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`

For example, for the Wald confidence interval for a proportion for a binary variable y that has the values 0 and 1, use the command

```
ci y, binomial wald
```

If you already have the sample size and count in the category of interest, Stata can construct the interval using them, with the `cii` command[1], such as

```
cii 1200 396, wald
```

for the Wald interval with 396 successes in 1200 trials, and

```
cii 1200 396, agresti
```

for the Agresti–Coull interval. See `www.stata.com/manuals14/rci.pdf` for further details and options.

To conduct a significance test of whether a categorical variable y that takes values 0 and 1 in the data file has a population proportion of 0.50 with the value 1:

```
prtest y == 0.50
```

If you already have summary statistics, you can use the `prtesti` command, by entering n, $\hat{\pi}$, and π_0, such as:

```
prtesti 1200 0.52 0.50
```

For further details and options, see `www.stata.com/manuals14/rprtest.pdf`.

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

`www.stata.com/help.cgi?bitest`

[1] Here, *i* following *ci* stands for *immediate*.

Chapters 2–3: Two-Way Contingency Tables

To test equality of proportions between binary variables $y1$ and $y2$ (that each take values 0 and 1) in a data file, use the command

```
prtest y1 == y2
```

This also shows the 95% confidence interval for the difference. To test equality of proportions for variable y between groups defined by a variable called *group* (such as gender), use

```
prtest y, by(group)
```

If you have summary statistics, you can find the inferences directly from the sample size and count in the category of interest for each group, such as

```
prtesti 604 0.522 597 0.509
```

With the `tabulate` command (`tab` for short), you can contruct contingency tables, find percentages in the conditional distributions (within-row relative frequencies), get expected frequencies for H_0: independence, get the chi-squared statistic and its P-value, and conduct Fisher's exact test. For categorical variables x and y in a data file, for instance,

```
tab x y, row expected chi2 exact gamma
```

If you already have the cell counts, you can enter them by row, such as with

```
tabi 495 590 272 \ 330 498 265, row expected chi2 exact gamma
```

To get standardized residuals, you currently must download a routine written by Nicholas Cox. Within Stata, use the command

```
ssc install tab_chi
```

then followed (if you have the cell counts) by

```
tabchii 495 590 272 \ 330 498 265, adjust
```

to get the standardized (adjusted) residuals.

For two variables $y1$ and $y2$, you can perform Fisher's exact test using the command

```
tab y1 y2, exact
```

You can enter the counts yourself from the contingency table that cross classifies $y1$ and $y2$ and request this test, such as using

```
tabi 3 1 \ 1 3, exact
```

See

`www.stata.com/help.cgi?tabulate_twoway` and

`www.stata.com/manuals14/rtabulatetwoway.pdf`

for a summary and a list of options.

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

```
csi 2509 409 189 2245 , or woolf
```

gives the 95% Woolf confidence interval for the odds ratio. See `www.stata.com/help.cgi?cs` for details, 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,

`ideas.repec.org/c/boc/bocode/s425301.html`

with an example at `www.ats.ucla.edu/stat/stata/examples/icda/icdast3.htm`. See also `www.stata.com/statalist/archive/2009-04/msg00893.html` and `www.stata.com/statalist/archive/2010-04/msg00800.html`.

Nicholas Cox has a package *tabchi* for basic analyses of contingency tables. See

`http://www.nd.edu/~rwilliam/stats1/Categorical-Stata.pdf`

for a document by Richard Williams that describes this and the use of Stata for basic analyses for categorical data analysis. In particular, it can generate standardized (adjusted) residuals, as shown in the example in `www.ats.ucla.edu/stat/stata/examples/icda/icdast2.htm`.

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:

`www.stata.com/help.cgi?glm`

The link functions (with keywords in parentheses) include log (log), identity (i), logit (l), probit (p), complementary log-log (c). The families include binomial (b), Poisson (p), and negative binomial (nb). Newton-Raphson fitting is the default. Code takes the form

```
. glm y x1 x2, family(poisson) link(log) lnoffset(time)
```

for a Poisson model with explanatory variables $x1$ and $x2$, and for a binomial variate y based on n successes,

```
. 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`.

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:

`www.stata-journal.com/sjpdf.html?articlenum=st0132`

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`.

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` and `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`,
and for the horseshoe crab data and AIDS/AZT examples of Chapter 5, see
`www.ats.ucla.edu/stat/stata/examples/icda/icdast5.htm`.
For a special command for grouped data, see `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

`www.stata.com/help.cgi?probit`

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

`www.stata.com/help.cgi?clogit`.

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

`www.stata.com/help.cgi?exlogistic`

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`

See also `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:

`www.stata.com/help.cgi?mlogit`

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`.

The command *mprobit* fits multinomial probit models, for the case of independent

normal error terms. See

`http://www.stata.com/help.cgi?mprobit`

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:

`www.stata.com/help.cgi?ologit`

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`. The corresponding command *oprobit* can fit cumulative probit models. See

`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`. 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` 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

`www.stata.com/help.cgi?xtgee` and `www.stata.com/capabilities/xtgee.html`,

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` 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`

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`.

Negative binomial regression models can be fitted with the *nbreg* command. See

`www.stata.com/help.cgi?nbreg` and `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`.

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`.

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`

and `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` and `www.stata.com/help.cgi?cluster`.

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

The *ci* command can construct small-sample confidence intervals for proportions, including Clopper–Pearson intervals. See

`www.stata.com/help.cgi?ci`

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`

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

`www.stata.com/help.cgi?exlogistic`