

An Introduction to Modern Econometrics Using Stata

CHRISTOPHER F. BAUM
Department of Economics
Boston College



A Stata Press Publication
StataCorp LP
College Station, Texas

Contents

Illustrations	xv
Preface	xvii
Notation and typography	xix
1 Introduction	1
1.1 An overview of Stata's distinctive features	1
1.2 Installing the necessary software	4
1.3 Installing the support materials	5
2 Working with economic and financial data in Stata	7
2.1 The basics	7
2.1.1 The use command	7
2.1.2 Variable types	8
2.1.3 <code>_n</code> and <code>_N</code>	9
2.1.4 <code>generate</code> and <code>replace</code>	10
2.1.5 <code>sort</code> and <code>gsort</code>	10
2.1.6 <code>if exp</code> and <code>in range</code>	11
2.1.7 Using <code>if exp</code> with indicator variables	13
2.1.8 Using <code>if exp</code> versus <code>by varlist: with statistical commands</code> . . .	15
2.1.9 Labels and notes	17
2.1.10 The <code>varlist</code>	20
2.1.11 <code>drop</code> and <code>keep</code>	20
2.1.12 <code>rename</code> and <code>renvars</code>	21
2.1.13 The <code>save</code> command	21
2.1.14 <code>insheet</code> and <code>infile</code>	21

2.2	Common data transformations	22
2.2.1	The cond() function	22
2.2.2	Recoding discrete and continuous variables	23
2.2.3	Handling missing data	24
	mvdecode and mvencode	25
2.2.4	String-to-numeric conversion and vice versa	26
2.2.5	Handling dates	27
2.2.6	Some useful functions for generate or replace	29
2.2.7	The egen command	30
	Official egen functions	30
	egen functions from the user community	31
2.2.8	Computation for by-groups	33
2.2.9	Local macros	36
2.2.10	Looping over variables: forvalues and foreach	37
2.2.11	Scalars and matrices	39
2.2.12	Command syntax and return values	39
3	Organizing and handling economic data	43
3.1	Cross-sectional data and identifier variables	43
3.2	Time-series data	44
	3.2.1 Time-series operators	45
3.3	Pooled cross-sectional time-series data	45
3.4	Panel data	46
	3.4.1 Operating on panel data	47
3.5	Tools for manipulating panel data	49
	3.5.1 Unbalanced panels and data screening	50
	3.5.2 Other transforms of panel data	53
	3.5.3 Moving-window summary statistics and correlations	53
3.6	Combining cross-sectional and time-series datasets	55
3.7	Creating long-format datasets with append	56
	3.7.1 Using merge to add aggregate characteristics	57

3.7.2	The dangers of many-to-many merges	58
3.8	The reshape command	58
3.8.1	The xpose command	62
3.9	Using Stata for reproducible research	62
3.9.1	Using do-files	62
3.9.2	Data validation: assert and duplicates	63
4	Linear regression	69
4.1	Introduction	69
4.2	Computing linear regression estimates	70
4.2.1	Regression as a method-of-moments estimator	72
4.2.2	The sampling distribution of regression estimates	73
4.2.3	Efficiency of the regression estimator	74
4.2.4	Numerical identification of the regression estimates	75
4.3	Interpreting regression estimates	75
4.3.1	Research project: A study of single-family housing prices	76
4.3.2	The ANOVA table: ANOVA F and R-squared	77
4.3.3	Adjusted R-squared	78
4.3.4	The coefficient estimates and beta coefficients	80
4.3.5	Regression without a constant term	81
4.3.6	Recovering estimation results	82
4.3.7	Detecting collinearity in regression	84
4.4	Presenting regression estimates	87
4.4.1	Presenting summary statistics and correlations	90
4.5	Hypothesis tests, linear restrictions, and constrained least squares	91
4.5.1	Wald tests with test	94
4.5.2	Wald tests involving linear combinations of parameters	96
4.5.3	Joint hypothesis tests	98
4.5.4	Testing nonlinear restrictions and forming nonlinear combinations	99
4.5.5	Testing competing (nonnested) models	100

4.6	Computing residuals and predicted values	102
4.6.1	Computing interval predictions	103
4.7	Computing marginal effects	107
4.A	Appendix: Regression as a least-squares estimator	112
4.B	Appendix: The large-sample VCE for linear regression	113
5	Specifying the functional form	115
5.1	Introduction	115
5.2	Specification error	115
5.2.1	Omitting relevant variables from the model	116
	Specifying dynamics in time-series regression models	117
5.2.2	Graphically analyzing regression data	117
5.2.3	Added-variable plots	119
5.2.4	Including irrelevant variables in the model	121
5.2.5	The asymmetry of specification error	121
5.2.6	Misspecification of the functional form	122
5.2.7	Ramsey's RESET	122
5.2.8	Specification plots	124
5.2.9	Specification and interaction terms	125
5.2.10	Outlier statistics and measures of leverage	126
	The DFITS statistic	128
	The DFBETA statistic	130
5.3	Endogeneity and measurement error	132
6	Regression with non-i.i.d. errors	133
6.1	The generalized linear regression model	134
6.1.1	Types of deviations from i.i.d. errors	134
6.1.2	The robust estimator of the VCE	136
6.1.3	The cluster estimator of the VCE	138
6.1.4	The Newey–West estimator of the VCE	139
6.1.5	The generalized least-squares estimator	142
	The FGLS estimator	143

6.2	Heteroskedasticity in the error distribution	143
6.2.1	Heteroskedasticity related to scale	144
	Testing for heteroskedasticity related to scale	145
	FGLS estimation	147
6.2.2	Heteroskedasticity between groups of observations	149
	Testing for heteroskedasticity between groups of observations .	150
	FGLS estimation	151
6.2.3	Heteroskedasticity in grouped data	152
	FGLS estimation	153
6.3	Serial correlation in the error distribution	154
6.3.1	Testing for serial correlation	155
6.3.2	FGLS estimation with serial correlation	159
7	Regression with indicator variables	161
7.1	Testing for significance of a qualitative factor	161
7.1.1	Regression with one qualitative measure	162
7.1.2	Regression with two qualitative measures	165
	Interaction effects	167
7.2	Regression with qualitative and quantitative factors	168
	Testing for slope differences	170
7.3	Seasonal adjustment with indicator variables	174
7.4	Testing for structural stability and structural change	179
7.4.1	Constraints of continuity and differentiability	179
7.4.2	Structural change in a time-series model	183
8	Instrumental-variables estimators	185
8.1	Introduction	185
8.2	Endogeneity in economic relationships	185
8.3	2SLS	188
8.4	The ivreg command	189
8.5	Identification and tests of overidentifying restrictions	190
8.6	Computing IV estimates	192

8.7	ivreg2 and GMM estimation	194
8.7.1	The GMM estimator	195
8.7.2	GMM in a homoskedastic context	196
8.7.3	GMM and heteroskedasticity-consistent standard errors	197
8.7.4	GMM and clustering	198
8.7.5	GMM and HAC standard errors	199
8.8	Testing overidentifying restrictions in GMM	200
8.8.1	Testing a subset of the overidentifying restrictions in GMM	201
8.9	Testing for heteroskedasticity in the IV context	205
8.10	Testing the relevance of instruments	207
8.11	Durbin–Wu–Hausman tests for endogeneity in IV estimation	211
8.A	Appendix: Omitted-variables bias	216
8.B	Appendix: Measurement error	216
8.B.1	Solving errors-in-variables problems	218
9	Panel-data models	219
9.1	FE and RE models	220
9.1.1	One-way FE	221
9.1.2	Time effects and two-way FE	224
9.1.3	The between estimator	226
9.1.4	One-way RE	227
9.1.5	Testing the appropriateness of RE	230
9.1.6	Prediction from one-way FE and RE	231
9.2	IV models for panel data	232
9.3	Dynamic panel-data models	232
9.4	Seemingly unrelated regression models	236
9.4.1	SUR with identical regressors	241
9.5	Moving-window regression estimates	242
10	Models of discrete and limited dependent variables	247
10.1	Binomial logit and probit models	247
10.1.1	The latent-variable approach	248

10.1.2	Marginal effects and predictions	250
	Binomial probit	251
	Binomial logit and grouped logit	253
10.1.3	Evaluating specification and goodness of fit	254
10.2	Ordered logit and probit models	256
10.3	Truncated regression and tobit models	259
10.3.1	Truncation	259
10.3.2	Censoring	262
10.4	Incidental truncation and sample-selection models	266
10.5	Bivariate probit and probit with selection	271
10.5.1	Binomial probit with selection	272
A	Getting the data into Stata	277
A.1	Inputting data from ASCII text files and spreadsheets	277
A.1.1	Handling text files	278
	Free format versus fixed format	278
	The insheet command	280
A.1.2	Accessing data stored in spreadsheets	281
A.1.3	Fixed-format data files	281
A.2	Importing data from other package formats	286
B	The basics of Stata programming	289
B.1	Local and global macros	290
B.1.1	Global macros	293
B.1.2	Extended macro functions and list functions	293
B.2	Scalars	294
B.3	Loop constructs	295
B.3.1	foreach	297
B.4	Matrices	299
B.5	return and ereturn	301
B.5.1	ereturn list	305

B.6	The program and syntax statements	307
B.7	Using Mata functions in Stata programs	313
	References	321
	Author index	329
	Subject index	333