Appendix A

SPSS and SAS for Statistical Analyses

Major statistical software packages have procedures for nearly all the methods presented in this text. This appendix illustrates software use for these methods.

There is insufficient space to discuss all the major packages, and we focus on SPSS and SAS. We discuss basic use of the software rather than the great variety of options provided by the procedures. For ease of reference, the material both for SAS and SPSS is organized by chapter of presentation in this text. The full data files for examples and exercises having large data sets are available at
`http://www.stat.ufl.edu/~aa/social/data.html`

Introduction to SPSS

SPSS for Windows has a graphical user interface that makes requesting statistical procedures simple. In this windows environment, SPSS provides menus and dialog boxes to save you the work of preparing code to request analyses. Our discussion below applies to version 15.0.

Starting and Using SPSS

When you start a session, you see a ***Data Editor window*** that contains a menu bar with a wide variety of separate menus. These include a FILE menu for creating a new file or opening an existing one, an ANALYZE menus that displays options for selecting a statistical method, a GRAPHS menu for creating a graph of some type, and menus for choosing other special features. The *Data Editor* window displays the contents of the data file. You use this to enter the data or edit existing data. The ***Output window*** shows results of the analyses after you request them from the menus. You can edit and save this for later use or printing. Other windows are available for charts and plotting data. Using the windows environment is mostly self-explanatory, but on-line help is readily available. It also helps to have access to a manual such as Norusis (2006) or Green and Salkind (2007).

Data files have the form of a spreadsheet, each column containing a variable and each row the observations for a particular subject. You can enter the data set while in SPSS. Or you can simply call up an existing data set, which might be an ASCII file that you created with a separate editor or a file created with a spreadsheet or database program. Select FILE on the menu bar in the application window and indicate whether you want to read in a new data file (such as crime.dat) or open an existing file (such as crime.sav) or enter a new data set on the spreadsheet.

Let's go through the steps of reading a text file into the SPSS data editor. In the *Data Editor* window, go to the FILE menu and select OPEN and then DATA. Enter the FILE NAME and select the TYPE

OF FILE (such as .txt file). Click on OPEN. This opens the TEXT IMPORT WIZARD. You will now be led through some steps to create the SPSS data file.

The first step asks if your text file matches a predefined format. Unless you had created a format to read the data, select NO and click NEXT. The second step asks how your variables are arranged. For files at the text website that are delimited by a space between each column of the data file, you select DELIMITED, tab, or space. You should also indicate here whether the variable names are at the top of the file (as at the text website). The third step asks whether each line represents a case (as is true for the data files at the text website) and whether you want to import all the cases (as you usually would). You also see here a preview of the data. The fourth step asks you to identify the delimiter between variables, which is a SPACE for the text website. Check the data preview and make sure it looks correct. In step five you have the option to change variable names and to identify the data format. SPSS should identify quantitative variables as NUMERIC and categorical variables (with labels for the categories) as STRING. At step 6, clicking on FINISH reads the data into the SPSS data editor. Check this to make sure that unintended spaces in your text file did not result in an added variable or case that needs to be deleted. A period in a cell represents a missing observation.

For your new data file, you can re-define names and characteristics for each variable after clicking on VARIABLE VIEW. (In the Measure column, make sure SPSS has not inappropriately labelled a variable as NOMINAL that should be SCALE (interval) or ORDINAL.) You save the file by selecting the FILE menu with the *Save As* option. The name you select receives the *.sav* extension. You can access the created file later from the FILE menu.

At this point, you can select a statistical procedure from the ANALYZE menu on the *Data Editor*. When you select a procedure, a *dialog box* opens that shows you the source variables in your data set. You highlight the ones you want to use currently and click on the arrow to the right of the list to move them to the selected variables list further to the right. You then click on OK and the procedure runs, showing results in the output window. For many procedures, you can click on *Options* and an additional *subdialog box* will open that displays extra available options for the method. To save output, in the OUTPUT window, use the FILE menu and the *Save As* option. One can later access the output in the FILE menu by selecting OPEN and then OUTPUT.

Chapter 3: Descriptive Statistics

To construct frequency distributions, histograms, and basic summary statistics, on the ANALYZE menu of the *Data Editor* window select the DESCRIPTIVE option with the FREQUENCIES suboption. A FREQUENCIES dialog box will open. Select the variables you want from the list for your file. Then, clicking on *OK* provides a frequency distribution in the *Output* window. Clicking on CHARTS in the FREQUENCIES dialog box presents you with a FREQUENCIES: CHARTS dialog box containing a histogram option for quantitative variables and a bar chart option for categorical variables. You can also construct a histogram from the GRAPHS menu on the *Data Editor* window by selecting INTERACTIVE and then HISTOGRAM. Drag a variable to the x-axis and then select whether you want to plot counts or percentages.

To construct a stem-and-leaf plot or a box plot, from the STATISTICS menu select the SUMMARIZE option with the EXPLORE suboption. The EXPLORE dialog box contains a *Plots* option; clicking on it reveals the various plot options. (The GRAPHS menu also has the option of a box plot.)

Clicking on STATISTICS in the FREQUENCIES dialog box presents you with a FREQUENCIES: STATISTICS dialog box containing options of various statistics to compute, such as measures of central tendency and variability. You can also get basic descriptive statistic from the ANALYZE menu, by selecting the DESCRIPTIVE STATISTICS option with the DESCRIPTIVES suboption.

To construct a box plot, on the GRAPHS menu in the *Data Editor* window, select CHARTBUILDER and then drag the box plot icon into the open canvas. Select the variable for the box plot and click OK. CHARTBUILDER also has options for side-by-side box plots according to the value of a categorical variable. You can also do this on the GRAPHS menu by selecting INTERACTIVE and then BOXPLOT.

Chapters 5 and 6: Estimation and Significance Tests

The ANALYZE menu on the *Data Editor* window has a COMPARE MEANS option with a ONE-SAMPLE T TEST suboption. The default with that option is a 95% confidence interval for the mean and a t test that the true mean equals 0. The options permit one to select a different confidence level. To test that the mean equals a constant μ_0, supply that number as in the Test Value box on the ONE-SAMPLE T TEST dialog box.

Here's how to find the t score corresponding to a certain tail probability. We illustrate with left-tail probability = 0.025, for $df = 226$. In an empty data editor window, put 0.025 in the first column and 226 in the second column. Then on the *Transform* menu select *Compute variable*.

In the Numeric Expression window, type IDF.T(var1,var2), where var1 and var2 are the names of the variables for the first two columns. Enter a name for the Target Variable. Click on OK, and the solution (−1.97052) will show up in a new column of the data file labeled with the name you gave the target variable. The code for other distributions is listed under the Functions and Special Variables menu in this dialog box, when you click on Inverse DF in the Function group menu (e.g., IDF.normal for the normal distribution).

Likewise, you can find the tail probability for a certain t score. To find the left-tail (i.e., cumulative) probability for a t score of 1.97 when $df = 226$, put 1.97 in the first column and 226 in the second column. Then on the *Transform* menu select *Compute variable*. In the Numeric Expression window, type CDF.T(var1,var2), where var1 and var2 are the names of the variables for the first two columns. Enter a name for the Target Variable. Click on OK, and the solution (0.97497) will show up in a new column of the data file labeled with the name you gave the target variable. To get the right-tail probability above the given t score, subtract this cumulative probability from 1. The code for other distributions is listed under the Functions and Special Variables menu in this dialog box, when you click on CDF in the Function group menu (e.g., CDF.normal for the normal distribution).

Chapter 7: Comparison of Two Groups

The ANALYZE menu has a COMPARE MEANS option with a INDEPENDENT-SAMPLES T TEST suboption. One selects the response variable (labeled the *Test variable*) and the variable that defines the two groups to be compared (labeled the *Grouping variable*). With Define Groups under the Grouping Variable label, identify the two levels of the grouping variable that specify the groups to be compared. Click OK, and results are shown in the *Output* window. There, in the *Equal variances* row, this procedure provides the results of the t test assuming the two populations have the same standard deviation value. The output provides the test statistic (labeled t), df, standard error (labeled *Std. Error Difference*), and two-sided P-value (labeled *Sig. (2-tailed)*). The procedure also provides the 95% confidence interval for comparing the means. Options allow one to change the confidence level. The output also shows results for the method that does not assume equal variances, in the *Unequal variances* row of the output.

The COMPARE MEANS option on the ANALYZE menu also has a PAIRED-SAMPLES T TEST suboption, which supplies the dependent-samples comparisons of means described in Section 7.4. For Fisher's exact test, see the description for the following chapter.

Chapter 8: Analyzing Association Between Categorical Variables

The DESCRIPTIVE STATISTICS option on the ANALYZE menu has a suboption 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.

Chapter 9: Linear Regression and Correlation

To construct a scatter diagram, enter the GRAPH menu on the *Data Editor* and choose the INTERACTIVE option and then SCATTERPLOT. In the CREATE SCATTERPLOT dialog box, drag the appropriate variables to the y and x axes. Click on FIT in this box and then as your method select REGRESSION if you want the regression line plotted over

the points or click NONE if you want only the scatterplot. Click on OK and you will see the graph in the *Output* window.

To fit the regression line, on the ANALYZE menu select REGRESSION and then LINEAR. You identify the response (Dependent) variable and the explanatory (Independent) variable. Various options are available by clicking on *Statistics* in the LINEAR REGRESSION dialog box, including estimates of the model parameters, confidence intervals for the parameters, and model fit statistics. After selecting what you want, click on CONTINUE and then back in the LINEAR REGRESSION dialog box click on OK. You'll see results in the *Output* window. Output for the estimates option includes the estimates for the prediction equation (labeled B), their standard errors (labeled *Std. Error*), the t statistic for testing that a regression parameter equals 0 and the associated P-value (labeled *Sig*), and a standardized regression coefficient (labeled as *Beta*) that in this bivariate model is simply the Pearson correlation.

Output for the model fit option in a *Model Summary* table includes the r^2 value (labeled *R Square*) and the estimate s of the conditional standard deviation (rather confusingly labeled *Std. Error of the Estimate*).

Chapter 11: Multiple Regression and Correlation

For a multiple regression analysis, again choose REGRESSION from the ANALYZE menu with the LINEAR suboption, and add additional variables to the list of independent variables. Among the options provided by clicking on *Statistics* in the dialog box are estimates of the coefficients and confidence intervals based on them and detail about the model fit. For the *Estimates* option, the output includes standard errors of the estimates, the t statistic for testing that the regression parameter equals zero and its associated two-sided P-value, and the estimated standardized regression coefficient (labeled *Beta*).

Requesting the *Model fit* option in the STATISTICS sub-dialog box provides additional information. For instance, the F statistic provided in the *ANOVA* table is the F test statistic for testing that the coefficients of the explanatory variables all equal 0. The probability labeled as *Sig* is the P-value for that test. Also provided in a *Model Summary* table are the multiple correlation R, R^2, and the estimate s of the conditional standard deviation (poorly labeled as *Std. Error of the Estimate*).

In the ANALYZE menu in the *Data Editor* window, selecting CORRELATE and then BIVARIATE gives a BIVARIATE CORRELATIONS sub-dialog box. You select the variables you want, check PEARSON CORRELATION and then click OK, and the output window shows a correlation matrix with the P-values for testing the significance of each.

To construct a scatterplot matrix, from the GRAPHS menu in the

Data Editor choose the CHART BUILDER option. Then click the GALLERY tab and select SCATTER/DOT in the Choose From list. Drag the Scatterplot Matrix icon onto the blank canvas. Drag the wanted variables to the Scattermatrix drop zone, and then click on OK. You'll see the graph in the *Output* window.

To produce all partial regression plots, click on PLOTS in the LINEAR REGRESSION dialog window and then click on *Produce all partial plots* in the LINEAR REGRESSION: PLOTS dialog box.

To obtain a partial correlation analysis, choose the PART AND PARTIAL CORRELATIONS option in the STATISTICS option box in the LINEAR REGRESSION window. Or, in the ANALYZE menu choose the CORRELATE option with the PARTIAL suboption. In the resulting PARTIAL CORRELATIONS dialog box, select the variables to correlate and select at least one variable to control. The output also provides tests of significance for these partial correlations.

To model interaction, you can construct an interaction variable within the SPSS data editor by selecting the COMPUTE VARIABLE option on the TRANSFORM menu. Provide a name for the new variable in the Target Variable box. Create the mathematical formula for the interaction term in the Numeric Expressions box, such as LIFE*SES for the explanatory variables LIFE and SES (the * symbol represents multiplication). Click OK, and in the data file in the *Data Editor* you will see a new column of observations for the new variable. This variable can then be entered into the model formula when requesting a regression equation.

Here's a second way to build an interaction term in a model, one that is especially useful for models in following chapters that also have categorical predictors. This method requires the student to use the general linear model function rather than the multiple linear regression function within SPSS. This second method is well suited for forming multiple interaction terms but presents output in a slightly different form and offers fewer options for data analysis. Choose the GENERAL LINEAR MODEL in the ANALYZE menu and select the UNIVARIATE suboption. Enter the response variable into the Dependent Variable box and the explanatory variables into the Covariate(s) box. Now click on the *Model* box and select the *Custom* option. Using the Build Term(s) arrow, enter the covariates as *Main effects*. Highlight a pair of variables for which you want a cross-product and enter them by selecting *Interaction* on the Build Term(s) arrow. Or, you can select the *All 2-way* option for the Build Term(s) arrow to request interaction terms for all pairs of variables. After specifying the terms for the model, click Continue and return to the UNIVARIATE dialog box. To display model parameter estimates, select the Options box and check the Parameter Estimates option. Click Continue to return to the UNIVARIATE dialog box and

then click OK to perform the regression analysis.

Chapter 12: Comparing Groups: Analysis of Variance Methods

To conduct a one-way ANOVA, on the ANALYZE menu select the COMPARE MEANS option with the ONE-WAY ANOVA suboption. Select the dependent variable and select the factor that defines the groups to be compared. (This must be coded numerically for SPSS to display it as a potential factor, even though it is treated as nominal scale! Otherwise, use the approach in the following paragraph.) Results provided include the F test statistic and its P-value, and sums of squares and mean squares for between-groups and within-groups variation. Clicking on *Post Hoc* in the ONE-WAY ANOVA dialog box provides a variety of options for multiple comparison procedures, including the Bonferroni and Tukey methods. The LSD (least significant difference) option provides ordinary confidence intervals with the confidence level applying to each interval. Clicking on *Options* in the ONE-WAY ANOVA dialog box provides the *Descriptive statistics* option of additional descriptive statistics, including the mean, standard deviation, standard error, and a 95% confidence interval for each group.

You can also conduct a one-way ANOVA on the ANALYZE menu by selecting the GENERAL LINEAR MODEL option with the UNIVARIATE suboption. With this approach, unlike the one just described, the categorical variable that is selected as the Fixed Factor can be coded with labels rather than numerically (i.e., a *string* variable in SPSS). In the UNIVARIATE dialog box, click on *Options* and you can request Descriptive statistics and Parameter estimates for displaying the regression parameter estimates from viewing the analysis as a special case of a regression analysis. Return to the UNIVARIATE dialog box and click on *Post Hoc* to select ordinary confidence intervals for comparing means (LSD) or multiple comparison intervals such as Bonferrroni or Tukey.

To conduct a two-way or higher-way factorial ANOVA, on the ANALYZE menu select the GENERAL LINEAR MODEL option with the UNIVARIATE suboption. Select the dependent variable and select the Fixed Factor(s) that define the cross-classification for the means. (If you have set up dummy variables yourself, they would be entered as Covariates.) The default model is a full factorial model containing all interactions. Click on *Model* to build a customized model that contains only some or none of the interactions. Highlight variables, select Interaction or Main Effects from the Build Term(s) list, and click on the arrow to move the terms to the model list on the right. Return to the UNIVARIATE dialog box and click on *Options.* You can request Descriptive statistics, Parameter estimates, and you can select particular

factors and request Display Means to see the observed and predicted means for subgroups defined by the factors. Return to the UNIVARIATE dialog box and click on *Contrasts* to display parameter estimates with standard errors, t statistics, and confidence intervals for comparing means for levels of each factor. Change the contrast type to *Simple* to compare each level to a baseline level, either the last (such as in setting up (1, 0) dummy variables for all levels but the last one) or the first. Return to the UNIVARIATE dialog box and click on *Post Hoc* to select confidence intervals for comparing means (LSD) or multiple comparison intervals such as Bonferrroni or Tukey.

Alternatively, for 1-way or factorial ANOVA, you could set up dummy variables in your data file and then use ordinary regression. On the ANALYZE menu, you would then select the REGRESSION option and LINEAR suboption, as in Chapter 11.

You can conduct repeated measures ANOVA using the GENERAL LINEAR MODEL option on the ANALYZE menu, with the REPEATED MEASURES suboption. This assumes that for each subject, each outcome for the response falls in a different column. For Example 12.8 on three influences, in a given row you would put the response for Movies in one column, for TV in a separate column, and for Rock in a third column. In the REPEATED MEASURES DEFINE FACTOR(S) dialog window, type the name and number of levels of the within-subjects factor (such as *influence* and 3) and click on *Add*. Then click on *Define* to define the model. Now, in the REPEATED MEASURES dialog box, select the between-subjects factors (if there are any), and select the response variable for each level of the within-subjects factor (such as Movies, TV, Rock). The default is a model containing all the factor interactions. Click on *Model*, and customize the model if you want to delete an interaction. Return to the REPEATED MEASURES dialog box and click on *Contrasts*, and options are provided for displaying parameter estimates and confidence intervals for contrasts comparing means in different factor levels, and for individual or Bonferroni confidence intervals. Change the contrast type to *Simple* for estimates of the between-subjects factors to refer to comparing each factor level to the first or last level. Return to the REPEATED MEASURES dialog box and click on *Options*, and you can request between-subjects observed and estimated means and various model diagnostics.

For repeated measures analyses, SPSS also reports results of standard multivariate (MANOVA) tests that do not make the assumption of sphericity for the joint distribution of the repeated responses (see Section 16.1). They are less powerful than the repeated measures ANOVA methods when the sphericity assumption is not violated.

Chapter 13: Combining Regression and ANOVA: Quantitative and Categorical Predictors

To fit an analysis of covariance model, you can set up dummy variables for categorical predictors and use ordinary regression procedures, such as described earlier for Chapter 11. To create cross-product terms for interactions, after creating the data file, you can select COMPUTE VARIABLE on the TRANSFORM menu and create products of appropriate variables.

Alternatively, on the ANALYZE menu select the GENERAL LINEAR MODEL option with the UNIVARIATE suboption. Proceed as described above for Chapter 11, now adding quantitative covariates in the Covariate(s) box. As in ANOVA, add categorical predictors to the Fixed Factor(s) box. Click on Model to build a custom model that contains only some or none of the interactions. Select *Interaction* or *Main Effects* from the Build Term(s) list, and click on the arrow to move the terms to the model list on the right.

Chapter 14: Model Building with Multiple Regression

In the LINEAR REGRESSION dialog window for the REGRESSION choice on the ANALYZE menu, you can select a *Method* for fitting the model, among which are options such as BACKWARD, FORWARD, and STEPWISE for selecting predictors in the model (or ENTER for adding all of them).

In the LINEAR REGRESSION dialog window, you can plot studentized residuals (labelled SRESID) and request all partial regression plots by clicking on *Plots* and then making appropriate selections in the PLOTS dialog box. To obtain predicted values, residuals, studentized residuals, leverage values, and influence diagnostics, click on *Save* in the LINEAR REGRESSION dialog box. The resulting LINEAR REGRESSION: SAVE dialog box contains options for these, such as such as *Standardized DfBeta(s)* for DFBETAS and *Standardized DfFit* for DFFITS. To find variance inflation factors, click on *Statistics* in the LINEAR REGRESSION dialog box and select *Collinearity diagnostics*.

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. (If you build a model assuming the gamma distribution, make

sure that in the Estimation dialog box you pick Maximum Likelihood Estimate for the Scale Parameter Method.)

To fit a quadratic regression model, on the ANALYZE menu select the REGRESSION option with the CURVE ESTIMATION suboption. Then, in the CURVE ESTIMATION dialog box, select the variables and choose the *Quadratic* model. The PLOT MODELS option provides a plot of the fitted curve. It can be useful to choose the Linear and Quadratic models so this plot shows the comparison.

To obtain a smoothing curve (a "kernel" method similar to LOESS mentioned in the text at the end of Section 14.5), on the GRAPHS menu of the data editor select INTERACTIVE and then SCATTERPLOT. After selecting the variables, click on FIT on the top of the CREATE SCATTERPLOT window and then pick SMOOTHER from the METHOD menu. Using a normal kernel provides the most smoothing, and is usually a sensible choice. The Bandwidth option determines how much smoothing is done, with larger numbers providing more smoothing. After selecting the options, click on OK. You can try a few bandwidth choices, such as 1, 2, and 5, to see its effect on the smoothing.

To fit the exponential regression model, on the ANALYZE menu select the GENERALIZED LINEAR MODELS option and the GENERALIZED LINEAR MODELS suboption. Enter the response variable and select the Log link function. Click on the Predictor tab and add the quantitative predictor as a Covariate. Specify the predictor as an effect after clicking on the Model tab, and click OK to see the fit.

There is also an option for an exponential regression model by selecting the CURVE ESTIMATION suboption under the REGRESSION option in the ANALYZE menu. However, this provides a somewhat different fit than using GLM software, since it assumes the log of y, rather than y, is normally distributed with constant variance. As discussed following Example 14.7, it fits the model $E[\log(Y)] = \alpha + \beta x$ rather than the model $\log[E(Y)] = \alpha + \beta x$.

Chapter 15: Logistic Regression

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.

One can also fit such models using the LOGLINEAR option with the LOGIT suboption in the ANALYZE menu. One identifies the dependent variable, selects categorical predictors as factors, and selects 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*, one 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, one weights 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, one uses the LOGLINEAR option with GENERAL suboption in the ANALYZE menu. One enters 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.

Introduction to SAS

The **SAS** language consists of DATA steps that name the variables and input the data and PROC steps that request the statistical procedures. All SAS statements must end with a semicolon. The first statement, the DATA statement, assigns a name to the data set. The next statement, the INPUT statement, tells SAS the variable names and the order in which the variables are listed in the data set. The data follow the DATALINES statement, one line per subject, unless the INPUT statement ends with @@. After the data lines, a line containing only a semicolon ends the data set.

Following the data entry, PROC statements invoke the statistical procedures. A typical PROC statement lists the procedure, such as MEANS, and then also may select some options for providing greater detail than SAS provides with the default statement. The text by Schlotzhauer and Littell (1997) introduces SAS and its use for basic statistical methods.

Chapter 3: Descriptive Techniques

Table A.1 shows the format for entering the data and performing some very basic analyses, using the data set in Table 3.1 on murder rates for the 50 states. For each state, we list the state label and its value on murder rate. When you input characters rather than numbers for a variable, such as the state labels, the variable has an accompanying $ label in the INPUT statement. We enter the 50 observations as 50 lines of data, or we can enter multiple observations on a line if we enter @@ at the end of the input line, as shown in Table A.1.

The first procedure, PROC PRINT, prints the data set. PROC FREQ provides a frequency distribution for the variable listed following TABLES. PROC CHART provides a histogram of the variable listed in the VBAR statement. Options exist for choosing the number of bars (e.g., VBAR MURDER / LEVELS = 5) or their midpoints and for forming horizontal rather than vertical bars (HBAR instead of VBAR).

PROC MEANS provides the mean and standard deviation of the variables listed after VAR (in this case, the murder rate). The PROC UNIVARIATE statement requests a greater variety of basic statistics for a variable, including the sample size, mean, standard deviation, median, mode, range, and quartiles. The ID statement, which is optional, names STATE as the variable to identify some of the extreme observations in part of the display from this procedure. Listing the PLOT option in PROC UNIVARIATE requests stem-and-leaf and box plots for the variables listed.

Table A.1: SAS for Printing Data, Computing Basic Summary Statistics, and Preparing Plots

```
data crime ;
input  state $  murder @@;
datalines;
  AL  11.6  AK   9.0  AZ  8.6
  AR  10.2  CA  13.2  CO  5.8
  ...
;
proc print ;  var  state  murder ;
proc freq;  tables  murder ;
proc chart;  vbar  murder ;
proc means;  var  murder ;
proc univariate   plot;  var  murder ;  id  state;
run ;
```

Chapters 5 and 6: Estimation and Significance Tests

The estimated standard error for a sample mean is one of the statistics provided by PROC UNIVARIATE. It is labeled as *Std Mean* in the output. We construct a confidence interval for a population mean by taking the sample mean and adding and subtracting the appropriate t-score times the standard error.

Table A.2 shows how to obtain the standard error and the t-score for the data from Example 6.4, for which $n = 29$ and $df = 28$. The two arguments for the TINV function are half the error probability and the df value. For instance, the statement in Table A.2 requests the t-score with left tail probability equal to 0.025 (for a 95% confidence interval) when $df = 28$, which equals -2.048. That table also shows how to input data for two dependent samples (WEIGHT1 and WEIGHT2 being the weights of anorexic girls at two times) and create a new variable (DIFF) that is the difference between WEIGHT2 and WEIGHT1.

Chapter 7: Comparison of Two Groups

Table A.3 uses SAS to perform a two-sample t test for comparing two means (Section 7.3), using the data in Example 7.7. The input variables are THERAPY, the levels of which are the two groups to be compared, and CHANGE, the change in weight, which is the response variable. PROC SORT sorts the data into groups, according to the levels of therapy, and then PROC MEAN finds means and standard deviations for

Table A.2: SAS for Obtaining Standard Errors and t-Scores

```
data anorexia ;
input weight1 weight 2 ;
diff = weight2 - weight1;
datalines;
  80.5  82.2
  84.9  85.6
  ...
;
proc univariate ; var diff ;
data findt;
       tvalue = tinv(.025, 28) ;
proc print   data = findt ;
run ;
```

the observations in each group, when one uses BY followed by the group variable. The BY statement is used with SAS procedures when you want to do an analysis separately for each level of the variable specified in the BY statement.

PROC TTEST is a procedure for a two-sample t test with independent samples. The CLASS statement names the variable that identifies the groups to be compared, and the VAR statement identifies the response variable for the analysis. The output shows the mean, standard deviation, and standard error for the response variable in each group, and provides the t test statistic, its *df* value, and a two-sided P-value, which is labeled by $Prob > |T|$. The approximate test is also provided that does not assume equal population variances for the two groups.

Chapter 8: Analyzing Association Between Categorical Variables

Table A.4 illustrates SAS for analyzing two-way contingency tables, using data from Table 8.1. PROC FREQ conducts chi-squared tests of independence using the CHISQ option and provides the expected frequencies for the test with the EXPECTED option. The MEASURES option provides a wide assortment of measures of association (including gamma for ordinal data) and their standard errors. For 2×2 tables this option provides confidence intervals for the odds ratio (labeled "case-control" on output) and the relative risk. The EXACT option provides Fisher's exact test. SAS lists the category levels in alphanumeric order

Table A.3: SAS for Two-Sample t Test for Example 7.7 (see Table 12.21 for the data)

```
data depress;
input therapy $ change @@ ;
datalines;
 cogbehav 1.7
 cogbehav 0.7
...
 control -0.5
 control -9.3
...
;
proc sort; by therapy ;
proc means; by therapy ; var change ;
proc ttest; class therapy ; var change ;
run;
```

unless you state ORDER=DATA in the PROC directive, in which case the levels have the order in which they occur in the input data.

You can also perform chi-squared tests using PROC GENMOD, as Table A.4 shows. This procedure, discussed in greater detail in the discussion below for Chapter 14, uses a generalized linear modeling approach introduced in Section 14.4. (The code in Table A.4 views the independence hypothesis as a "loglinear model" for Poisson counts with main effects of gender and party but no interaction.) The OBSTATS and RESIDUALS options in GENMOD provide cell residuals; the output labeled *StReschi* is the standardized residual.

Chapter 9: Linear Regression and Correlation

Table A.5 uses SAS to perform linear regression, using the "statewide crime 2" dataset, shown partly in Table 9.1. The PROC PLOT statement requests a scatterplot for murder rate and poverty rate; the first variable listed goes on the y axis. The PROC REG statement requests a regression analysis, predicting murder rate using poverty rate. The P option following this model statement requests the predicted values and residuals for all observations. The PROC CORR statement requests the correlation between each pair of variables listed in the VAR list.

Table A.4: SAS for Chi-Squared Test with Table 8.1

```
data politics;
input  gender  $  party  $  count  @@;
datalines;
  Female  Democ  573   Female  Indep  516  Female  Repub  422
  Male    Democ  386   Male    Indep  475  Male    Repub  399
;
proc freq;  weight count ;
      tables  gender*party / chisq  expected  measures ;
proc genmod;  class gender party;
      model  count = gender  party / dist=poi  link=log  obstats  residuals;
run;
```

Chapter 11: Multiple Regression and Correlation

Table A.6 uses SAS to perform multiple regression with the mental impairment data of Table 11.1. You list every explanatory variable in the model to the right of the equal sign in the model statement. The PARTIAL option provides partial regression scatterplots. The PCORR2 option provides squared partial correlations, and the STB option provides standardized regression coefficients. Following the input statement, we define a variable life_ses to be the cross-product of life events and ses. We enter that variable in the second regression model to permit interaction in the model.

Chapter 12: Comparing Groups: Analysis of Variance Methods

Table A.7 uses SAS to perform one-way ANOVA with Table 12.1 and two-way ANOVA with Table 12.10. The first PROC MEANS statement requests sample means on ideology for the data grouped by party. PROC GLM is a procedure for *general linear models*. It is similar in many ways to the regression procedure, PROC REG, except that PROC GLM can use CLASS statements to create dummy variables in order to include categorical predictors in the model.

The first GLM statement requests a one-way ANOVA, comparing ideology by party. The CLASS statement requests dummy variables for the levels of party. The MEANS option provides multiple comparison confidence intervals. Here, we request the Bonferroni and Tukey methods and specify alpha = .10 for overall 90% confidence. The SOLUTION

Table A.5: SAS for Regression Analysis with Table 9.1

```
data crime ;
input state $  violent  murder  metro  white  hs  poverty  single ;
datalines;
  AK   761    9.0  41.8   75.2   86.6    9.1   14.3
  AL   780   11.6  67.4   73.5   66.9   17.4   11.5
  AR   593   10.2  44.7   82.9   66.3   20.0   10.7
  AZ   715    8.6  84.7   88.6   78.7   15.4   12.1
  CA  1078   13.1  96.7   79.3   76.2   18.2   12.5
  ....
;
proc print  data=crime;
proc plot; plot murder*poverty ;
proc reg;  model  murder = poverty / p;
proc corr;  var  violent  murder  metro  white  hs  poverty  single ;
run;
```

option requests the estimates for the prediction equation.

The second PROC MEANS requests sample means on ideology for each of the combinations of party and gender. Following that is a GLM statement to conduct a two-way ANOVA using party and gender as

Table A.6: SAS for Multiple Regression Analysis with Table 11.1

```
data mental ;
input  impair  life   ses ;
life_ses = life*ses;
datalines;
  17   46   84
  19   39   97
  ....
;
proc print;  var  impair   life   ses ;
proc plot ;  plot  impair*life  impair*ses ;
proc reg;  model  impair =  life  ses  /  partial  stb  pcorr2 ;
proc reg;  model  impair =  life  ses  life_ses ;
run;
```

Table A.7: SAS for One-Way ANOVA with Table 12.1 and Two-Way ANOVA with Table 12.10

```
data anova;
input party  $   gender  $   ideology ;
datalines;
  Dem  F  1
  Dem  F  1
  Dem  F  1
  Dem  F  1
  Dem  F  1
  Dem  F  2
  ...
  Rep  M  7
;
proc means;   by party;   var ideology;
proc glm;   class party ;
    model ideology = party   /   solution;
    means party   /   bon   tukey   alpha=.10;
proc means;   by   party   gender;   var   ideology;
proc glm;   class   party   gender;
    model ideology = party gender   /   solution;
    means party /   bon   tukey;
proc glm;   class   party   gender;
    model ideology = party gender party*gender;
run;
```

predictors of ideology, setting up dummy variables for each predictor with the CLASS statement. This is followed by a MEANS option requesting multiple comparisons across the levels of party. This analysis assumes a lack of interaction. The final GLM statement adds an interaction term to the model.

Table A.8 shows SAS for the repeated measures ANOVA with Table 12.17. Each row of the data provides the opinion responses on the three influences (movies, TV, rock) for a particular subject. You can use PROC REG or else PROC ANOVA for the modeling. The latter applies for "balanced" analyses in which the same number of responses occur at each level of a factor. This model looks like that for a standard two-way ANOVA, except that one effect is the subject effect. The analysis is followed by a multiple comparison of means across the levels of influence type.

Table A.9 shows an alternative way of inputing data for a repeated

measures ANOVA. This table refers to Table 12.21, in which the groups refer to three therapies and the response is weight. Each line identifies the group into which a subject falls, and then lists successively the repeated responses by the subject, labeling them by a name such as RESP1-RESP3 if there are three repeated responses. This table provides the analysis for a between-subjects effect (therapy) and a within-subject effect (the repeated responses on weight). The model statement indicates that the repeated responses are modeled as a function of *therapy* and that the levels at which the repeated measurements occur refer to a variable labeled as *occasion*. The analysis is followed by a Bonferroni multiple comparison of the response means by category of therapy.

This extends to more complex designs. For instance, suppose we had three factors, A and B being between-subjects factors and the repeated measures on a variable y taken at the four levels of a factor C. We could use the SAS code:

```
PROC ANOVA; CLASS  A  B ;
MODEL  Y1 - Y4 = A  B  A*B ; REPEATED C ;
```

You can also conduct repeated measures ANOVA in SAS using PROC MIXED. This is a more advanced procedure that provides additional options for the covariance structure of the random effect (see Section 16.1). There are a variety of options in addition to the sphericity form of the standard analysis, and the results of tests for fixed effects depend on the choice. This procedure, unlike PROC ANOVA or GLM, can use data from subjects that have missing observations. Other advantages of PROC MIXED are that you can use continuous variables in within-subject effects, instead of only classification variables, and you can omit the between–within interaction effects from the model. See Littell et al. (2006) for details.

Chapter 13: Combining Regression and ANOVA: Quantitative and Categorical Predictors

Table A.10 uses SAS to fit analysis of covariance models to Table 13.1. The PLOT statement requests a plot of income by education, with symbols indicating which race each observation has. The first GLM statement fits the analysis of covariance model, assuming no interaction, using a CLASS statement, to provide dummy variables for levels of race. This is followed by a request for adjusted means (also called "least squares means" and abbreviated by SAS as LSMEANS) on the response for the different levels of race, with Bonferroni multiple comparisons of them. The second GLM statement adds an interaction of race and education to the model.

Chapter 14: Model Building with Multiple Regression

Table A.11 shows a variety of analyses for the house sales data. In fitting a multiple regression model, the BACKWARD, FORWARD, STEPWISE, and CP choices for the SELECTION option yield these selection procedures. The P option yields predicted values and the PRESS model diagnostic. The INFLUENCE option yields studentized residuals, leverage values, and measures of influence such as DFFITS and DFBETAS. The PLOT option following the second model statement requests plots of residuals against the predicted values and against size of home. The code sets up an artificial variable *size_2* that is the square of size. Entering it in the model, as in the third regression statement, provides a quadratic regression model.

PROC GENMOD in SAS fits generalized linear models. GENMOD specifies the distribution of the random component in the DIST option ("nor" for normal, "gam" for gamma, "poi" for Poisson, "bin" for binomial) and specifies the link in the LINK option (including "log", "identity", and "logit"). The first GENMOD statement in Table A.11 fits the ordinary bivariate regression model to price and size. This gives the same results as using least squares with PROC REG or GLM. The second GENMOD statement fits the same type of model but instead assumes a gamma distribution for price.

Table A.12 uses GENMOD to fit an exponential regression model to the population growth data of Table 14.8. This model uses the log link.

Chapter 15: Logistic Regression

You can fit logistic regression models either using software for generalized linear models or specialized software for logistic regression. Table A.13 applies PROC GENMOD and PROC LOGISTIC to Table 15.1. In the code, *credit* is a dummy variable indicating whether the subject has a credit card. There would be 100 lines of data for the 100 subjects, with a 1 in the column for credit whenever a subject had a travel credit card.

The GENMOD statement requests the binomial distribution and logit link options, which is logistic regression. PROC LOGISTIC also fits the logistic regression model. These procedures order the levels of the response variable alphanumerically, forming the logit, for instance, as

$$\log\left[\frac{P(Y=0)}{P(Y=1)}\right]$$

The DESCENDING option reverses the order. Following the LOGISTIC model fit, Table A.13 requests predicted probabilities and lower and upper 95% confidence limits for the true probabilities.

For PROC GENMOD and PROC LOGISTIC with binomial models, the response in the model statements can have the form of the number

of successes divided by the number of cases (as data were presented in Table 15.1 and other tables). Table A.14 fits a logistic model with categorical predictors to the death penalty data in Table 15.3. Here, we set up dummy variables for the predictors when we input the data, but you can automatically do this for factors by declaring them in a CLASS statement. The OBSTATS option in GENMOD provides predicted probabilities and their confidence limits, and the RESIDUALS option provides standardized residuals (labeled *StReschi*). In models with multiple predictors, the TYPE3 option in GENMOD provides likelihood-ratio tests for testing the significance of each individual predictor in the model.

You can fit loglinear models using either software for generalized linear models. Table A.15 uses GENMOD to fit model (AC, AM, CM) to the student survey data of Table 15.11. The CLASS statement generates dummy variables for the classification factors. The AM association is represented by $A * M$. The OBSTATS and RESIDUALS options provide expected frequencies (predicted values) and diagnostics, including standardized residuals.

For ordinal responses, PROC LOGISTIC provides ML fitting of the proportional odds version of cumulative logit models. PROC GENMOD fits this model using options DIST=MULTINOMIAL and LINK=CLOGIT. PROC LOGISTIC fits the baseline-category logit model for nominal responses with the option LINK=GLOGIT.

A.0.2 Bibliography

Freund, R. J., and Littell, R. C. (2000). *SAS System for Regression*, 3rd ed. SAS Institute.

Green, S. B., and Salkind, N. J. (2007). *Using SPSS for Windows and Macintosh*, 5th ed. Prentice Hall.

Littell, R., Milliken, G., Stroup, W., Wolfinger, R., and Schabenberger, O. (2006). *SAS for Mixed Models*, 2nd ed. SAS Institute.

Littell, R., Stroup, W., and Freund, R. (2002). *SAS for Linear Models*, 4th ed. SAS Institute.

Norusis, M. (2006). *SPSS 14.0 Guide to Data Analysis*. Prentice Hall.

Schlotzhauer, S. S., and Littell, R. C. (1997). *SAS System for Elementary Statistical Analysis*, 2nd ed. SAS Institute.

Table A.8: SAS for Repeated Measures ANOVA with Table 12.17

```
data repeat;
input subject  $   influ  $   opinion @@;
datalines;
  1  M  -1   1   T   0   1  R  -1
  2  M   1   2   T   0   2  R   0
 ....
 12  M  -1  12   T  -1  12  R  -2
;
proc print;
proc anova;    classes subject influ ;
    model opinion = influ subject ;
    means influ   /   tukey bon;
run;
```

Table A.9: SAS for Two-Way Repeated Measures ANOVA with Table 12.21

```
data repeat2;
input subject $ therapy $ weight1-weight2;
datalines;
  1   CB   80.5   82.2
  2   CB   84.9   85.6
  3   CB   81.5   81.4
  ....
 72    C   89.0   78.8
;
proc anova; class therapy ;
model weight1-weight2 = therapy ;
  repeated occasion  /  short printe;
means therapy / bon ;
run;
```

Table A.10: SAS for Analysis of Covariance Models with Table 13.1

```
data ancova ;
input  income  educ  race $ ;
datalines;
 16  10  black
 18   7  black
 26   9  black
 ....
 56  20  white
;
proc plot;   plot income*educ = race;
proc glm; class race;   model income = educ race / solution;
lsmeans race   adjust=bon ;
proc glm; class race;   model income = educ race educ*race / solution;
run;
```

Table A.11: SAS for Various Analyses Conducted with House Sales Data

```
data housing ;
input  price   size   bed   bath   new;
size_2 = size*size;
datalines;
  279900   2048   4   2   0
  146500    912   2   2   0
  ....
;
proc reg;   model price = size bed bath new   /   selection=backward;
proc reg;   model price = size bath new   /   p influence partial;
plot r.*p. r.*size ;
proc reg;   model price = size size_2 ;
proc genmod;   model price = size  /  dist = nor   link = identity;
proc genmod;   model price = size  /  dist = gam   link = identity ;
run;
```

Table A.12: SAS for Fitting Exponential Regression Model as a Generalized Linear Model to Table 14.8

```
data growth ;
input   decade   popul ;
datalines;
 0   62.95
 1   75.99
 ....
;
proc genmod;  model popul = decade  /  dist = nor  link = log ;
run;
```

Table A.13: SAS for Fitting Logistic Regression Model to Table 15.1

```
data binary ;
input   income   credit ;
datalines;
 12   0
 13   0
 ....
;
proc genmod descending;
    model credit/n = income  /  dist = bin  link = logit ;
proc logistic descending; model credit = income  / influence;
    output  out=predict   p=pi_hat   lower=LCL  upper=UCL;
proc print  data = predict;
run;
```

Table A.14: SAS for Fitting Logistic Model to Table 15.3

```
data death ;
input   vic   def   yes   n ;
datalines;
  1   1   53  467
  1   0   11   48
  0   1    0   16
  0   0    4  143
;
proc genmod; model yes/n = def vic / dist=bin link=logit residuals
obstats type3;
proc logistic;  model yes/n = def  vic;
```

Table A.15: SAS for Fitting Loglinear Models to Table 15.11

```
data drugs ;
input   a   $   c    $    m    $   count @@ ;
datalines;
 yes  yes  yes  911   yes  yes  no  538
 yes  no   yes   44   yes   no  no  456
 no   yes  yes    3    no  yes  no   43
 no   no   yes    2    no   no  no  279
;
proc genmod; class  a  c  m ;
 model count = a  c  m  a*c  a*m  c*m / dist=poi link=log obstats residuals;
run;
```