HANDOUT on SPSS for CDA, prepared for STA 4504/5503

This is taken from my book *Statistical Methods for the Social Sciences, 4th ed.*, published 2008.

As of version 15.0, most of the methods described below are available in a special student version of SPSS. The GENERALIZED LINEAR MODELS options in the ANALYZE menu mentioned below is available in an add-on module, "SPSS Advanced Models." Particular options for modeling categorical responses (such as logistic regression and a multinomial logistic model for nominal responses) are available with another add-on, "SPSS Regression Models." These two add-ons are also part of the "SPSS Graduate Pack" that is more advanced than the ordinary student version.

Analyzing Association Between Categorical Variables

The DESCRIPTIVE STATISTICS option on the ANALYZE menu has a sub-option called CROSSTABS, which provides several methods for contingency tables. After identifying the row and column variables in CROSSTABS, clicking on STATISTICS provides a wide variety of options, including the chi-squared test and measures of association. The output lists the Pearson statistic, its degrees of freedom, and its P-value (labeled *Asymp. Sig.*).

If any expected frequencies in a 2×2 table are less than 5, Fisher's exact test results. It can also be requested by clicking on Exact in the CROSSTABS dialog box and selecting the exact test. SPSS also has an advanced module for small-sample inference (called SPSS Exact Tests) that provides exact P-values for various tests in CROSSTABS and NPAR TESTS procedures. For instance, the Exact Tests module provides exact tests of independence for $r \times c$ contingency tables with nominal or ordinal classifications. See the publication *SPSS Exact Tests 15.0 for Windows*.

In CROSSTABS, clicking on CELLS provides options for displaying observed and expected frequencies, as well as the standardized residuals, labeled as *Adjusted standardized*. Clicking on STATISTICS in CROSSTABS provides options of a wide variety of statistics other than chi-squared, including gamma and Kendall's tau-b. The output shows the measures and their standard errors (labeled *Asymp. Std. Error*), which you can use to construct confidence intervals. It also provides a test statistic for testing that the true measure equals zero, which is the ratio of the estimate to its standard error. This test uses a simpler standard error that only applies under independence and is inappropriate for confidence intervals. One option in the list of statistics, labeled *Risk*, provides as output the odds ratio and its confidence interval.

Suppose you enter the data as cell counts for the various combinations of the two variables, rather than as responses on the two variables for individual

subjects; for instance, perhaps you call COUNT the variable that contains these counts. Then, select the WEIGHT CASES option on the DATA menu in the *Data Editor* window, instruct SPSS to weight cases by COUNT.

Logistic Regression and Other GLMs

To fit generalized linear models, on the ANALYZE menu select the GENERALIZED LINEAR MODELS option and the GENERALIZED LINEAR MODELS suboption. Select the Dependent Variable and then the Distribution and Link Function. Click on the Predictors tab at the top of the dialog box and then enter quantitative variables as Covariates and categorical variables as Factors. Click on the Model tab at the top of the dialog box and enter these variables as main effects, and construct any interactions that you want in the model. Click on OK to run the model.

To fit logistic regression models, on the ANALYZE menu select the REGRESSION option and the BINARY LOGISTIC suboption. In the LOGISTIC REGRESSION dialog box, identify the binary response (dependent) variable and the explanatory predictors (covariates). Highlight variables in the source list and click on $a * b$ to create an interaction term. Identify the explanatory variables that are categorical and for which you want dummy variables by clicking on Categorical and declaring such a covariate to be a Categorical Covariate in the LOGISTIC REGRESSION: DEFINE CATEGORICAL VARIABLES dialog box. Highlight the categorical covariate and under Change Contrast you will see several options for setting up dummy variables. The *Simple* contrast constructs them as in this text, in which the final category is the baseline.

In the LOGISTIC REGRESSION dialog box, click on *Method* for stepwise model selection procedures, such as backward elimination. Click on *Save* to save predicted probabilities, measures of influence such as leverage values and DFBETAS, and standardized residuals. Click on *Options* to open a dialog box that contains an option to construct confidence intervals for exponentiated parameters.

Another way to fit logistic regression models is with the GENERALIZED LINEAR MODELS option and suboption on the ANALYZE menu. You pick the binomial distribution and logit link function. It is also possible there to enter the data as the number of successes out of a certain number of trials, which is useful when the data are in contingency table form such as with the death penalty example in Table 15.3. For example, suppose in one column you have the number of successes at each particular setting of predictors, and in a separate column you have the sample size that number of successes is based on. Then, you identify the dependent variable as the variable listing the number of successes, you click the box "Variable represents binary response or number of events," and then "Number of events occurring in a set of trials," entering the variable listing the sample sizes as the "Trials variable."

You can also fit such models using the LOGLINEAR option with the LOGIT suboption in the ANALYZE menu. You identify the dependent variable, select categorical predictors as factors, and select quantitative predictors as cell covariates. The default fit is the saturated model for the factors, without including any covariates. To change this, click on *Model* and select a *Custom* model, entering the predictors and relevant interactions as terms in a customized (unsaturated) model. Clicking on *Options*, you can also display standardized residuals (called adjusted residuals) for model fits. This approach is well suited for logit models with categorical predictors, such as discussed in Section 15.2, since standard output includes observed and expected frequencies. When the data file contains the data as cell counts, such as binomial numbers of successes and failures, you weight each cell by the cell count using the WEIGHT CASES option in the DATA menu.

SPSS can also fit logistic models for categorical response variables having several response categories. On the ANALYZE menu, choose the REGRESSION option and then the ORDINAL suboption for a cumulative logit model. Select the MULTINOMIAL LOGISTIC suboption for a baseline-category logit model. In the latter, click on *Statistics* and check Likelihood-ratio tests under Parameters to obtain results of likelihood-ratio tests for the effects of the predictors.

For loglinear models, use the LOGLINEAR option with GENERAL suboption in the ANALYZE menu. You enter the factors for the model. The default is the saturated model, so click on *Model* and select a *Custom* model. Enter the factors as terms in a customized (unsaturated) model and then select additional interaction effects. Click on *Options* to show options for displaying observed and expected frequencies and adjusted residuals. When the data file contains the data as cell counts for the various combinations of factors rather than as responses listed for individual subjects, weight each cell by the cell count using the WEIGHT CASES option in the DATA menu.