

# Microeconometrics Using Stata

## Volume I: Cross-Sectional and Panel Regression Methods

Second Edition

A. COLIN CAMERON  
*Department of Economics*  
*University of California, Davis, CA*  
*and*  
*School of Economics*  
*University of Sydney, Sydney, Australia*

PRAVIN K. TRIVEDI  
*School of Economics*  
*University of Queensland, Brisbane, Australia*  
*and*  
*Department of Economics*  
*Indiana University, Bloomington, IN*



**Stata**® *Press*

A Stata Press Publication  
StataCorp LLC  
College Station, Texas

# Contents

	List of tables	xiii
	List of figures	xv
	Preface to the Second Edition	xvii
	Preface to the First Edition	xix
<b>1</b>	<b>Stata basics</b>	<b>1</b>
	1.1 Interactive use . . . . .	1
	1.2 Documentation . . . . .	2
	1.3 Command syntax and operators . . . . .	5
	1.4 Do-files and log files . . . . .	14
	1.5 Scalars and matrices . . . . .	19
	1.6 Using results from Stata commands . . . . .	20
	1.7 Global and local macros . . . . .	23
	1.8 Looping commands . . . . .	26
	1.9 Mata and Python in Stata . . . . .	29
	1.10 Some useful commands . . . . .	29
	1.11 Template do-file . . . . .	30
	1.12 Community-contributed commands . . . . .	30
	1.13 Additional resources . . . . .	31
	1.14 Exercises . . . . .	31
<b>2</b>	<b>Data management and graphics</b>	<b>33</b>
	2.1 Introduction . . . . .	33
	2.2 Types of data . . . . .	33
	2.3 Inputting data . . . . .	36
	2.4 Data management . . . . .	43

2.5	Manipulating datasets . . . . .	60
2.6	Graphical display of data . . . . .	67
2.7	Additional resources . . . . .	83
2.8	Exercises . . . . .	83
<b>3</b>	<b>Linear regression basics</b>	<b>85</b>
3.1	Introduction . . . . .	85
3.2	Data and data summary . . . . .	85
3.3	Transformation of data before regression . . . . .	94
3.4	Linear regression . . . . .	96
3.5	Basic regression analysis . . . . .	102
3.6	Specification analysis . . . . .	123
3.7	Specification tests . . . . .	132
3.8	Sampling weights . . . . .	140
3.9	OLS using Mata . . . . .	145
3.10	Additional resources . . . . .	147
3.11	Exercises . . . . .	147
<b>4</b>	<b>Linear regression extensions</b>	<b>149</b>
4.1	Introduction . . . . .	149
4.2	In-sample prediction . . . . .	149
4.3	Out-of-sample prediction . . . . .	157
4.4	Predictive margins . . . . .	161
4.5	Marginal effects . . . . .	175
4.6	Regression decomposition analysis . . . . .	186
4.7	Shapley decomposition of relative regressor importance . . . . .	193
4.8	Difference-in-differences estimators . . . . .	195
4.9	Additional resources . . . . .	204
4.10	Exercises . . . . .	204
<b>5</b>	<b>Simulation</b>	<b>207</b>
5.1	Introduction . . . . .	207
5.2	Pseudorandom-number generators . . . . .	208

5.3	Distribution of the sample mean . . . . .	214
5.4	Pseudorandom-number generators: Further details . . . . .	220
5.5	Computing integrals . . . . .	227
5.6	Simulation for regression: Introduction . . . . .	232
5.7	Additional resources . . . . .	242
5.8	Exercises . . . . .	242
<b>6</b>	<b>Linear regression with correlated errors</b>	<b>245</b>
6.1	Introduction . . . . .	245
6.2	Generalized least-squares and FGLS regression . . . . .	246
6.3	Modeling heteroskedastic data . . . . .	250
6.4	OLS for clustered data . . . . .	256
6.5	FGLS estimators for clustered data . . . . .	265
6.6	Fixed-effects estimator for clustered data . . . . .	269
6.7	Linear mixed models for clustered data . . . . .	277
6.8	Systems of linear regressions . . . . .	286
6.9	Survey data: Weighting, clustering, and stratification . . . . .	295
6.10	Additional resources . . . . .	301
6.11	Exercises . . . . .	302
<b>7</b>	<b>Linear instrumental-variables regression</b>	<b>305</b>
7.1	Introduction . . . . .	305
7.2	Simultaneous equations model . . . . .	306
7.3	Instrumental-variables regression . . . . .	310
7.4	Instrumental-variables example . . . . .	316
7.5	Weak instruments . . . . .	330
7.6	Diagnostics and tests for weak instruments . . . . .	339
7.7	Inference with weak instruments . . . . .	353
7.8	Finite sample inference with weak instruments . . . . .	362
7.9	Other estimators . . . . .	363
7.10	Three-stage least-squares systems estimation . . . . .	367

7.11	Additional resources . . . . .	368
7.12	Exercises . . . . .	369
<b>8</b>	<b>Linear panel-data models: Basics</b>	<b>373</b>
8.1	Introduction . . . . .	373
8.2	Panel-data methods overview . . . . .	373
8.3	Summary of panel data . . . . .	379
8.4	Pooled or population-averaged estimators . . . . .	394
8.5	Fixed-effects or within estimator . . . . .	397
8.6	Between estimator . . . . .	401
8.7	Random-effects estimator . . . . .	402
8.8	Comparison of estimators . . . . .	406
8.9	First-difference estimator . . . . .	412
8.10	Panel-data management . . . . .	414
8.11	Additional resources . . . . .	418
8.12	Exercises . . . . .	419
<b>9</b>	<b>Linear panel-data models: Extensions</b>	<b>421</b>
9.1	Introduction . . . . .	421
9.2	Panel instrumental-variables estimation . . . . .	421
9.3	Hausman–Taylor estimator . . . . .	425
9.4	Arellano–Bond estimator . . . . .	428
9.5	Long panels . . . . .	445
9.6	Additional resources . . . . .	456
9.7	Exercises . . . . .	456
<b>10</b>	<b>Introduction to nonlinear regression</b>	<b>459</b>
10.1	Introduction . . . . .	459
10.2	Binary outcome models . . . . .	459
10.3	Probit model . . . . .	462
10.4	MEs and coefficient interpretation . . . . .	466
10.5	Logit model . . . . .	472
10.6	Nonlinear least squares . . . . .	474

10.7	Other nonlinear estimators . . . . .	476
10.8	Additional resources . . . . .	477
10.9	Exercises . . . . .	477
<b>11</b>	<b>Tests of hypotheses and model specification</b>	<b>479</b>
11.1	Introduction . . . . .	479
11.2	Critical values and p-values . . . . .	480
11.3	Wald tests and confidence intervals . . . . .	485
11.4	Likelihood-ratio tests . . . . .	498
11.5	Lagrange multiplier test (or score test) . . . . .	502
11.6	Multiple testing . . . . .	505
11.7	Test size and power . . . . .	512
11.8	The power onemean command for multiple regression . . . . .	519
11.9	Specification tests . . . . .	529
11.10	Permutation tests and randomization tests . . . . .	532
11.11	Additional resources . . . . .	534
11.12	Exercises . . . . .	534
<b>12</b>	<b>Bootstrap methods</b>	<b>537</b>
12.1	Introduction . . . . .	537
12.2	Bootstrap methods . . . . .	537
12.3	Bootstrap pairs using the vce(bootstrap) option . . . . .	539
12.4	Bootstrap pairs using the bootstrap command . . . . .	547
12.5	Percentile-t bootstraps with asymptotic refinement . . . . .	555
12.6	Wild bootstrap with asymptotic refinement . . . . .	560
12.7	Bootstrap pairs using bsample and simulate . . . . .	569
12.8	Alternative resampling schemes . . . . .	570
12.9	The jackknife . . . . .	575
12.10	Additional resources . . . . .	576
12.11	Exercises . . . . .	577

<b>13</b>	<b>Nonlinear regression methods</b>	<b>579</b>
13.1	Introduction . . . . .	579
13.2	Nonlinear example: Doctor visits . . . . .	580
13.3	Nonlinear regression methods . . . . .	582
13.4	Different estimates of the VCE . . . . .	597
13.5	Prediction . . . . .	604
13.6	Predictive margins . . . . .	609
13.7	Marginal effects . . . . .	612
13.8	Model diagnostics . . . . .	629
13.9	Clustered data . . . . .	632
13.10	Additional resources . . . . .	640
13.11	Exercises . . . . .	640
<b>14</b>	<b>Flexible regression: Finite mixtures and nonparametric</b>	<b>643</b>
14.1	Introduction . . . . .	643
14.2	Models based on finite mixtures . . . . .	644
14.3	FMM example: Earnings of doctors . . . . .	650
14.4	Global polynomials . . . . .	665
14.5	Regression splines . . . . .	668
14.6	Nonparametric regression . . . . .	675
14.7	Partially parametric regression . . . . .	680
14.8	Additional resources . . . . .	681
14.9	Exercises . . . . .	681
<b>15</b>	<b>Quantile regression</b>	<b>683</b>
15.1	Introduction . . . . .	683
15.2	Conditional quantile regression . . . . .	684
15.3	CQR for medical expenditures data . . . . .	688
15.4	CQR for generated heteroskedastic data . . . . .	699
15.5	Quantile treatment effects for a binary treatment . . . . .	703
15.6	Additional resources . . . . .	706
15.7	Exercises . . . . .	707

<b>A</b>	<b>Programming in Stata</b>	<b>709</b>
A.1	Stata matrix commands . . . . .	709
A.2	Programs . . . . .	716
A.3	Program debugging . . . . .	722
A.4	Additional resources . . . . .	725
<b>B</b>	<b>Mata</b>	<b>727</b>
B.1	How to run Mata . . . . .	727
B.2	Mata matrix commands . . . . .	729
B.3	Programming in Mata . . . . .	738
B.4	Additional resources . . . . .	740
<b>C</b>	<b>Optimization in Mata</b>	<b>741</b>
C.1	Mata moptimize() function . . . . .	741
C.2	Mata optimize() function . . . . .	751
C.3	Additional resources . . . . .	754
	<b>Glossary of abbreviations</b>	<b>755</b>
	<b>References</b>	<b>761</b>
	<b>Author index</b>	<b>777</b>
	<b>Subject index</b>	<b>783</b>



# Microeconometrics Using Stata

## Volume II: Nonlinear Models and Causal Inference Methods

Second Edition

A. COLIN CAMERON  
*Department of Economics*  
*University of California, Davis, CA*  
and  
*School of Economics*  
*University of Sydney, Sydney, Australia*

PRAVIN K. TRIVEDI  
*School of Economics*  
*University of Queensland, Brisbane, Australia*  
and  
*Department of Economics*  
*Indiana University, Bloomington, IN*



**Stata**® *Press*

A Stata Press Publication  
StataCorp LLC  
College Station, Texas

# Contents

	List of tables	xiii
	List of figures	xv
<b>16</b>	<b>Nonlinear optimization methods</b>	<b>819</b>
16.1	Introduction . . . . .	819
16.2	Newton–Raphson method . . . . .	819
16.3	Gradient methods . . . . .	824
16.4	Overview of ml, moptimize(), and optimize() . . . . .	829
16.5	The ml command: lf method . . . . .	831
16.6	Checking the program . . . . .	837
16.7	The ml command: lf0–lf2, d0–d2, and gf0 methods . . . . .	844
16.8	Nonlinear instrumental-variables (GMM) example . . . . .	851
16.9	Additional resources . . . . .	854
16.10	Exercises . . . . .	854
<b>17</b>	<b>Binary outcome models</b>	<b>857</b>
17.1	Introduction . . . . .	857
17.2	Some parametric models . . . . .	858
17.3	Estimation . . . . .	860
17.4	Example . . . . .	862
17.5	Goodness of fit and prediction . . . . .	869
17.6	Marginal effects . . . . .	877
17.7	Clustered data . . . . .	880
17.8	Additional models . . . . .	881
17.9	Endogenous regressors . . . . .	887
17.10	Grouped and fractional data . . . . .	895

17.11	Additional resources . . . . .	898
17.12	Exercises . . . . .	898
<b>18</b>	<b>Multinomial models</b>	<b>901</b>
18.1	Introduction . . . . .	901
18.2	Multinomial models overview . . . . .	901
18.3	Multinomial example: Choice of fishing mode . . . . .	905
18.4	Multinomial logit model . . . . .	908
18.5	Alternative-specific conditional logit model . . . . .	914
18.6	Nested logit model . . . . .	922
18.7	Multinomial probit model . . . . .	929
18.8	Alternative-specific random-parameters logit . . . . .	934
18.9	Ordered outcome models . . . . .	938
18.10	Clustered data . . . . .	942
18.11	Multivariate outcomes . . . . .	943
18.12	Additional resources . . . . .	946
18.13	Exercises . . . . .	946
<b>19</b>	<b>Tobit and selection models</b>	<b>949</b>
19.1	Introduction . . . . .	949
19.2	Tobit model . . . . .	950
19.3	Tobit model example . . . . .	953
19.4	Tobit for lognormal data . . . . .	961
19.5	Two-part model in logs . . . . .	970
19.6	Selection models . . . . .	974
19.7	Nonnormal models of selection . . . . .	982
19.8	Prediction from models with outcome in logs . . . . .	986
19.9	Endogenous regressors . . . . .	989
19.10	Missing data . . . . .	991
19.11	Panel attrition . . . . .	995
19.12	Additional resources . . . . .	1019
19.13	Exercises . . . . .	1019

<b>20</b>	<b>Count-data models</b>	<b>1021</b>
20.1	Introduction . . . . .	1021
20.2	Modeling strategies for count data . . . . .	1022
20.3	Poisson and negative binomial models . . . . .	1026
20.4	Hurdle model . . . . .	1044
20.5	Finite-mixture models . . . . .	1050
20.6	Zero-inflated models . . . . .	1069
20.7	Endogenous regressors . . . . .	1079
20.8	Clustered data . . . . .	1089
20.9	Quantile regression for count data . . . . .	1090
20.10	Additional resources . . . . .	1096
20.11	Exercises . . . . .	1096
<b>21</b>	<b>Survival analysis for duration data</b>	<b>1099</b>
21.1	Introduction . . . . .	1099
21.2	Data and data summary . . . . .	1100
21.3	Survivor and hazard functions . . . . .	1104
21.4	Semiparametric regression model . . . . .	1109
21.5	Fully parametric regression models . . . . .	1118
21.6	Multiple-records data . . . . .	1129
21.7	Discrete-time hazards logit model . . . . .	1132
21.8	Time-varying regressors . . . . .	1135
21.9	Clustered data . . . . .	1136
21.10	Additional resources . . . . .	1137
21.11	Exercises . . . . .	1137
<b>22</b>	<b>Nonlinear panel models</b>	<b>1139</b>
22.1	Introduction . . . . .	1139
22.2	Nonlinear panel-data overview . . . . .	1139
22.3	Nonlinear panel-data example . . . . .	1145
22.4	Binary outcome and ordered outcome models . . . . .	1148
22.5	Tobit and interval-data models . . . . .	1167

22.6	Count-data models . . . . .	1172
22.7	Panel quantile regression . . . . .	1184
22.8	Endogenous regressors in nonlinear panel models . . . . .	1187
22.9	Additional resources . . . . .	1188
22.10	Exercises . . . . .	1188
<b>23</b>	<b>Parametric models for heterogeneity and endogeneity</b>	<b>1191</b>
23.1	Introduction . . . . .	1191
23.2	Finite mixtures and unobserved heterogeneity . . . . .	1192
23.3	Empirical examples of FMMs . . . . .	1195
23.4	Nonlinear mixed-effects models . . . . .	1224
23.5	Linear structural equation models . . . . .	1231
23.6	Generalized structural equation models . . . . .	1251
23.7	ERM commands for endogeneity and selection . . . . .	1261
23.8	Additional resources . . . . .	1266
23.9	Exercises . . . . .	1266
<b>24</b>	<b>Randomized control trials and exogenous treatment effects</b>	<b>1269</b>
24.1	Introduction . . . . .	1269
24.2	Potential outcomes . . . . .	1271
24.3	Randomized control trials . . . . .	1272
24.4	Regression in an RCT . . . . .	1282
24.5	Treatment evaluation with exogenous treatment . . . . .	1290
24.6	Treatment evaluation methods and estimators . . . . .	1292
24.7	Stata commands for treatment evaluation . . . . .	1302
24.8	Oregon Health Insurance Experiment example . . . . .	1305
24.9	Treatment-effect estimates using the OHIE data . . . . .	1312
24.10	Multilevel treatment effects . . . . .	1323
24.11	Conditional quantile TEs . . . . .	1332
24.12	Additional resources . . . . .	1334
24.13	Exercises . . . . .	1335

<b>25</b>	<b>Endogenous treatment effects</b>	<b>1337</b>
25.1	Introduction . . . . .	1337
25.2	Parametric methods for endogenous treatment . . . . .	1338
25.3	ERM commands for endogenous treatment . . . . .	1341
25.4	ET commands for binary endogenous treatment . . . . .	1348
25.5	The LATE estimator for heterogeneous effects . . . . .	1356
25.6	Difference-in-differences and synthetic control . . . . .	1363
25.7	Regression discontinuity design . . . . .	1369
25.8	Conditional quantile regression with endogenous regressors . . . . .	1388
25.9	Unconditional quantiles . . . . .	1394
25.10	Additional resources . . . . .	1401
25.11	Exercises . . . . .	1402
<b>26</b>	<b>Spatial regression</b>	<b>1405</b>
26.1	Introduction . . . . .	1405
26.2	Overview of spatial regression models . . . . .	1406
26.3	Geospatial data . . . . .	1407
26.4	The spatial weighting matrix . . . . .	1411
26.5	OLS regression and test for spatial correlation . . . . .	1413
26.6	Spatial dependence in the error . . . . .	1414
26.7	Spatial autocorrelation regression models . . . . .	1417
26.8	Spatial instrumental variables . . . . .	1427
26.9	Spatial panel-data models . . . . .	1428
26.10	Additional resources . . . . .	1429
26.11	Exercises . . . . .	1430
<b>27</b>	<b>Semiparametric regression</b>	<b>1433</b>
27.1	Introduction . . . . .	1433
27.2	Kernel regression . . . . .	1434
27.3	Series regression . . . . .	1438
27.4	Nonparametric single regressor example . . . . .	1440
27.5	Nonparametric multiple regressor example . . . . .	1450

27.6	Partial linear model . . . . .	1453
27.7	Single-index model . . . . .	1456
27.8	Generalized additive models . . . . .	1458
27.9	Additional resources . . . . .	1461
27.10	Exercises . . . . .	1462
<b>28</b>	<b>Machine learning for prediction and inference</b>	<b>1465</b>
28.1	Introduction . . . . .	1465
28.2	Measuring the predictive ability of a model . . . . .	1466
28.3	Shrinkage estimators . . . . .	1477
28.4	Prediction using lasso, ridge, and elasticnet . . . . .	1482
28.5	Dimension reduction . . . . .	1493
28.6	Machine learning methods for prediction . . . . .	1496
28.7	Prediction application . . . . .	1501
28.8	Machine learning for inference in partial linear model . . . . .	1507
28.9	Machine learning for inference in other models . . . . .	1516
28.10	Additional resources . . . . .	1523
28.11	Exercises . . . . .	1524
<b>29</b>	<b>Bayesian methods: Basics</b>	<b>1527</b>
29.1	Introduction . . . . .	1527
29.2	Bayesian introductory example . . . . .	1528
29.3	Bayesian methods overview . . . . .	1532
29.4	An i.i.d. example . . . . .	1538
29.5	Linear regression . . . . .	1549
29.6	A linear regression example . . . . .	1552
29.7	Modifying the MH algorithm . . . . .	1560
29.8	RE model . . . . .	1562
29.9	Bayesian model selection . . . . .	1567
29.10	Bayesian prediction . . . . .	1569
29.11	Probit example . . . . .	1572

29.12	Additional resources . . . . .	1576
29.13	Exercises . . . . .	1576
<b>30</b>	<b>Bayesian methods: Markov chain Monte Carlo algorithms</b>	<b>1579</b>
30.1	Introduction . . . . .	1579
30.2	User-provided log likelihood . . . . .	1579
30.3	MH algorithm in Mata . . . . .	1584
30.4	Data augmentation and the Gibbs sampler in Mata . . . . .	1589
30.5	Multiple imputation . . . . .	1595
30.6	Multiple-imputation example . . . . .	1599
30.7	Additional resources . . . . .	1608
30.8	Exercises . . . . .	1608
	<b>Glossary of abbreviations</b>	<b>1611</b>
	<b>References</b>	<b>1617</b>
	<b>Author index</b>	<b>1635</b>
	<b>Subject index</b>	<b>1641</b>