A.3 Stata

For examples of categorical data analyses with Stata for many data sets in my text *An Introduction to Categorical Data Analysis*, see the useful site

www.ats.ucla.edu/stat/examples/icda/

set up by the UCLA Statistical Computing Center. Specific examples are linked below. See also *Regression Models for Categorical Dependent Variables Using Stata* by J. S. Long and J. Freese (Stata Press 2006) and *A Handbook of Statistical Analyses Using Stata*, 4th ed., by S. Rabe-Hesketh and B. Everitt (CRC Press, 2006). Many examples of the use of Stata for various generalized linear models are in *Generalized Linear Models and Extensions*, 2nd edition, by J. Hardin and J. Hilbe (Stata Press, 2007). A listing of the extensive selection of categorical data methods available as of 2002 in Stata was given in Table 3 of the article by R. A. Oster in the August 2002 issue of *The American Statistician* (pp. 235-246); the main focus of that article is on methods for small-sample exact analysis.

Chapter 1: Introduction

The *ci* command can construct confidence intervals for proportions, including Wald, score (Wilson), Agresti–Coull, Jeffreys Bayes, and Clopper–Pearson small-sample methods. See

www.stata.com/help.cgi?ci

The *bitest* command can conduct small-sample tests about a binomial parameter. See

www.stata.com/help.cgi?bitest

Chapters 2–3: Two-Way Contingency Tables

The *tabulate* command can construct two-way contingency tables, conduct chi-squared tests and Fisher’s exact test, and find various measures of association and their standard errors, including Goodman and Kruskal’s gamma and Kendall’s tau-b. See

www.stata.com/help.cgi?tabulate_twoway

for a summary and a list of options. See

www.ats.ucla.edu/stat/stata/examples/icda/icdast2.htm

for an example.

The *cs* command can construct confidence intervals for the difference of proportions, relative risk, and odds ratio. See

www.stata.com/help.cgi?cs

and for an example, see www.ats.ucla.edu/stat/stata/examples/icda/icdast2.htm

which also shows how to use *logit* to obtain an interval for the odds ratio. The *cc* command can also construct confidence intervals for odds ratios. See

www.stata.com/help.cgi?cc

and for an example, see www.ats.ucla.edu/stat/stata/examples/icda/icdast3.htm

For a Stata module for three-way tables, one can use the *tab3way* command,
Chapter 4: Generalized Linear Models

For the use of Stata for various generalized linear models, see Generalized Linear Models and Extensions, 2nd edition, by J. Hardin and J. Hilbe (Stata Press, 2007).

The `glm` command can fit generalized linear models such as logistic regression and loglinear models:

```stata
    glm y x1 x2, family(poisson) link(log) lnoffset(time)
```
for a Poisson model with explanatory variables \( x_1 \) and \( x_2 \), and for a binomial variate \( y \) based on \( n \) successes,

```stata
    glm y x1 x2, family(binomial n) link(logit)
```
for a logistic model. For examples, see [www.ats.ucla.edu/stat/stata/examples/icda/icdast4.htm](http://www.ats.ucla.edu/stat/stata/examples/icda/icdast4.htm).

Profile likelihood confidence intervals are available with the `pllf` and `logprof` (for logistic regression) commands in Stata. For `pllf`, see article by P. Royston in Stata Journal, vol. 7, pp. 376–387:


Chapters 5–7: Logistic Regression and Binary Response Methods

For a summary of all the Stata commands that can perform logistic regression, see [www.stata.com/capabilities/logistic.html](http://www.stata.com/capabilities/logistic.html)

Once a model has been fitted, the `predict` command has various options, including fitted values, the Hosmer–Lemeshow statistic, standardized residuals, and influence diagnostics.

In particular, other than with the `glm` command, logistic models can be fitting using the `logistic` and `logit` commands. See [www.stata.com/help.cgi?logistic](http://www.stata.com/help.cgi?logistic) and [www.stata.com/help.cgi?logit](http://www.stata.com/help.cgi?logit).
Code has the form

```
.logit y x [fw=count]
```

with the option of frequency weights. For examples, see

[www.ats.ucla.edu/stat/stata/examples/icda/icdast4.htm](http://www.ats.ucla.edu/stat/stata/examples/icda/icdast4.htm),

and for the horseshoe crab data and AIDS/AZT examples of Chapter 5, see

[www.ats.ucla.edu/stat/stata/examples/icda/icdast5.htm](http://www.ats.ucla.edu/stat/stata/examples/icda/icdast5.htm).

For a special command for grouped data, see [www.stata.com/help.cgi?glogit](http://www.stata.com/help.cgi?glogit).

In the `glm` command, other links, such as probit and cloglog, can be substituted for the logit. Probit models can also be fitting using `probit`. See


Conditional logistic regression can be conducted using the `clogit` command. See


The `exlogistic` command performs exact conditional logistic regression. See


`FIRTHLOGIT` is a Stata module to use Firth’s method for bias reduction in logistic regression. See

[ideas.repec.org/c/boc/bocode/s456948.html](http://ideas.repec.org/c/boc/bocode/s456948.html)

See also [http://www.homepages.ucl.ac.uk/~ucakgam/stata.html](http://www.homepages.ucl.ac.uk/~ucakgam/stata.html) for information about a package of penalized logistic regression programs that also includes the lasso as a special case.

Stata does not seem to currently have Bayesian capability.

### Chapter 8: Multinomial Response Models

The command `mlogit` can fit baseline-category logit models:


The code for a baseline-category logit model takes the form

```
mlogit y x1 x2 [fweight=freq]
```

For the alligator food choice example of the text, but using three outcome categories, see [www.ats.ucla.edu/stat/stata/examples/icda/icdast8.htm](http://www.ats.ucla.edu/stat/stata/examples/icda/icdast8.htm).

The command `mprobit` fits multinomial probit models, for the case of independent normal error terms. See


for details. The command `asmprobit` allows more general structure for the error terms.

The command `ologit` can fit ordinal models, such as cumulative logit models:


The code for the proportional odds version of cumulative logit models has form

```
. ologit y x [fweight=freq]
```

For an example, see [www.ats.ucla.edu/stat/stata/examples/icda/icdast8.htm](http://www.ats.ucla.edu/stat/stata/examples/icda/icdast8.htm). The corresponding command `oprobit` can fit cumulative probit models. See

[www.nd.edu/~rwilliam/oglm](http://www.nd.edu/~rwilliam/oglm)
for discussion of a new *oglm* command by Richard Williams for ordinal models that include as a special case cumulative link models with logit, probit, and complementary log-log link. Continuation-ratio logit models can be fitted with the *ocratio* module. See

www.stata.com/search.cgi?query=ocratio

A command *omodel* is available from the Stata website for fitting these models and testing the assumption of the same effects for each cumulative probability (i.e., the proportional odds assumption for cumulative logit models).

Chapters 9–10: Loglinear Models

Loglinear models can be fitted as generalized linear models using the *glm* command. For examples, including the high school student survey of alcohol, cigarette, and marijuana use from Chapter 9, see

www.ats.ucla.edu/stat/stata/examples/icda/icdast6.htm

That source also describes use of a special *ipf* command for iterative proportional fitting.

For an example of using *glm* to fit an association model such as linear-by-linear association, see [www.ats.ucla.edu/stat/stata/examples/icda/icdast7.htm](http://www.ats.ucla.edu/stat/stata/examples/icda/icdast7.htm). An example shown is the text example from Chapter 10 on opinions about premarital sex and birth control.

Chapter 11: Models for Matched Pairs

Most models in this chapter can be fitted as special cases of logistic or loglinear models, which are themselves special cases of generalized linear models with the *glm* command. Some specialized commands are also available. For example, *symmetry* tests symmetry and marginal homogeneity in square tables, and thus gives McNemar’s test for the special case of 2×2 tables. See

http://www.stata.com/help.cgi?symmetry

and see also [www.ats.ucla.edu/stat/stata/examples/icda/icdast9.htm](http://www.ats.ucla.edu/stat/stata/examples/icda/icdast9.htm) for an example and the use of the *mcc* command for McNemar’s test. That location also shows analyses of the coffee choice example from the text, and also the use of *glm* for fitting the Bradley–Terry model, with a tennis example.

The command *clogit* performs conditional logistic regression.

Chapters 12–14: Clustered Categorical Responses

For information about using GEE in Stata, see

www.stata.com/meeting/1nasug/gee.pdf

by Nicholas Horton (in 2001). The GEE method can be conducted using the *xtgee* command, see

with the usual distributions and link functions for the marginal models. Code has form such as

```
.xtgee y x1 x2, family(poisson) link(log) corr(exchangeable) robust
```

For ML fitting of generalized linear mixed models, the GLLAMM module described at [www.gllamm.org](http://www.gllamm.org) can fit a very wide variety of models, including logistic and cumulative logit models with random effects. For further details, see [www.stata.com/search.cgi?query=gllamm](http://www.stata.com/search.cgi?query=gllamm) and Chapter 5 of *Multilevel and Longitudinal Modeling Using Stata* by S. Rabe-Hesketh and A. Skrondal (Stata Press, 2005). For a discussion of its use of adaptive Gauss-Hermite quadrature, see [www.stata-journal.com/sjpdf.html?articlenum=st0005](http://www.stata-journal.com/sjpdf.html?articlenum=st0005).

Negative binomial regression models can be fitted with the `nbreg` command. See [www.stata.com/help.cgi?nbreg](http://www.stata.com/help.cgi?nbreg) and [www.ats.ucla.edu/stat/stata/dae/nbreg.htm](http://www.ats.ucla.edu/stat/stata/dae/nbreg.htm).

It is also possible to fit these models with the `glm` command, with the nbinomial option for the family. See [www.ats.ucla.edu/stat/stata/library/count.htm](http://www.ats.ucla.edu/stat/stata/library/count.htm).

### Chapter 15: Non-Model-Based Classification and Clustering

There is a `cart` module for classification trees, prepared by Wim van Putten. See [econpapers.repec.org/software/bocbocode/s456776.htm](http://econpapers.repec.org/software/bocbocode/s456776.htm).

Discriminant analysis is available with the `discrim` command. Options include linear discriminant analysis (subcommand `lda`, that is, the full command is `discrim lda`), quadratic discriminant analysis with subcommand `qda`, k nearest neighbor with subcommand `knn`, and logistic with subcommand `logistic`. See [www.stata.com/help.cgi?discrim](http://www.stata.com/help.cgi?discrim) and [http://www.stata.com/help.cgi?candisc](http://www.stata.com/help.cgi?candisc) for the canonical linear discriminant function.

For a summary of Stata capabilities for cluster analysis with the `cluster` command, see [www.stata.com/capabilities/cluster.html](http://www.stata.com/capabilities/cluster.html) and [www.stata.com/help.cgi?cluster](http://www.stata.com/help.cgi?cluster).

### Chapter 16: Large- and Small-Sample Theory for Multinomial Models


The `cc` command constructs small-sample confidence intervals for the odds ratio, unless one requests a different option. See [www.stata.com/help.cgi?cc](http://www.stata.com/help.cgi?cc).