

Regression Models for Categorical Dependent Variables Using Stata

Second Edition

J. SCOTT LONG
Department of Sociology
Indiana University
Bloomington, Indiana

JEREMY FREESE
Department of Sociology
University of Wisconsin-Madison
Madison, Wisconsin



A Stata Press Publication
StataCorp LP
College Station, Texas

Contents

Preface	xxix
I General Information	1
1 Introduction	3
1.1 What is this book about?	3
1.2 Which models are considered?	4
1.3 Whom is this book for?	5
1.4 How is the book organized?	5
1.5 What software do you need?	7
1.5.1 Updating Stata 9	8
1.5.2 Installing SPost	9
Installing SPost using search	9
Installing SPost using net install	11
1.5.3 What if commands do not work?	11
1.5.4 Uninstalling SPost	12
1.5.5 Using spex to load data and run examples	12
1.5.6 More files available on the web site	13
1.6 Where can I learn more about the models?	13
2 Introduction to Stata	15
2.1 The Stata interface	16
Changing the scrollbar buffer size	18
Changing the display of variable names in the Variables window	19
2.2 Abbreviations	19

2.3	How to get help	20
2.3.1	Online help	20
2.3.2	Manuals	21
2.3.3	Other resources	21
2.4	The working directory	21
2.5	Stata file types	22
2.6	Saving output to log files	22
	Options	23
2.6.1	Closing a log file	23
2.6.2	Viewing a log file	24
2.6.3	Converting from SMCL to plain text or PostScript	24
2.7	Using and saving datasets	24
2.7.1	Data in Stata format	24
2.7.2	Data in other formats	25
2.7.3	Entering data by hand	26
2.8	Size limitations on datasets*	26
2.9	Do-files	26
2.9.1	Adding comments	28
2.9.2	Long lines	29
2.9.3	Stopping a do-file while it is running	29
2.9.4	Creating do-files	29
	Using Stata's Do-file Editor	29
	Using other editors to create do-files	30
2.9.5	Recommended structure for do-files	30
2.10	Using Stata for serious data analysis	31
2.11	Syntax of Stata commands	33
2.11.1	Commands	34
2.11.2	Variable lists	35
2.11.3	if and in qualifiers	36
	Examples of if qualifier	37

2.11.4	Options	37
2.12	Managing data	37
2.12.1	Looking at your data	37
2.12.2	Getting information about variables	38
2.12.3	Missing values	41
2.12.4	Selecting observations	41
2.12.5	Selecting variables	42
2.13	Creating new variables	42
2.13.1	generate command	42
2.13.2	replace command	44
2.13.3	recode command	44
2.13.4	Common transformations for RHS variables	45
	Breaking a categorical variable into a set of binary variables	45
	More examples of creating binary variables	47
	Nonlinear transformations	48
	Interaction terms	49
2.14	Labeling variables and values	49
2.14.1	Variable labels	49
2.14.2	Value labels	50
2.14.3	notes command	52
2.15	Global and local macros	52
2.16	Graphics	54
2.16.1	graph command	56
2.16.2	Displaying previously drawn graphs	63
2.16.3	Printing graphs	63
2.16.4	Combining graphs	63
2.17	A brief tutorial	65
	A batch version	72

3	Estimation, testing, fit, and interpretation	75
3.1	Estimation	76
3.1.1	Stata's output for ML estimation	76
3.1.2	ML and sample size	77
3.1.3	Problems in obtaining ML estimates	77
3.1.4	Syntax of estimation commands	78
	Variable lists	78
	Specifying the estimation sample	79
	Weights	84
	Options	85
3.1.5	Reading the output	87
	Header	87
	Estimates and standard errors	88
	Confidence intervals	88
3.1.6	Storing estimation results	89
3.1.7	Reformatting output with estimates table	89
3.1.8	Reformatting output with estout	91
3.1.9	Alternative output with listcoef	94
	Options for types of coefficients	95
	Options for mlogit, mprobit, and slogit	95
	Other options	96
	Standardized coefficients	96
	Factor and percent change	98
3.2	Postestimation analysis	99
3.3	Testing	99
3.3.1	Wald tests	99
	The accumulate option	101
3.3.2	LR tests	101
	Avoiding invalid LR tests	102
3.4	estat command	103

3.5	Measures of fit	104
	Syntax of fitstat	104
	Options	105
	Models and measures	105
	Example of fitstat	107
	Methods and formulas for fitstat	108
3.6	Interpretation	113
3.6.1	Approaches to interpretation	116
3.6.2	Predictions using predict	116
3.6.3	Overview of prvalue, prchange, prtab, and prgen	118
	Specifying the levels of variables	118
	Options controlling output	119
3.6.4	Syntax for prvalue	120
	Options	120
	Options for confidence intervals	120
	Options used for bootstrapped confidence intervals	121
3.6.5	Syntax for prchange	122
	Options	122
3.6.6	Syntax for prtab	122
	Options	123
3.6.7	Syntax for prgen	123
	Options	123
	Options for confidence intervals and marginals	124
	Variables generated	124
3.6.8	Computing marginal effects using mfx	125
3.7	Confidence intervals for predictions	126
3.8	Next steps	128

II Models for Specific Kinds of Outcomes	129
4 Models for binary outcomes	131
4.1 The statistical model	132
4.1.1 A latent-variable model	132
4.1.2 A nonlinear probability model	135
4.2 Estimation using logit and probit	136
Variable lists	136
Specifying the estimation sample	136
Weights	136
Options	137
Example	137
4.2.1 Observations predicted perfectly	140
4.3 Hypothesis testing with test and lrtest	140
4.3.1 Testing individual coefficients	140
One- and two-tailed tests	141
Testing single coefficients using test	142
Testing single coefficients using lrtest	142
4.3.2 Testing multiple coefficients	143
Testing multiple coefficients using test	143
Testing multiple coefficients using lrtest	144
4.3.3 Comparing LR and Wald tests	144
4.4 Residuals and influence using predict	145
4.4.1 Residuals	147
Example	147
4.4.2 Influential cases	151
4.4.3 Least likely observations	152
Syntax	152
Options	152
Options controlling the list of values	153

4.5	Measuring fit	154
4.5.1	Scalar measures of fit using fitstat	154
4.5.2	Hosmer–Lemeshow statistic	155
4.6	Interpretation using predicted values	157
4.6.1	Predicted probabilities with predict	158
4.6.2	Individual predicted probabilities with prvalue	160
4.6.3	Tables of predicted probabilities with prtab	162
4.6.4	Graphing predicted probabilities with prgen	163
4.6.5	Plotting confidence intervals	166
4.6.6	Changes in predicted probabilities	168
	Marginal change	168
	Discrete change	170
4.7	Interpretation using odds ratios with listcoef	177
	Multiplicative coefficients	179
	Effect of the base probability	179
	Percent change in the odds	180
4.8	Other commands for binary outcomes	181
5	Models for ordinal outcomes	183
5.1	The statistical model	184
5.1.1	A latent-variable model	184
5.1.2	A nonlinear probability model	187
5.2	Estimation using ologit and oprobit	188
	Variable lists	188
	Specifying the estimation sample	188
	Weights	188
	Options	189
5.2.1	Example of attitudes toward working mothers	189
5.2.2	Predicting perfectly	192

5.3	Hypothesis testing with <code>test</code> and <code>lrtest</code>	193
5.3.1	Testing individual coefficients	193
5.3.2	Testing multiple coefficients	194
5.4	Scalar measures of fit using <code>fitstat</code>	195
5.5	Converting to a different parameterization*	196
5.6	The parallel regression assumption	197
5.7	Residuals and outliers using <code>predict</code>	200
5.8	Interpretation	202
5.8.1	Marginal change in y^*	203
5.8.2	Predicted probabilities	204
5.8.3	Predicted probabilities with <code>predict</code>	204
5.8.4	Individual predicted probabilities with <code>prvalue</code>	205
5.8.5	Tables of predicted probabilities with <code>prtab</code>	207
5.8.6	Graphing predicted probabilities with <code>prgen</code>	209
5.8.7	Changes in predicted probabilities	211
	Marginal change with <code>prchange</code>	212
	Marginal change with <code>mfx</code>	212
	Discrete change with <code>prchange</code>	213
	Confidence intervals for discrete changes	215
	Computing discrete change for a 10-year increase in age	216
5.8.8	Odds ratios using <code>listcoef</code>	217
5.9	Less common models for ordinal outcomes	220
5.9.1	The stereotype model	220
5.9.2	The generalized ordered logit model	220
5.9.3	The continuation ratio model	221
6	Models for nominal outcomes with case-specific data	223
6.1	The multinomial logit model	224
6.1.1	Formal statement of the model	227

6.2	Estimation using mlogit	228
	Variable lists	228
	Specifying the estimation sample	229
	Weights	229
	Options	229
6.2.1	Example of occupational attainment	230
6.2.2	Using different base categories	231
6.2.3	Predicting perfectly	234
6.3	Hypothesis testing of coefficients	234
6.3.1	mlogtest for tests of the MNLM	235
	Options	235
6.3.2	Testing the effects of the independent variables	236
	A likelihood-ratio test	236
	A Wald test	237
	Testing multiple independent variables	238
6.3.3	Tests for combining alternatives	239
	A Wald test for combining alternatives	239
	Using test [category]*	240
	An LR test for combining alternatives	241
	Using constraint with lrtest*	241
6.4	Independence of irrelevant alternatives	243
	Hausman test of IIA	244
	Small-Hsiao test of IIA	245
6.5	Measures of fit	246
6.6	Interpretation	246
6.6.1	Predicted probabilities	247
6.6.2	Predicted probabilities with predict	247
	Using predict to compare mlogit and ologit	248
6.6.3	Predicted probabilities and discrete change with prvalue	249
6.6.4	Tables of predicted probabilities with prtab	249

6.6.5	Graphing predicted probabilities with <code>prgen</code>	250
	Plotting probabilities for one outcome and two groups	251
	Graphing probabilities for all outcomes for one group	252
6.6.6	Changes in predicted probabilities	254
	Computing marginal and discrete change with <code>prchange</code>	255
	Marginal change with <code>mfx</code>	257
6.6.7	Plotting discrete changes with <code>prchange</code> and <code>mlogview</code>	257
6.6.8	Odds ratios using <code>listcoef</code> and <code>mlogview</code>	260
	Listing odds ratios with <code>listcoef</code>	261
	Plotting odds ratios	262
6.6.9	Using <code>mlogplot*</code>	267
6.6.10	Plotting estimates from matrices with <code>mlogplot*</code>	268
	Options for using matrices with <code>mlogplot</code>	269
	Global macros and matrices used by <code>mlogplot</code>	269
	Example	270
6.7	Multinomial probit model with IIA	272
6.8	Stereotype logistic regression	277
6.8.1	Formal statement of the one-dimensional SLM	279
6.8.2	Fitting the SLM with <code>slogit</code>	280
	Options	280
	Example	281
6.8.3	Interpretation using predicted probabilities	281
6.8.4	Interpretation using odds ratios	283
6.8.5	Distinguishability and the ϕ parameters	286
6.8.6	Ordinality in the one-dimensional SLM	288
	Higher-dimension SLM	291
7	Models for nominal outcomes with alternative-specific data	293
7.1	Alternative-specific data organization	294
7.1.1	Syntax for <code>case2alt</code>	296

7.2	The conditional logit model	297
7.2.1	Fitting the conditional logit model	298
	Example of the clogit model	298
7.2.2	Interpreting odds ratios from clogit	299
7.2.3	Interpreting probabilities from clogit	299
	Using predict	299
	Using asprvalue	300
7.2.4	Fitting the multinomial logit model using clogit	304
	Setting up the data with case2alt	304
	Fitting multinomial logit with clogit	306
7.2.5	Using clogit with case- and alternative-specific variables	307
	Example of a mixed model	308
	Interpretation of odds ratios using listcoef	308
	Interpretation of predicted probabilities using asprvalue	310
	Allowing the effects of alternative-specific variables to vary over the alternatives	312
7.3	Alternative-specific multinomial probit	313
7.3.1	The model	314
7.3.2	Informal explanation of estimation by simulation	315
7.3.3	Alternative-based data with uncorrelated errors	319
	Options	319
	Examples	320
7.3.4	Alternative-based data with correlated errors	322
7.4	The structural covariance matrix	325
7.4.1	Interpretation using probabilities	329
	Using predict	329
	Using asprvalue	330
7.4.2	Identification, discrete change, and marginal effects	332
7.4.3	Testing for IIA	336
7.4.4	Adding case-specific data	337

7.5	Rank-ordered logistic regression	339
7.5.1	Fitting the rank-ordered logit model	341
	Options	341
	Example of the rank-ordered logit model	342
7.5.2	Interpreting results from rologit	343
	Interpretation using odds ratios	343
	Interpretation using predicted probabilities	345
7.6	Conclusions	347
8	Models for count outcomes	349
8.1	The Poisson distribution	349
8.1.1	Fitting the Poisson distribution with the poisson command	350
8.1.2	Computing predicted probabilities with prcounts	352
	Syntax	352
	Options	352
	Variables generated	352
8.1.3	Comparing observed and predicted counts with prcounts	354
8.2	The Poisson regression model	356
8.2.1	Fitting the PRM with poisson	357
	Variable lists	357
	Specifying the estimation sample	358
	Weights	358
	Options	358
8.2.2	Example of fitting the PRM	358
8.2.3	Interpretation using the rate, μ	359
	Factor change in $E(y \mid x)$	359
	Percent change in $E(y \mid x)$	360
	Example of factor and percent change	360
	Marginal change in $E(y \mid x)$	361
	Example of marginal change using prchange	362

	Example of marginal change using mfx	362
	Discrete change in $E(y x)$	362
	Example of discrete change using prchange	363
	Example of discrete change with confidence intervals	364
8.2.4	Interpretation using predicted probabilities	365
	Example of predicted probabilities using prvalue	365
	Example of predicted probabilities using prgen	367
	Example of predicted probabilities using prcounts	368
8.2.5	Exposure time*	370
8.3	The negative binomial regression model	372
8.3.1	Fitting the NBRM with nbreg	374
	NB1 and NB2 variance functions	374
8.3.2	Example of fitting the NBRM	375
	Comparing the PRM and NBRM using estimates table	375
8.3.3	Testing for overdispersion	376
8.3.4	Interpretation using the rate μ	377
8.3.5	Interpretation using predicted probabilities	378
8.4	Models for truncated counts	381
8.4.1	Fitting zero-truncated models	383
8.4.2	Example of fitting zero-truncated models	383
8.4.3	Interpretation of parameters	384
8.4.4	Interpretation using predicted probabilities and rates	386
8.4.5	Computing predicted rates and probabilities in the estimation sample	387
8.5	The hurdle regression model*	387
8.5.1	In-sample predictions for the hurdle model	388
8.5.2	Predictions for user-specified values	391
8.6	Zero-inflated count models	394
8.6.1	Fitting zero-inflated models with zinb and zip	396
	Variable lists	397

- Options 397
- 8.6.2 Example of fitting the ZIP and ZINB models 397
- 8.6.3 Interpretation of coefficients 398
- 8.6.4 Interpretation of predicted probabilities 400
 - Predicted probabilities with prvalue 400
 - Confidence intervals with prvalue 401
 - Predicted probabilities with prgen 404
- 8.7 Comparisons among count models 405
 - 8.7.1 Comparing mean probabilities 405
 - 8.7.2 Tests to compare count models 407
 - LR tests of α 407
 - Vuong test of nonnested models 408
- 8.8 Using countfit to compare count models 409
- 9 More topics 415**
- 9.1 Ordinal and nominal independent variables 415
 - 9.1.1 Coding a categorical independent variable as a set of dummy variables 415
 - 9.1.2 Estimation and interpretation with categorical independent variables 417
 - 9.1.3 Tests with categorical independent variables 418
 - Testing the effect of membership in one category versus the reference category 418
 - Testing the effect of membership in two nonreference categories 419
 - Testing that a categorical independent variable has no effect . 420
 - Testing whether treating an ordinal variable as interval loses information 421
 - 9.1.4 Discrete change for categorical independent variables 422
 - Computing discrete change with prchange 422
 - Computing discrete change with prvalue 423
- 9.2 Interactions 423
 - 9.2.1 Computing sex differences in predictions with interactions . . 425

9.2.2	Computing sex differences in discrete change with interactions	426
9.3	Nonlinear nonlinear models	427
9.3.1	Adding nonlinearities to linear predictors	428
9.3.2	Discrete change in nonlinear models	429
9.4	Using praccum and forvalues to plot predictions	430
	Options	431
9.4.1	Example using age and age-squared	432
9.4.2	Using forvalues with praccum	434
9.4.3	Using praccum for graphing a transformed variable	435
9.4.4	Using praccum to graph interactions	436
9.4.5	Using forvalues with prvalue to create tables	438
9.4.6	A more advanced example*	441
9.4.7	Using forvalues to create tables with other commands	442
9.5	Extending SPost to other estimation commands	444
9.6	Using Stata more efficiently	444
9.6.1	profile.do	444
9.6.2	Changing screen fonts and window preferences	446
9.6.3	Using ado-files for changing directories	446
9.6.4	me.help file	446
9.7	Conclusions	447
A	Syntax for SPost commands	449
A.1	asprvalue	450
	Syntax	450
	Description	450
	Options	450
	Examples	451
A.2	brant	452
	Syntax	452
	Description	452

	Option	452
	Examples	453
	Saved results	454
A.3	case2alt	454
	Syntax	454
	Description	454
	Options	454
	Example	455
A.4	countfit	455
	Syntax	456
	Description	456
	Options for specifying the model	456
	Options to select the models to fit	456
	Options to label and save results	456
	Options to control what is printed	457
	Example	457
A.5	fitstat	459
	Syntax	459
	Description	459
	Options	459
	Examples	459
	Saved results	461
A.6	leastlikely	461
	Syntax	461
	Description	461
	Options	462
	Options for listing	462
	Examples	462
A.7	listcoef	464
	Syntax	464

	Description	464
	Options	464
	Options for nominal outcomes	465
	Examples	465
	Saved results	467
A.8	misschk	468
	Syntax	468
	Options	468
	Example	468
A.9	mlogplot	470
	Syntax	470
	Description	470
	Options	471
	Examples	472
A.10	mlogtest	473
	Syntax	473
	Description	473
	Options	473
	Examples	474
	Saved results	476
	Acknowledgment	476
A.11	mlogview	477
	Syntax	477
	Description	477
	Dialog box controls	477
A.12	Overview of prchange, prgen, prtab, and prvalue	478
	Syntax	478
	Examples	479
A.13	praccum	480
	Syntax	480

	Description	480
	Options	480
	Examples	481
	Variables generated	482
A.14	prchange	483
	Syntax	483
	Description	483
	Options	483
	Examples	484
A.15	prcounts	485
	Syntax	485
	Description	485
	Options	485
	Variables generated	486
	Examples	486
A.16	prgen	487
	Syntax	487
	Description	487
	Options	488
	Options for confidence intervals and marginals	488
	Examples	488
	Variables generated	489
A.17	prtab	490
	Syntax	490
	Description	490
	Options	491
	Examples	491
A.18	prvalue	493
	Syntax	493
	Description	493

Options	494
Options for confidence intervals	494
Options used for bootstrapped confidence intervals	494
Examples	495
Saved results	497
A.19 spex	498
Syntax	498
Description	498
Options	498
Examples	498
B Description of datasets	499
B.1 binlfp2	499
B.2 couart2	500
B.3 gsskidvalue2	501
B.4 nomocc2	502
B.5 ordwarm2	503
B.6 science2	504
B.7 travel2	506
B.8 wlsrnk	507
References	509
Author index	515
Subject index	517