has the same impact on medical expenditures as having an activity limitation. The test of $H_0: \beta_{\texttt{phylim}} = \beta_{\texttt{actlim}}$ against $H_a: \beta_{\texttt{phylim}} \neq \beta_{\texttt{actlim}}$ is implemented as