

Contents

	List of tables	xxxv
	List of figures	xxxvii
	Preface	xxxix
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	Abbreviations, case sensitivity, and wildcards	9
1.3.5	Arithmetic, relational, and logical operators	9
1.3.6	Error messages	10
1.4	Do-files and log files	10
1.4.1	Writing a do-file	10
1.4.2	Running do-files	11
1.4.3	Log files	12
1.4.4	A three-step process	13
1.4.5	Comments and long lines	13
1.4.6	Different implementations of Stata	14

1.5	Scalars and matrices	15
1.5.1	Scalars	15
1.5.2	Matrices	15
1.6	Using results from Stata commands	16
1.6.1	Using results from the r-class command summarize	16
1.6.2	Using results from the e-class command regress	17
1.7	Global and local macros	19
1.7.1	Global macros	19
1.7.2	Local macros	20
1.7.3	Scalar or macro?	21
1.8	Looping commands	22
1.8.1	The foreach loop	23
1.8.2	The forvalues loop	23
1.8.3	The while loop	24
1.8.4	The continue command	24
1.9	Some useful commands	24
1.10	Template do-file	25
1.11	User-written commands	25
1.12	Stata resources	26
1.13	Exercises	26
2	Data management and graphics	29
2.1	Introduction	29
2.2	Types of data	29
2.2.1	Text or ASCII data	30
2.2.2	Internal numeric data	30
2.2.3	String data	31
2.2.4	Formats for displaying numeric data	31
2.3	Inputting data	32
2.3.1	General principles	32
2.3.2	Inputting data already in Stata format	33

2.3.3	Inputting data from the keyboard	34
2.3.4	Inputting nontext data	34
2.3.5	Inputting text data from a spreadsheet	35
2.3.6	Inputting text data in free format	36
2.3.7	Inputting text data in fixed format	36
2.3.8	Dictionary files	37
2.3.9	Common pitfalls	37
2.4	Data management	38
2.4.1	PSID example	38
2.4.2	Naming and labeling variables	41
2.4.3	Viewing data	42
2.4.4	Using original documentation	43
2.4.5	Missing values	43
2.4.6	Imputing missing data	45
2.4.7	Transforming data (generate, replace, egen, recode)	45
	The generate and replace commands	46
	The egen command	46
	The recode command	47
	The by prefix	47
	Indicator variables	47
	Set of indicator variables	48
	Interactions	49
	Demeaning	50
2.4.8	Saving data	51
2.4.9	Selecting the sample	51
2.5	Manipulating datasets	53
2.5.1	Ordering observations and variables	53
2.5.2	Preserving and restoring a dataset	53
2.5.3	Wide and long forms for a dataset	54

2.5.4	Merging datasets	54
2.5.5	Appending datasets	56
2.6	Graphical display of data	57
2.6.1	Stata graph commands	57
	Example graph commands	57
	Saving and exporting graphs	58
	Learning how to use graph commands	59
2.6.2	Box-and-whisker plot	60
2.6.3	Histogram	61
2.6.4	Kernel density plot	62
2.6.5	Twoway scatterplots and fitted lines	64
2.6.6	Lowess, kernel, local linear, and nearest-neighbor regression	65
2.6.7	Multiple scatterplots	67
2.7	Stata resources	68
2.8	Exercises	68
3	Linear regression basics	71
3.1	Introduction	71
3.2	Data and data summary	71
3.2.1	Data description	71
3.2.2	Variable description	72
3.2.3	Summary statistics	73
3.2.4	More-detailed summary statistics	74
3.2.5	Tables for data	75
3.2.6	Statistical tests	78
3.2.7	Data plots	78
3.3	Regression in levels and logs	79
3.3.1	Basic regression theory	79
3.3.2	OLS regression and matrix algebra	80
3.3.3	Properties of the OLS estimator	81
3.3.4	Heteroskedasticity-robust standard errors	82

3.3.5	Cluster-robust standard errors	82
3.3.6	Regression in logs	83
3.4	Basic regression analysis	84
3.4.1	Correlations	84
3.4.2	The regress command	85
3.4.3	Hypothesis tests	86
3.4.4	Tables of output from several regressions	87
3.4.5	Even better tables of regression output	88
3.5	Specification analysis	90
3.5.1	Specification tests and model diagnostics	90
3.5.2	Residual diagnostic plots	91
3.5.3	Influential observations	92
3.5.4	Specification tests	93
	Test of omitted variables	93
	Test of the Box-Cox model	94
	Test of the functional form of the conditional mean	95
	Heteroskedasticity test	96
	Omnibus test	97
3.5.5	Tests have power in more than one direction	98
3.6	Prediction	100
3.6.1	In-sample prediction	100
3.6.2	Marginal effects	102
3.6.3	Prediction in logs: The retransformation problem	103
3.6.4	Prediction exercise	104
3.7	Sampling weights	105
3.7.1	Weights	106
3.7.2	Weighted mean	106
3.7.3	Weighted regression	107
3.7.4	Weighted prediction and MEs	109
3.8	OLS using Mata	109

3.9	Stata resources	111
3.10	Exercises	111
4	Simulation	113
4.1	Introduction	113
4.2	Pseudorandom-number generators: Introduction	114
4.2.1	Uniform random-number generation	114
4.2.2	Draws from normal	116
4.2.3	Draws from t, chi-squared, F, gamma, and beta	117
4.2.4	Draws from binomial, Poisson, and negative binomial	118
	Independent (but not identically distributed) draws from binomial	118
	Independent (but not identically distributed) draws from Poisson	119
	Histograms and density plots	120
4.3	Distribution of the sample mean	121
4.3.1	Stata program	122
4.3.2	The simulate command	123
4.3.3	Central limit theorem simulation	123
4.3.4	The postfile command	124
4.3.5	Alternative central limit theorem simulation	125
4.4	Pseudorandom-number generators: Further details	125
4.4.1	Inverse-probability transformation	126
4.4.2	Direct transformation	127
4.4.3	Other methods	127
4.4.4	Draws from truncated normal	128
4.4.5	Draws from multivariate normal	129
	Direct draws from multivariate normal	129
	Transformation using Cholesky decomposition	130
4.4.6	Draws using Markov chain Monte Carlo method	130
4.5	Computing integrals	132
4.5.1	Quadrature	133

4.5.2	Monte Carlo integration	133
4.5.3	Monte Carlo integration using different S	134
4.6	Simulation for regression: Introduction	135
4.6.1	Simulation example: OLS with χ^2 errors	135
4.6.2	Interpreting simulation output	138
	Unbiasedness of estimator	138
	Standard errors	138
	t statistic	138
	Test size	139
	Number of simulations	140
4.6.3	Variations	140
	Different sample size and number of simulations	140
	Test power	140
	Different error distributions	141
4.6.4	Estimator inconsistency	141
4.6.5	Simulation with endogenous regressors	142
4.7	Stata resources	144
4.8	Exercises	144
5	GLS regression	147
5.1	Introduction	147
5.2	GLS and FGLS regression	147
5.2.1	GLS for heteroskedastic errors	147
5.2.2	GLS and FGLS	148
5.2.3	Weighted least squares and robust standard errors	149
5.2.4	Leading examples	149
5.3	Modeling heteroskedastic data	150
5.3.1	Simulated dataset	150
5.3.2	OLS estimation	151
5.3.3	Detecting heteroskedasticity	152
5.3.4	FGLS estimation	154

5.3.5	WLS estimation	156
5.4	System of linear regressions	156
5.4.1	SUR model	156
5.4.2	The sureg command	157
5.4.3	Application to two categories of expenditures	158
5.4.4	Robust standard errors	160
5.4.5	Testing cross-equation constraints	161
5.4.6	Imposing cross-equation constraints	162
5.5	Survey data: Weighting, clustering, and stratification	163
5.5.1	Survey design	164
5.5.2	Survey mean estimation	167
5.5.3	Survey linear regression	167
5.6	Stata resources	169
5.7	Exercises	169
6	Linear instrumental-variables regression	171
6.1	Introduction	171
6.2	IV estimation	171
6.2.1	Basic IV theory	171
6.2.2	Model setup	173
6.2.3	IV estimators: IV, 2SLS, and GMM	174
6.2.4	Instrument validity and relevance	175
6.2.5	Robust standard-error estimates	176
6.3	IV example	177
6.3.1	The ivregress command	177
6.3.2	Medical expenditures with one endogenous regressor	178
6.3.3	Available instruments	179
6.3.4	IV estimation of an exactly identified model	180
6.3.5	IV estimation of an overidentified model	181
6.3.6	Testing for regressor endogeneity	182
6.3.7	Tests of overidentifying restrictions	185

6.3.8	IV estimation with a binary endogenous regressor	186
6.4	Weak instruments	188
6.4.1	Finite-sample properties of IV estimators	188
6.4.2	Weak instruments	189
	Diagnostics for weak instruments	189
	Formal tests for weak instruments	190
6.4.3	The estat firststage command	191
6.4.4	Just-identified model	191
6.4.5	Overidentified model	193
6.4.6	More than one endogenous regressor	195
6.4.7	Sensitivity to choice of instruments	195
6.5	Better inference with weak instruments	197
6.5.1	Conditional tests and confidence intervals	197
6.5.2	LIML estimator	199
6.5.3	Jackknife IV estimator	199
6.5.4	Comparison of 2SLS, LIML, JIVE, and GMM	200
6.6	3SLS systems estimation	201
6.7	Stata resources	203
6.8	Exercises	203
7	Quantile regression	205
7.1	Introduction	205
7.2	QR	205
7.2.1	Conditional quantiles	206
7.2.2	Computation of QR estimates and standard errors	207
7.2.3	The qreg, bsqreg, and sqreg commands	207
7.3	QR for medical expenditures data	208
7.3.1	Data summary	208
7.3.2	QR estimates	209
7.3.3	Interpretation of conditional quantile coefficients	210
7.3.4	Retransformation	211

7.3.5	Comparison of estimates at different quantiles	212
7.3.6	Heteroskedasticity test	213
7.3.7	Hypothesis tests	214
7.3.8	Graphical display of coefficients over quantiles	215
7.4	QR for generated heteroskedastic data	216
7.4.1	Simulated dataset	216
7.4.2	QR estimates	219
7.5	QR for count data	220
7.5.1	Quantile count regression	221
7.5.2	The qcount command	222
7.5.3	Summary of doctor visits data	222
7.5.4	Results from QCR	224
7.6	Stata resources	226
7.7	Exercises	226
8	Linear panel-data models: Basics	229
8.1	Introduction	229
8.2	Panel-data methods overview	229
8.2.1	Some basic considerations	230
8.2.2	Some basic panel models	231
	Individual-effects model	231
	Fixed-effects model	231
	Random-effects model	232
	Pooled model or population-averaged model	232
	Two-way-effects model	232
	Mixed linear models	233
8.2.3	Cluster-robust inference	233
8.2.4	The xtreg command	233
8.2.5	Stata linear panel-data commands	234
8.3	Panel-data summary	234
8.3.1	Data description and summary statistics	234

8.3.2	Panel-data organization	236
8.3.3	Panel-data description	237
8.3.4	Within and between variation	238
8.3.5	Time-series plots for each individual	241
8.3.6	Overall scatterplot	242
8.3.7	Within scatterplot	243
8.3.8	Pooled OLS regression with cluster-robust standard errors .	244
8.3.9	Time-series autocorrelations for panel data	245
8.3.10	Error correlation in the RE model	247
8.4	Pooled or population-averaged estimators	248
8.4.1	Pooled OLS estimator	248
8.4.2	Pooled FGLS estimator or population-averaged estimator .	248
8.4.3	The xtreg, pa command	249
8.4.4	Application of the xtreg, pa command	250
8.5	Within estimator	251
8.5.1	Within estimator	251
8.5.2	The xtreg, fe command	251
8.5.3	Application of the xtreg, fe command	252
8.5.4	Least-squares dummy-variables regression	253
8.6	Between estimator	254
8.6.1	Between estimator	254
8.6.2	Application of the xtreg, be command	255
8.7	RE estimator	255
8.7.1	RE estimator	255
8.7.2	The xtreg, re command	256
8.7.3	Application of the xtreg, re command	256
8.8	Comparison of estimators	257
8.8.1	Estimates of variance components	257
8.8.2	Within and between R-squared	258
8.8.3	Estimator comparison	258

8.8.4	Fixed effects versus random effects	259
8.8.5	Hausman test for fixed effects	260
	The hausman command	260
	Robust Hausman test	261
8.8.6	Prediction	262
8.9	First-difference estimator	263
8.9.1	First-difference estimator	263
8.9.2	Strict and weak exogeneity	264
8.10	Long panels	265
8.10.1	Long-panel dataset	265
8.10.2	Pooled OLS and PFGLS	266
8.10.3	The xtpcse and xtgls commands	267
8.10.4	Application of the xtgls, xtpcse, and xtscs commands	268
8.10.5	Separate regressions	270
8.10.6	FE and RE models	271
8.10.7	Unit roots and cointegration	272
8.11	Panel-data management	274
8.11.1	Wide-form data	274
8.11.2	Convert wide form to long form	274
8.11.3	Convert long form to wide form	275
8.11.4	An alternative wide-form data	276
8.12	Stata resources	278
8.13	Exercises	278
9	Linear panel-data models: Extensions	281
9.1	Introduction	281
9.2	Panel IV estimation	281
9.2.1	Panel IV	281
9.2.2	The xtivreg command	282
9.2.3	Application of the xtivreg command	282
9.2.4	Panel IV extensions	284

9.3	Hausman–Taylor estimator	284
9.3.1	Hausman–Taylor estimator	284
9.3.2	The <code>xthtaylor</code> command	285
9.3.3	Application of the <code>xthtaylor</code> command	285
9.4	Arellano–Bond estimator	287
9.4.1	Dynamic model	287
9.4.2	IV estimation in the FD model	288
9.4.3	The <code>xtabond</code> command	289
9.4.4	Arellano–Bond estimator: Pure time series	290
9.4.5	Arellano–Bond estimator: Additional regressors	292
9.4.6	Specification tests	294
9.4.7	The <code>xtdpdsys</code> command	295
9.4.8	The <code>xtdpd</code> command	297
9.5	Mixed linear models	298
9.5.1	Mixed linear model	298
9.5.2	The <code>xtmixed</code> command	299
9.5.3	Random-intercept model	300
9.5.4	Cluster–robust standard errors	301
9.5.5	Random-slopes model	302
9.5.6	Random-coefficients model	303
9.5.7	Two-way random-effects model	304
9.6	Clustered data	306
9.6.1	Clustered dataset	306
9.6.2	Clustered data using nonpanel commands	306
9.6.3	Clustered data using panel commands	307
9.6.4	Hierarchical linear models	310
9.7	Stata resources	311
9.8	Exercises	311
10	Nonlinear regression methods	313
10.1	Introduction	313

10.2	Nonlinear example: Doctor visits	314
10.2.1	Data description	314
10.2.2	Poisson model description	315
10.3	Nonlinear regression methods	316
10.3.1	MLE	316
10.3.2	The poisson command	317
10.3.3	Postestimation commands	318
10.3.4	NLS	319
10.3.5	The nl command	319
10.3.6	GLM	321
10.3.7	The glm command	321
10.3.8	Other estimators	322
10.4	Different estimates of the VCE	323
10.4.1	General framework	323
10.4.2	The vce() option	324
10.4.3	Application of the vce() option	324
10.4.4	Default estimate of the VCE	326
10.4.5	Robust estimate of the VCE	326
10.4.6	Cluster-robust estimate of the VCE	327
10.4.7	Heteroskedasticity- and autocorrelation-consistent estimate of the VCE	328
10.4.8	Bootstrap standard errors	328
10.4.9	Statistical inference	329
10.5	Prediction	329
10.5.1	The predict and predictnl commands	329
10.5.2	Application of predict and predictnl	330
10.5.3	Out-of-sample prediction	331
10.5.4	Prediction at a specified value of one of the regressors	332
10.5.5	Prediction at a specified value of all the regressors	332
10.5.6	Prediction of other quantities	333

10.6	Marginal effects	333
10.6.1	Calculus and finite-difference methods	334
10.6.2	MEs estimates AME, MEM, and MER	334
10.6.3	Elasticities and semielasticities	335
10.6.4	Simple interpretations of coefficients in single-index models	336
10.6.5	The mfx command	337
10.6.6	MEM: Marginal effect at mean	337
	Comparison of calculus and finite-difference methods	338
10.6.7	MER: Marginal effect at representative value	338
10.6.8	AME: Average marginal effect	339
10.6.9	Elasticities and semielasticities	340
10.6.10	AME computed manually	342
10.6.11	Polynomial regressors	343
10.6.12	Interacted regressors	344
10.6.13	Complex interactions and nonlinearities	344
10.7	Model diagnostics	345
10.7.1	Goodness-of-fit measures	345
10.7.2	Information criteria for model comparison	346
10.7.3	Residuals	347
10.7.4	Model-specification tests	348
10.8	Stata resources	349
10.9	Exercises	349
11	Nonlinear optimization methods	351
11.1	Introduction	351
11.2	Newton–Raphson method	351
11.2.1	NR method	351
11.2.2	NR method for Poisson	352
11.2.3	Poisson NR example using Mata	353
	Core Mata code for Poisson NR iterations	353
	Complete Stata and Mata code for Poisson NR iterations	353

11.3	Gradient methods	355
11.3.1	Maximization options	355
11.3.2	Gradient methods	356
11.3.3	Messages during iterations	357
11.3.4	Stopping criteria	357
11.3.5	Multiple maximums	357
11.3.6	Numerical derivatives	358
11.4	The ml command: lf method	359
11.4.1	The ml command	360
11.4.2	The lf method	360
11.4.3	Poisson example: Single-index model	361
11.4.4	Negative binomial example: Two-index model	362
11.4.5	NLS example: Nonlikelihood model	363
11.5	Checking the program	364
11.5.1	Program debugging using ml check and ml trace	365
11.5.2	Getting the program to run	366
11.5.3	Checking the data	366
11.5.4	Multicollinearity and near collinearity	367
11.5.5	Multiple optimums	368
11.5.6	Checking parameter estimation	369
11.5.7	Checking standard-error estimation	370
11.6	The ml command: d0, d1, and d2 methods	371
11.6.1	Evaluator functions	371
11.6.2	The d0 method	373
11.6.3	The d1 method	374
11.6.4	The d1 method with the robust estimate of the VCE	374
11.6.5	The d2 method	375
11.7	The Mata optimize() function	376
11.7.1	Type d and v evaluators	376
11.7.2	Optimize functions	377

11.7.3	Poisson example	377
	Evaluator program for Poisson MLE	377
	The optimize() function for Poisson MLE	378
11.8	Generalized method of moments	379
11.8.1	Definition	380
11.8.2	Nonlinear IV example	380
11.8.3	GMM using the Mata optimize() function	381
11.9	Stata resources	383
11.10	Exercises	383
12	Testing methods	385
12.1	Introduction	385
12.2	Critical values and p-values	385
12.2.1	Standard normal compared with Student's t	386
12.2.2	Chi-squared compared with F	386
12.2.3	Plotting densities	386
12.2.4	Computing p-values and critical values	388
12.2.5	Which distributions does Stata use?	389
12.3	Wald tests and confidence intervals	389
12.3.1	Wald test of linear hypotheses	389
12.3.2	The test command	391
	Test single coefficient	392
	Test several hypotheses	392
	Test of overall significance	393
	Test calculated from retrieved coefficients and VCE	393
12.3.3	One-sided Wald tests	394
12.3.4	Wald test of nonlinear hypotheses (delta method)	395
12.3.5	The testnl command	395
12.3.6	Wald confidence intervals	396
12.3.7	The lincom command	396
12.3.8	The nlcom command (delta method)	397

12.3.9	Asymmetric confidence intervals	398
12.4	Likelihood-ratio tests	399
12.4.1	Likelihood-ratio tests	399
12.4.2	The lrtest command	401
12.4.3	Direct computation of LR tests	401
12.5	Lagrange multiplier test (or score test)	402
12.5.1	LM tests	402
12.5.2	The estat command	403
12.5.3	LM test by auxiliary regression	403
12.6	Test size and power	405
12.6.1	Simulation DGP: OLS with chi-squared errors	405
12.6.2	Test size	406
12.6.3	Test power	407
12.6.4	Asymptotic test power	410
12.7	Specification tests	411
12.7.1	Moment-based tests	411
12.7.2	Information matrix test	411
12.7.3	Chi-squared goodness-of-fit test	412
12.7.4	Overidentifying restrictions test	412
12.7.5	Hausman test	412
12.7.6	Other tests	413
12.8	Stata resources	413
12.9	Exercises	413
13	Bootstrap methods	415
13.1	Introduction	415
13.2	Bootstrap methods	415
13.2.1	Bootstrap estimate of standard error	415
13.2.2	Bootstrap methods	416
13.2.3	Asymptotic refinement	416
13.2.4	Use the bootstrap with caution	416

13.3	Bootstrap pairs using the <code>vce(bootstrap)</code> option	417
13.3.1	Bootstrap-pairs method to estimate VCE	417
13.3.2	The <code>vce(bootstrap)</code> option	418
13.3.3	Bootstrap standard-errors example	418
13.3.4	How many bootstraps?	419
13.3.5	Clustered bootstraps	420
13.3.6	Bootstrap confidence intervals	421
13.3.7	The postestimation <code>estat bootstrap</code> command	422
13.3.8	Bootstrap confidence-intervals example	423
13.3.9	Bootstrap estimate of bias	423
13.4	Bootstrap pairs using the <code>bootstrap</code> command	424
13.4.1	The <code>bootstrap</code> command	424
13.4.2	Bootstrap parameter estimate from a Stata estimation command	425
13.4.3	Bootstrap standard error from a Stata estimation command	426
13.4.4	Bootstrap standard error from a user-written estimation command	426
13.4.5	Bootstrap two-step estimator	427
13.4.6	Bootstrap Hausman test	429
13.4.7	Bootstrap standard error of the coefficient of variation . . .	430
13.5	Bootstraps with asymptotic refinement	431
13.5.1	Percentile-t method	431
13.5.2	Percentile-t Wald test	432
13.5.3	Percentile-t Wald confidence interval	433
13.6	Bootstrap pairs using <code>bsample</code> and <code>simulate</code>	434
13.6.1	The <code>bsample</code> command	434
13.6.2	The <code>bsample</code> command with <code>simulate</code>	434
13.6.3	Bootstrap Monte Carlo exercise	436
13.7	Alternative resampling schemes	436
13.7.1	Bootstrap pairs	437
13.7.2	Parametric bootstrap	437

13.7.3	Residual bootstrap	439
13.7.4	Wild bootstrap	440
13.7.5	Subsampling	441
13.8	The jackknife	441
13.8.1	Jackknife method	441
13.8.2	The vce(jackknife) option and the jackknife command . . .	442
13.9	Stata resources	442
13.10	Exercises	442
14	Binary outcome models	445
14.1	Introduction	445
14.2	Some parametric models	445
14.2.1	Basic model	445
14.2.2	Logit, probit, linear probability, and clog-log models	446
14.3	Estimation	446
14.3.1	Latent-variable interpretation and identification	447
14.3.2	ML estimation	447
14.3.3	The logit and probit commands	448
14.3.4	Robust estimate of the VCE	448
14.3.5	OLS estimation of LPM	448
14.4	Example	449
14.4.1	Data description	449
14.4.2	Logit regression	450
14.4.3	Comparison of binary models and parameter estimates . . .	451
14.5	Hypothesis and specification tests	452
14.5.1	Wald tests	453
14.5.2	Likelihood-ratio tests	453
14.5.3	Additional model-specification tests	454
	Lagrange multiplier test of generalized logit	454
	Heteroskedastic probit regression	455
14.5.4	Model comparison	456

14.6	Goodness of fit and prediction	457
14.6.1	Pseudo-R ² measure	457
14.6.2	Comparing predicted probabilities with sample frequencies	457
14.6.3	Comparing predicted outcomes with actual outcomes	459
14.6.4	The predict command for fitted probabilities	460
14.6.5	The prvalue command for fitted probabilities	461
14.7	Marginal effects	462
14.7.1	Marginal effect at a representative value (MER)	462
14.7.2	Marginal effect at the mean (MEM)	463
14.7.3	Average marginal effect (AME)	464
14.7.4	The prchange command	464
14.8	Endogenous regressors	465
14.8.1	Example	465
14.8.2	Model assumptions	466
14.8.3	Structural-model approach	467
	The ivprobit command	467
	Maximum likelihood estimates	468
	Two-step sequential estimates	469
14.8.4	IVs approach	471
14.9	Grouped data	472
14.9.1	Estimation with aggregate data	473
14.9.2	Grouped-data application	473
14.10	Stata resources	475
14.11	Exercises	475
15	Multinomial models	477
15.1	Introduction	477
15.2	Multinomial models overview	477
15.2.1	Probabilities and MEs	477
15.2.2	Maximum likelihood estimation	478
15.2.3	Case-specific and alternative-specific regressors	479

15.2.4	Additive random-utility model	479
15.2.5	Stata multinomial model commands	480
15.3	Multinomial example: Choice of fishing mode	480
15.3.1	Data description	480
15.3.2	Case-specific regressors	483
15.3.3	Alternative-specific regressors	483
15.4	Multinomial logit model	484
15.4.1	The mlogit command	484
15.4.2	Application of the mlogit command	485
15.4.3	Coefficient interpretation	486
15.4.4	Predicted probabilities	487
15.4.5	MEs	488
15.5	Conditional logit model	489
15.5.1	Creating long-form data from wide-form data	489
15.5.2	The asclogit command	491
15.5.3	The clogit command	491
15.5.4	Application of the asclogit command	492
15.5.5	Relationship to multinomial logit model	493
15.5.6	Coefficient interpretation	493
15.5.7	Predicted probabilities	494
15.5.8	MEs	494
15.6	Nested logit model	496
15.6.1	Relaxing the independence of irrelevant alternatives assumption	497
15.6.2	NL model	497
15.6.3	The nlogit command	498
15.6.4	Model estimates	499
15.6.5	Predicted probabilities	501
15.6.6	MEs	501
15.6.7	Comparison of logit models	502

15.7	Multinomial probit model	503
15.7.1	MNP	503
15.7.2	The mprobit command	503
15.7.3	Maximum simulated likelihood	504
15.7.4	The asmprobit command	505
15.7.5	Application of the asmprobit command	505
15.7.6	Predicted probabilities and MEs	507
15.8	Random-parameters logit	508
15.8.1	Random-parameters logit	508
15.8.2	The mixlogit command	508
15.8.3	Data preparation for mixlogit	509
15.8.4	Application of the mixlogit command	509
15.9	Ordered outcome models	510
15.9.1	Data summary	511
15.9.2	Ordered outcomes	512
15.9.3	Application of the ologit command	512
15.9.4	Predicted probabilities	513
15.9.5	MEs	513
15.9.6	Other ordered models	514
15.10	Multivariate outcomes	514
15.10.1	Bivariate probit	515
15.10.2	Nonlinear SUR	517
15.11	Stata resources	518
15.12	Exercises	518
16	Tobit and selection models	521
16.1	Introduction	521
16.2	Tobit model	521
16.2.1	Regression with censored data	521
16.2.2	Tobit model setup	522
16.2.3	Unknown censoring point	523

16.2.4	Tobit estimation	523
16.2.5	ML estimation in Stata	524
16.3	Tobit model example	524
16.3.1	Data summary	524
16.3.2	Tobit analysis	525
16.3.3	Prediction after tobit	526
16.3.4	Marginal effects	527
	Left-truncated, left-censored, and right-truncated examples	527
	Left-censored case computed directly	528
	Marginal impact on probabilities	529
16.3.5	The ivtobit command	530
16.3.6	Additional commands for censored regression	530
16.4	Tobit for lognormal data	531
16.4.1	Data example	531
16.4.2	Setting the censoring point for data in logs	532
16.4.3	Results	533
16.4.4	Two-limit tobit	534
16.4.5	Model diagnostics	534
16.4.6	Tests of normality and homoskedasticity	535
	Generalized residuals and scores	535
	Test of normality	536
	Test of homoskedasticity	537
16.4.7	Next step?	538
16.5	Two-part model in logs	538
16.5.1	Model structure	538
16.5.2	Part 1 specification	539
16.5.3	Part 2 of the two-part model	540
16.6	Selection model	541
16.6.1	Model structure and assumptions	541
16.6.2	ML estimation of the sample-selection model	543

16.6.3	Estimation without exclusion restrictions	543
16.6.4	Two-step estimation	545
16.6.5	Estimation with exclusion restrictions	546
16.7	Prediction from models with outcome in logs	547
16.7.1	Predictions from tobit	548
16.7.2	Predictions from two-part model	548
16.7.3	Predictions from selection model	549
16.8	Stata resources	550
16.9	Exercises	550
17	Count-data models	553
17.1	Introduction	553
17.2	Features of count data	553
17.2.1	Generated Poisson data	554
17.2.2	Overdispersion and negative binomial data	555
17.2.3	Modeling strategies	556
17.2.4	Estimation methods	557
17.3	Empirical example 1	557
17.3.1	Data summary	557
17.3.2	Poisson model	558
	Poisson model results	559
	Robust estimate of VCE for Poisson MLE	560
	Test of overdispersion	561
	Coefficient interpretation and marginal effects	562
17.3.3	NB2 model	562
	NB2 model results	563
	Fitted probabilities for Poisson and NB2 models	565
	The countfit command	565
	The prvalue command	567
	Discussion	567
	Generalized NB model	567

17.3.4	Nonlinear least-squares estimation	568
17.3.5	Hurdle model	569
	Variants of the hurdle model	571
	Application of the hurdle model	571
17.3.6	Finite-mixture models	575
	FMM specification	575
	Simulated FMM sample with comparisons	575
	ML estimation of the FMM	577
	The fmm command	578
	Application: Poisson finite-mixture model	578
	Interpretation	579
	Comparing marginal effects	580
	Application: NB finite-mixture model	582
	Model selection	584
	Cautionary note	585
17.4	Empirical example 2	585
17.4.1	Zero-inflated data	585
17.4.2	Models for zero-inflated data	586
17.4.3	Results for the NB2 model	587
	The prcounts command	588
17.4.4	Results for ZINB	589
17.4.5	Model comparison	590
	The countfit command	590
	Model comparison using countfit	590
17.5	Models with endogenous regressors	591
17.5.1	Structural-model approach	592
	Model and assumptions	592
	Two-step estimation	593
	Application	593
17.5.2	Nonlinear IV method	596

17.6	Stata resources	598
17.7	Exercises	598
18	Nonlinear panel models	601
18.1	Introduction	601
18.2	Nonlinear panel-data overview	601
18.2.1	Some basic nonlinear panel models	601
	FE models	602
	RE models	602
	Pooled models or population-averaged models	602
	Comparison of models	603
18.2.2	Dynamic models	603
18.2.3	Stata nonlinear panel commands	603
18.3	Nonlinear panel-data example	604
18.3.1	Data description and summary statistics	604
18.3.2	Panel-data organization	606
18.3.3	Within and between variation	606
18.3.4	FE or RE model for these data?	607
18.4	Binary outcome models	607
18.4.1	Panel summary of the dependent variable	607
18.4.2	Pooled logit estimator	608
18.4.3	The xtlogit command	609
18.4.4	The xtgee command	610
18.4.5	PA logit estimator	610
18.4.6	RE logit estimator	611
18.4.7	FE logit estimator	613
18.4.8	Panel logit estimator comparison	615
18.4.9	Prediction and marginal effects	616
18.4.10	Mixed-effects logit estimator	616
18.5	Tobit model	617
18.5.1	Panel summary of the dependent variable	617

18.5.2	RE tobit model	617
18.5.3	Generalized tobit models	618
18.5.4	Parametric nonlinear panel models	619
18.6	Count-data models	619
18.6.1	The xtpoisson command	619
18.6.2	Panel summary of the dependent variable	620
18.6.3	Pooled Poisson estimator	620
18.6.4	PA Poisson estimator	621
18.6.5	RE Poisson estimators	622
18.6.6	FE Poisson estimator	624
18.6.7	Panel Poisson estimators comparison	626
18.6.8	Negative binomial estimators	627
18.7	Stata resources	628
18.8	Exercises	629
A	Programming in Stata	631
A.1	Stata matrix commands	631
A.1.1	Stata matrix overview	631
A.1.2	Stata matrix input and output	631
	Matrix input by hand	631
	Matrix input from Stata estimation results	632
A.1.3	Stata matrix subscripts and combining matrices	633
A.1.4	Matrix operators	634
A.1.5	Matrix functions	634
A.1.6	Matrix accumulation commands	635
A.1.7	OLS using Stata matrix commands	636
A.2	Programs	637
A.2.1	Simple programs (no arguments or access to results)	637
A.2.2	Modifying a program	638
A.2.3	Programs with positional arguments	638
A.2.4	Temporary variables	639

A.2.5	Programs with named positional arguments	639
A.2.6	Storing and retrieving program results	640
A.2.7	Programs with arguments using standard Stata syntax	641
A.2.8	Ado-files	642
A.3	Program debugging	643
A.3.1	Some simple tips	644
A.3.2	Error messages and return code	644
A.3.3	Trace	645
B	Mata	647
B.1	How to run Mata	647
B.1.1	Mata commands in Mata	647
B.1.2	Mata commands in Stata	648
B.1.3	Stata commands in Mata	648
B.1.4	Interactive versus batch use	648
B.1.5	Mata help	648
B.2	Mata matrix commands	649
B.2.1	Mata matrix input	649
	Matrix input by hand	649
	Identity matrices, unit vectors, and matrices of constants . .	650
	Matrix input from Stata data	651
	Matrix input from Stata matrix	651
	Stata interface functions	652
B.2.2	Mata matrix operators	652
	Element-by-element operators	652
B.2.3	Mata functions	653
	Scalar and matrix functions	653
	Matrix inversion	654
B.2.4	Mata cross products	655
B.2.5	Mata matrix subscripts and combining matrices	655

B.2.6	Transferring Mata data and matrices to Stata	657
	Creating Stata matrices from Mata matrices	657
	Creating Stata data from a Mata vector	657
B.3	Programming in Mata	658
B.3.1	Declarations	658
B.3.2	Mata program	658
B.3.3	Mata program with results output to Stata	659
B.3.4	Stata program that calls a Mata program	659
B.3.5	Using Mata in ado-files	660
	Glossary of abbreviations	661
	References	665
	Author index	673
	Subject index	677