

CHAPTER 19

MONTECARLO COMMAND

In this chapter, the MONTECARLO command is discussed. The MONTECARLO command is used to set up and carry out a Monte Carlo simulation study.

THE MONTECARLO COMMAND

Following are the options for the MONTECARLO command:

MONTECARLO:		
NAMES =	names of variables;	
NOBSERVATIONS =	number of observations;	
NGROUPS =	number of groups;	1
NREPS =	number of replications;	1
SEED =	random seed for data generation;	0
GENERATE =	scale of dependent variables for data generation;	
CUTPOINTS =	thresholds to be used for categorization of covariates;	
GENCLASSES =	names of categorical latent variables (number of latent classes used for data generation);	
NCSIZES =	number of unique cluster sizes for each group separated by the symbol;	
CSIZES =	number (cluster size) for each group separated by the symbol;	
HAZARDC =	specifies the hazard for the censoring process;	
PATMISS =	missing data patterns and proportion missing for each dependent variable;	
PATPROBS =	proportion for each missing data pattern;	
MISSING =	names of dependent variables that have missing data;	
CENSORED ARE	names, censoring type, and inflation status for censored dependent variables;	
CATEGORICAL ARE	names of binary and ordered categorical (ordinal) dependent variables (model);	
NOMINAL ARE	names of unordered categorical (nominal) dependent variables;	
COUNT ARE	names of count variables (model);	

CHAPTER 19

CLASSES =	names of categorical latent variables (number of latent classes used for model estimation);
AUXILIARY =	names of auxiliary variables (R3STEP); names of auxiliary variables (R); names of auxiliary variables (BCH); names of auxiliary variables (DU3STEP); names of auxiliary variables (DCATEGORICAL); names of auxiliary variables (DE3STEP); names of auxiliary variables (DCONTINUOUS); names of auxiliary variables (E);
SURVIVAL =	names and time intervals for time-to-event variables;
TSCORES =	names, means, and standard deviations of observed variables with information on individually-varying times of observation;
WITHIN =	names of individual-level observed variables;
BETWEEN =	names of cluster-level observed variables;
POPULATION =	name of file containing population parameter values for data generation;
COVERAGE =	name of file containing population parameter values for computing parameter coverage;
STARTING =	name of file containing parameter values for use as starting values for the analysis;
REPSAVE =	numbers of the replications to save data from or ALL ;
SAVE =	name of file in which generated data are stored;
RESULTS =	name of file in which analysis results are stored;
BPARAMETERS =	name of file in which Bayesian posterior parameter values are stored;
LAGGED ARE	names of lagged variables (lag);

The MONTECARLO command is not a required command. When the MONTECARLO command is used, however, the NAMES and NOBSERVATIONS options are required. Default settings are shown in the last column. If the default settings are appropriate for the analysis, nothing besides the required options needs to be specified. Following is a description of the MONTECARLO command.

GENERAL SPECIFICATIONS

The NAMES, NOBSERVATIONS, NGROUPS, NREPS, and SEED options are used to give the basic specifications for a Monte Carlo simulation study. These options are described below.

NAMES

The NAMES option is used to assign names to the variables in the generated data sets. These names are used in the MODEL POPULATION and MODEL commands to specify the data generation and analysis models. As in regular analysis, the list feature can be used to generate variable names. Consider the following specification of the NAMES option:

```
NAMES = y1-y10 x1-x5;
```

which is the same as specifying:

```
NAMES = y1 y2 y3 y4 y5 y6 y7 y8 y9 y10 x1 x2 x3 x4 x5;
```

NOBSERVATIONS

The NOBSERVATIONS option is used to specify the sample size to be used for data generation and in the analysis. The NOBSERVATIONS option is specified as follows:

```
NOBSERVATIONS = 500;
```

where 500 is the sample size to be used for data generation and in the analysis.

If the data being generated are for a multiple group analysis, a sample size must be specified for each group. In multiple group analysis, the NOBSERVATIONS option is specified as follows:

```
NOBSERVATIONS = 500 1000;
```

where a sample size of 500 is used for data generation and in the analysis in the first group and a sample size of 1000 is used for data generation and in the analysis for the second group.

CHAPTER 19

The NOBSERVATION option can be specified as follows when there are many groups:

```
NOBSERVATIONS = 2(1000) 38(500);
```

which specifies 2 groups of size 1000 and 38 groups of size 500.

NGROUPS

The NGROUPS option is used to specify the number of groups to be used for data generation and in the analysis. The NGROUPS option is specified as follows:

```
NGROUPS = 3;
```

where 3 is the number of groups to be used for data generation and in the analysis. The default for the NGROUPS option is 1. The NGROUPS option is not available for TYPE=MIXTURE.

For Monte Carlo studies, the program automatically assigns the label g1 to the first group, g2 to the second group, etc. These labels are used with the MODEL POPULATION and MODEL commands to describe the data generation and analysis models for each group.

The NGROUPS option can be used with TYPE=MIXTURE to specify the number of known classes to be used for data generation and in the analysis. The label %g#1% is assigned to the first known class, %g#2% to the second known class, etc.. These labels are used in the MODEL POPULATION and MODEL commands.

NREPS

The NREPS option is used to specify the number of replications for a Monte Carlo study, that is, the number of samples that are drawn from the specified population and the number of analyses that are carried out. The NREPS option is specified as follows:

```
NREPS = 100;
```

where 100 is the number of samples that are drawn and the number of analyses that are carried out. The default for the NREPS option is 1.

SEED

The SEED option is used to specify the seed to be used for the random draws. The SEED option is specified as follows:

```
SEED = 23458256;
```

where 23458256 is the random seed to be used for the random draws. The default for the SEED option is zero.

DATA GENERATION

The GENERATE, CUTPOINTS, GENCLASSES, NCSIZES, CSIZES, and HAZARDC options are used in conjunction with the MODEL POPULATION command to specify how data are to be generated for a Monte Carlo simulation study. These options are described below.

GENERATE

The GENERATE option is used to specify the scales of the dependent variables for data generation. Variables not mentioned using the GENERATE option are generated as continuous variables. In addition to generating continuous variables which is the default, dependent variables can be generated as censored, binary, ordered categorical (ordinal), unordered categorical (nominal), count variables, and time-to-event variables.

Censored variables can be generated with censoring from above or from below and can be generated with or without inflation. The letters ca followed by a censoring limit in parentheses following a variable name indicate that the variable is censored from above. The letters cb followed by a censoring limit in parentheses following a variable name indicate that the variable is censored from below. The letters cai followed by a censoring limit in parentheses following a variable name indicate that the variable is censored from above with inflation. The letters cbi followed by a censoring limit in parentheses following a variable name indicate that the variable is censored from below with inflation.

For binary and ordered categorical (ordinal) variables using maximum likelihood estimation, the number of thresholds followed by the letter l

for a logistic model or the letter p for a probit model is put in parentheses following the variable name. If no letter is specified, a logistic model is used. The number of thresholds is equal to the number of categories minus one. For binary variables, the logistic model is the same as a two-parameter logistic model. To generate data for a three-parameter logistic model, the number 1 and the letters 3pl are put in parentheses following the variable name. To generate data for a four-parameter logistic model, the number 1 and the letters 4pl are put in parentheses following the variable name. For ordered categorical (ordinal) variables and a logistic model, the data are generated according to a proportional odds model which is the same as a graded response model. To generate data for a generalized partial credit model, the number of thresholds and the letters gpcm are put in parentheses following the variable name. For binary variables and a probit model, the data are generated according to a two-parameter normal ogive model. For ordered categorical (ordinal) variables and a probit model, the data are generated according to a graded response model.

For binary and ordered categorical (ordinal) variables using weighted least squares estimation, only a probit model is allowed. If p is not specified, a probit model is used. The number of thresholds is equal to the number of categories minus one. For binary variables and a probit model, the data are generated according to a two-parameter normal ogive model. For ordered categorical (ordinal) variables and a probit model, the data are generated according to a graded response model.

For unordered categorical (nominal) variables, the letter n followed by the number of intercepts is put in parentheses following the variable name. The number of intercepts is equal to the number of categories minus one because the intercepts are fixed to zero in the last category which is the reference category.

Count variables can be generated for the following six models: Poisson, zero-inflated Poisson, negative binomial, zero-inflated negative binomial, zero-truncated negative binomial, and negative binomial hurdle (Long, 1997; Hilbe, 2011). The letter c or p in parentheses following the variable name indicates that the variable is generated using a Poisson model. The letters ci or pi in parentheses following the variable name indicate that the variable is generated using a zero-inflated Poisson model. The letters nb in parentheses following the variable name indicate that the variable is generated using a negative binomial

model. The letters nbi in parentheses following the variable name indicate that the variable is generated using a zero-inflated negative binomial model. The letters nbt in parentheses following the variable name indicate that the variable is generated using a zero-truncated negative binomial model. The letters nbh in parentheses following the variable name indicate that the variable is generated using a negative binomial hurdle model.

For time-to-event variables in continuous-time survival analysis, the letter s and the number and length of time intervals of the baseline hazard function is put in parentheses following the variable name. When only s is in parentheses, the number of intervals is equal to the number of observations.

The GENERATE option is specified as follows:

```
GENERATE = u1-u2 (1) u3 (1 p) u4 (1 1) u5 u6 (2 p) y1 (ca 1) y2 (cbi 0)
u7 (n 2) u8 (ci) t1 (s 5*1);
```

where the information in parentheses following the variable name or list of variable names defines the scale of the dependent variables for data generation. In this example, the variables u1, u2, u3, and u4 are binary variables with one threshold. Variables u1, u2, and u4 are generated using the logistic model. This is specified by placing nothing or the letter l after the number of thresholds. Variable u3 is generated using the probit model. This is specified by placing the letter p after the number of thresholds. Variables u5 and u6 are three-category ordered categorical (ordinal) variables with two thresholds. The p in parentheses specifies that they are generated using the probit model. Note that if a variable has nothing in parentheses after it, the specification in the next set of parentheses is applied. This means that both u5 and u6 are ordered categorical (ordinal) variables with two thresholds. Variable y1 is a censored variable that is censored from above with a censoring limit of one. Variable y2 is a censored variable with an inflation part that is censored from below with a censoring limit of zero. Variable u7 is a three-category unordered categorical (nominal) variable with two intercepts. Variable u8 is a count variable with an inflation part. Variable t1 is a time-to-event variable. The numbers in parentheses specify that five time intervals of length one will be used for data generation.

In MODEL POPULATION, the inflation part of a censored or count variable is referred to by adding to the name of the censored or count variable the number sign (#) followed by the number 1. The baseline hazard parameters in continuous-time survival analysis are referred to by adding to the name of the time-to-event variable the number sign (#) followed by a number. There are as many baseline hazard parameters as there are time intervals plus one.

CUTPOINTS

The CUTPOINTS option is used to create binary independent variables from the multivariate normal independent variables generated by the program. The CUTPOINTS option specifies the value of the cutpoint to be used in categorizing an independent variable. Following is an example of how the CUTPOINTS option is specified:

```
CUTPOINTS = x1 (0) x2 ( 1);
```

where x1 has a cutpoint of 0 and x2 has a cutpoint of 1. For x1, observations having a value less than or equal to 0 are assigned the value of zero and observations having values greater than 0 are assigned the value of one. Any independent variable not mentioned using the CUTPOINTS option is assumed to be continuous.

In multiple group analysis, the CUTPOINT option is specified as follows where the cutpoints for the groups are separated using the | symbol:

```
CUTPOINTS = x1 (0) x2 ( 1) | x1 (1) x2 (0);
```

where the cutpoints before the | symbol are the cutpoints for group 1 and the cutpoints after the | symbol are the cutpoints for group 2.

GENCLASSES

The GENCLASSES option is used to assign names to the categorical latent variables in the data generation model and to specify the number of latent classes to be used for data generation. This option is used in conjunction with TYPE=MIXTURE. The GENCLASSES option is specified as follows:

```
GENCLASSES = c1 (3) c2 (2) c3 (3);
```

where *c1*, *c2*, and *c3* are the names of the three categorical latent variables in the data generation model. The numbers in parentheses are the number of classes that will be used for each categorical latent variable for data generation. Three classes will be used for data generation for *c1*, two classes for *c2*, and three classes for *c3*.

The letter *b* following the number of classes specifies that the categorical latent variable is a between-level variable. Following is an example of how to specify that a categorical latent variable being generated is a between-level variable:

```
GENCLASSES = cb (2 b);
```

Categorical latent variables that are to be treated as between-level variables in the analysis must be specified as between-level variables using the *BETWEEN* option.

NCSIZES

The *NCSIZES* option is used with *TYPE=TWOLEVEL*, *TYPE=THREELEVEL*, and *TYPE=CROSSCLASSIFIED* to specify the number of unique cluster sizes to be used for data generation. If the data being generated are for a multiple group analysis, the number of unique cluster sizes must be specified for each group.

For *TYPE=TWOLEVEL*, the *NCSIZES* option is specified as follows:

```
NCSIZES = 3;
```

where 3 is the number of unique cluster sizes to be used for data generation.

In multiple group analysis, the *NCSIZES* option is specified as follows where the number of unique cluster sizes for the groups are separated using the | symbol:

```
NCSIZES = 3 | 2;
```

where 3 is the number of unique cluster sizes to be used for data generation for group 1 and 2 is the number of unique cluster sizes for group 2.

CHAPTER 19

For TYPE=THREELEVEL, consider a model where students are nested in classrooms and classrooms are nested in schools. Level 1 is student; level 2 is classroom; and level 3 is school. For TYPE=THREELEVEL, the NCSIZES option is specified as follows:

```
NCSIZES = 3 [2];
```

where the numbers 3 and 2 are the unique cluster sizes to be used for data generation. In this example, 3 is the number of unique cluster sizes for level 3, school, and 2 is the number of unique cluster sizes for level 2, classroom.

In multiple group analysis, the NCSIZES option is specified as follows where the number of unique cluster sizes for the groups are separated using the | symbol:

```
NCSIZES = 3 [2] | 4 [3];
```

where the numbers 3 and 2 are the unique cluster sizes to be used for data generation in group 1 and the numbers 4 and 3 are the unique cluster sizes to be used for data generation in group 2. In this example, 3 is the number of unique cluster sizes for level 3, school, and 2 is the number of unique cluster sizes for level 2, classroom, for group 1 and 4 is the number of unique cluster sizes for level 3, school, and 3 is the number of unique cluster sizes for level 2, classroom, for group 2.

For TYPE=CROSSCLASSIFIED, consider a model where students are nested in schools crossed with neighborhoods. Level 1 is student; level 2a is school; and level 2b is neighborhood. For TYPE=CROSSCLASSIFIED, the NCSIZES option is specified as follows:

```
NCSIZES = 3 [2];
```

where the numbers 3 and 2 are the unique cluster sizes to be used for data generation. In this example, 3 is the number of unique cluster sizes for level 2a, school, and 2 is the number of unique cluster sizes for level 2b, neighborhood.

CSIZES

The **CSIZES** option is used with **TYPE=TWOLEVEL**, **TYPE=THREELEVEL**, and **TYPE=CROSSCLASSIFIED** to specify the number of clusters and the sizes of the clusters to be used for data generation.

For **TYPE=TWOLEVEL**, the **CSIZES** option is specified as follows:

```
CSIZES = 100 (10) 30 (5) 15 (1);
```

where 100 clusters of size 10, 30 clusters of size 5, and 15 clusters of size 1 will be used for data generation.

In multiple group analysis, the **CSIZES** option is specified as follows where the number of clusters and the sizes of the clusters to be used for the groups are separated by the | symbol:

```
CSIZES = 100 (10) 30 (5) 15 (1) | 80 (10) 20 (5);
```

where 100 clusters of size 10, 30 clusters of size 5, and 15 clusters of size 1 will be used for data generation for group 1 and 80 clusters of size 10 and 20 clusters of size 5 will be used for data generation for group 2.

For **TYPE=THREELEVEL**, consider a model where students are nested in classrooms and classrooms are nested in schools. Level 1 is student; level 2 is classroom; and level 3 is school. For **TYPE=THREELEVEL**, the **CSIZES** option is specified as follows:

```
CSIZES = 40 [15(2) 10(5)] 30 [6(8)] 7 [20(2)];
```

where the numbers 40, 30, and 7 are the number of level 3, school, clusters. There are a total of 77 level 3, school, clusters. The 40 level 3, school, clusters are made up of 15 level 2, class, clusters of size two and 10 level 2, class, clusters of size 5 for a total of 3200 observations. The 30 level 3, school, clusters are made up of 6 level 2, class, clusters of size 8 for a total of 1440 observations. The 7 level 3, school, clusters are made up of 20 level 2, class, clusters of size 2 for a total of 280 observations. The total sample size for data generation is 4920.

CHAPTER 19

In multiple group analysis, the CSIZES option is specified as follows where the number of clusters and the sizes of the clusters to be used for the groups are separated by the | symbol:

```
CSIZES = 30 [6(8)] 7 [20(2)] | 40 [5(6)] 20 [4(2)] ;
```

where the numbers 30 and 7 are the number of level 3, school, clusters for group 1 and the numbers 40 and 20 are the number of level 3, school, clusters for group 2. There are a total of 37 level 3, school, clusters for group 1 and 60 level 3, school, clusters for group 2. For group 1, the 30 level 3, school, clusters are made up of 6 level 2, class, clusters of size 8 for a total of 1440 observations. The 7 level 3, school, clusters are made up of 20 level 2, class, clusters of size 2 for a total of 280 observations. For group 1, the total sample size for data generation is 1720. For group 2, the 40 level 3, school, clusters are made up of 5 level 2, class, clusters of size 6 for a total of 1200 observations. The 20 level 3, school, clusters are made up of 4 level 2, class, clusters of size 2 for a total of 160 observations. For group 2, the total sample size for data generation is 1360.

For TYPE=CROSSCLASSIFIED, consider a model where students are nested in schools crossed with neighborhoods. Level 1 is student; level 2a is school; and level 2b is neighborhood. For TYPE=CROSSCLASSIFIED, the CSIZES option is specified as follows:

```
CSIZES = 40 [15(2) 10(5)] 30 [6(8)] 7 [20(2)];
```

where the numbers 40, 30, and 7 are the number of level 2b, neighborhood, clusters. There are a total of 77 level 2b, neighborhood, clusters. The numbers 15, 10, 6, and 20 are the number of level 2a, school, clusters. There are a total of 51 level 2a, school, clusters. The 40 level 2b, neighborhood, clusters are crossed with the 15 level 2a, school, clusters and the 10 level 2a, school, clusters. Each cell of the 40 by 15 cross-classification contains 2 students for a total of 1200 observations. Each cell of the 40 by 10 cross-classification contains 5 students for a total of 2000 observations. The 30 level 2b, neighborhood, clusters are crossed with 6 level 2a, school, clusters. Each cell of the 30 by 6 cross-classification contains 8 students for a total of 1440 observations. The 7 level 2b, neighborhood, clusters are crossed with the 20 level 2a, school, clusters. Each cell of the 7 by 20

cross-classification contains 2 students for a total of 280 observations. The total sample size for data generation is 4920.

HAZARDC

The HAZARDC option is used to specify the hazard for the censoring process in continuous-time survival analysis when time-to-event variables are generated. This information is used to create a censoring indicator variable where zero is not censored and one is right censored. The HAZARDC option is specified as follows:

```
HAZARDC = t1 (.5);
```

where t1 is the name of the time-to-event variable that is generated and .5 is the hazard for censoring.

MISSING DATA GENERATION

The PATMISS, PATPROBS, and MISSING options and the MODEL MISSING command are used to specify how missing data will be generated for a Monte Carlo simulation study. These options are described below. Missing data can be generated using two approaches. In the first approach, the PATMISS and PATPROBS options are used together to generate missing data. In the second approach, the MISSING option is used in conjunction with the MODEL MISSING command to generate missing data. The approaches cannot be used in combination. When generated data are saved, the missing value flag is 999. The PATMISS and PATPROBS options are not available for multiple group analysis. For multiple group analysis, missing data are generated using the MISSING option in conjunction with the MODEL MISSING command. These options are described below.

PATMISS

The PATMISS option is used to specify the missing data patterns and the proportion of data that are missing to be used in missing data generation for each dependent variable in the model. Any variable in the NAMES statement that is not listed in a missing data pattern is assumed to have no missing data for all individuals in that pattern. The PATMISS option is used in conjunction with the PATPROBS option. The PATMISS option is specified as follows:

```
PATMISS =  y1 (.2) y2 (.3) y3 (.1) |
           y2 (.2) y3 (.1) y4 (.3) |
           y3 (.1) y4 (.3);
```

The statement above specifies that there are three missing data patterns which are separated by the | symbol. The number in parentheses following each variable is the probability of missingness to be used for that variable in data generation. In the first pattern, y1, y2, y3 are observed with missingness probabilities of .2, .3, and .1, respectively. In the second pattern, y2, y3, y4 are observed with missingness probabilities of .2, .1, and .3, respectively. In the third pattern, y3 and y4 are observed with missingness probabilities of .1 and .3, respectively. Assuming that the NAMES statement includes variables y1, y2, y3, and y4, individuals in the first pattern have no missing data on variable y4; individuals in the second pattern have no missing data on variable y1; and individuals in the third pattern have no missing data on variables y1 and y2.

PATPROBS

The PATPROBS option is used to specify the proportion of individuals for each missing data pattern to be used in the missing data generation. The PATPROBS option is used in conjunction with the PATMISS option. The proportions are listed in the order of the missing data patterns in the PATMISS option and are separated by the | symbol. The PATPROBS option is specified as follows:

```
PATPROBS = .4 | .3 | .3;
```

where missing data pattern one has probability .40 of being observed in the data being generated, missing data pattern two has probability .30 of being observed in the data being generated, and missing data pattern three has probability .30 of being observed in the data being generated. The missing data pattern probabilities must sum to one.

MISSING

The MISSING option is used to identify the dependent variables for which missing data are to be generated. This option is used in conjunction with the MODEL MISSING command. Missing data are not allowed on

the observed independent variables. The MISSING option is specified as follows:

```
MISSING = y1 y2 u1;
```

which indicates that missing data will be generated for variables y1, y2, and u1. The probabilities of missingness are described using the MODEL MISSING command which is described in Chapter 17.

SCALE OF DEPENDENT VARIABLES FOR ANALYSIS

The CENSORED, CATEGORICAL, NOMINAL, and COUNT options are used to specify the scale of the dependent variables for analysis. These options are described below.

All observed dependent variables are assumed to be measured on a continuous scale for the analysis unless the CENSORED, CATEGORICAL, NOMINAL, and/or COUNT options are specified. The specification of the scale of the dependent variables determines how the variables are treated in the model and its estimation. Independent variables can be binary or continuous. The scales of the independent variables have no impact on the model or its estimation. The distinction between dependent and independent variables is described in the discussion of the MODEL command.

CENSORED

The CENSORED option is used to specify which dependent variables are treated as censored variables in the model and its estimation, whether they are censored from above or below, and whether a censored or censored-inflated model will be estimated.

The CENSORED option is specified as follows for a censored model:

```
CENSORED ARE y1 (a) y2 (b) y3 (a) y4 (b);
```

where y1, y2, y3, y4 are censored dependent variables in the analysis. The letter a in parentheses following the variable name indicates that the variable is censored from above. The letter b in parentheses following the variable name indicates that the variable is censored from below.

The lower and upper censoring limits are determined from the data generation.

The CENSORED option is specified as follows for a censored-inflated model:

```
CENSORED ARE y1 (ai) y2 (bi) y3 (ai) y4 (bi);
```

where y1, y2, y3, y4 are censored dependent variables in the analysis. The letters ai in parentheses following the variable name indicates that the variable is censored from above and that a censored-inflated model will be estimated. The letter bi in parentheses following the variable name indicates that the variable is censored from below and that a censored-inflated model will be estimated. The lower and upper censoring limits are determined from the data generation.

With a censored-inflated model, two variables are considered, a censored variable and an inflation variable. The censored variable takes on values for individuals who are able to assume values of the censoring point and beyond. The inflation variable is a binary latent variable for which the value one denotes that an individual is unable to assume any value except the censoring point. The inflation variable is referred to by adding to the name of the censored variable the number sign (#) followed by the number 1. In the example above, the censored variables available for use in the MODEL command are y1, y2, y3, and y4, and the inflation variables available for use in the MODEL command are y1#1, y2#1, y3#1, and y4#1.

CATEGORICAL

The CATEGORICAL option is used to specify which dependent variables are treated as binary or ordered categorical (ordinal) variables in the model and its estimation and the type of model to be estimated. Both probit and logistic regression models can be estimated for categorical variables. For binary variables, the following IRT models can be estimated: two-parameter normal ogive, two-parameter logistic, three-parameter logistic, and four-parameter logistic. For ordered categorical (ordinal) variables, the following IRT models can be estimated: generalized partial credit with logistic and graded-response with probit (normal ogive) and logistic. For a nominal IRT model, use the NOMINAL option.

For categorical dependent variables, there are as many thresholds as there are categories minus one. The thresholds are referred to in the MODEL command by adding to the variable name the dollar sign (\$) followed by a number. The threshold for a binary variable u1 is referred to as u1\$1. The two thresholds for a three-category variable u2 are referred to as u2\$1 and u2\$2. Ordered categorical dependent variables cannot have more than 10 categories. The number of categories is determined from the data generation.

The CATEGORICAL option is specified as follows:

```
CATEGORICAL ARE u2 u3 u7-u13;
```

where u2, u3, u7, u8, u9, u10, u11, u12, and u13 are binary or ordered categorical dependent variables in the analysis. With weighted least squares and Bayes estimation, a probit model is estimated. For binary variables, this is a two-parameter normal ogive model. For ordered categorical (ordinal) variables, this is a graded response model. With maximum likelihood estimation, a logistic model is estimated as the default. For binary variables, this is a two-parameter logistic model. For ordered categorical (ordinal) variables, this is a proportional odds model which is the same as a graded response model. Probit models can also be estimated with maximum likelihood estimation using the LINK option of the ANALYSIS command.

The CATEGORICAL option for a generalized partial credit model is specified as follows:

```
CATEGORICAL = u1 -u3 (gpcm) u10 (gpcm);
```

where the variables u1, u2, u3, and u10 are ordered categorical (ordinal) variables for which a generalized partial credit model will be estimated. The partial credit model has $c-1$ step parameters for an item with c categories and one slope parameter (Asparouhov & Muthén, 2016). The step parameters are referred to in the same way as thresholds. The first step parameter for a three-category ordered categorical (ordinal) variable u1 is referred to as u1\$1. The second step parameter is referred to as u1\$2.

The CATEGORICAL option for a three-parameter logistic model is specified as follows:

`CATEGORICAL = u1 -u3 (3pl) u10 (3pl);`

where the variables `u1`, `u2`, `u3`, and `u10` are binary variables for which a three-parameter logistic model will be estimated. The guessing parameter cannot be referred to directly. Instead a parameter related to the guessing parameter is referred to (Asparouhov & Muthén, 2016). This parameter is referred to as the second threshold. The first threshold for a binary variable `u1` is referred to as `u1$1`. The second threshold is referred to as `u1$2`.

The `CATEGORICAL` option for a four-parameter logistic model is specified as follows:

`CATEGORICAL = u1 -u3 (4pl) u10 (4pl);`

where the variables `u1`, `u2`, `u3`, and `u10` are binary variables for which a four-parameter logistic model will be estimated. The lower asymptote (guessing) and upper asymptote parameters cannot be referred to directly. Instead a parameter which is related to the lower asymptote (guessing) and a parameter which is related to the upper asymptote parameter are referred to (Asparouhov & Muthén, 2016). The parameter related to the lower asymptote (guessing) parameter is referred to as the second threshold. The parameter related to the upper asymptote parameter is referred to as the third threshold. The first threshold for a binary variable `u1` is referred to as `u1$1`. The second threshold is referred to as `u1$2`. The third threshold is referred to as `u1$3`.

NOMINAL

The `NOMINAL` option is used to specify which dependent variables are treated as unordered categorical (nominal) variables in the model and its estimation. Unordered categorical dependent variables cannot have more than 10 categories. The number of categories is determined from the data generation. The `NOMINAL` option is specified as follows:

`NOMINAL ARE u1 u2 u3 u4;`

where `u1`, `u2`, `u3`, `u4` are unordered categorical dependent variables in the analysis.

For nominal dependent variables, all categories but the last category can be referred to. The last category is the reference category. The categories are referred to in the MODEL command by adding to the variable name the number sign (#) followed by a number. The three categories of a four-category nominal variable are referred to as u1#1, u1#2, and u1#3.

COUNT

The COUNT option is used to specify which dependent variables are treated as count variables in the model and its estimation and the type of model to be estimated. The following models can be estimated for count variables: Poisson, zero-inflated Poisson, negative binomial, zero-inflated negative binomial, zero-truncated negative binomial, and negative binomial hurdle (Long, 1997; Hilbe, 2011). The negative binomial models use the NB-2 variance representation (Hilbe, 2011, p. 63). Count variables may not have negative or non-integer values.

The COUNT option can be specified in two ways for a Poisson model:

COUNT = u1 u2 u3 u4;

or

COUNT = u1 (p) u2 (p) u3 (p) u4 (p);

or using the list function:

COUNT = u1-u4 (p);

The COUNT option can be specified in two ways for a zero-inflated Poisson model:

COUNT = u1-u4 (i);

or

COUNT = u1-u4 (pi);

where u_1 , u_2 , u_3 , and u_4 are count dependent variables in the analysis. The letter i or pi in parentheses following the variable name indicates that a zero-inflated Poisson model will be estimated.

With a zero-inflated Poisson model, two variables are considered, a count variable and an inflation variable. The count variable takes on values for individuals who are able to assume values of zero and above following the Poisson model. The inflation variable is a binary latent variable with one denoting that an individual is unable to assume any value except zero. The inflation variable is referred to by adding to the name of the count variable the number sign (#) followed by the number 1. If the inflation parameter value is estimated at a large negative value corresponding to a probability of zero, the inflation part of the model is not needed.

Following is the specification of the COUNT option for a negative binomial model:

```
COUNT = u1 (nb) u2 (nb) u3 (nb) u4 (nb);
```

or using the list function:

```
COUNT = u1-u4 (nb);
```

With a negative binomial model, a dispersion parameter is estimated. The dispersion parameter is referred to by using the name of the count variable. If the dispersion parameter is estimated at zero, the model is a Poisson model.

Following is the specification of the COUNT option for a zero-inflated negative binomial model:

```
COUNT = u1- u4 (nbi);
```

With a zero-inflated negative binomial model, two variables are considered, a count variable and an inflation variable. The count variable takes on values for individuals who are able to assume values of zero and above following the negative binomial model. The inflation variable is a binary latent variable with one denoting that an individual is unable to assume any value except zero. The inflation variable is referred to by adding to the name of the count variable the number sign

(#) followed by the number 1. If the inflation parameter value is estimated at a large negative value corresponding to a probability of zero, the inflation part of the model is not needed.

Following is the specification of the COUNT option for a zero-truncated negative binomial model:

COUNT = u1-u4 (nbt);

Count variables for the zero-truncated negative binomial model must have values greater than zero.

Following is the specification of the COUNT option for a negative binomial hurdle model:

COUNT = u1-u4 (nbh);

With a negative binomial hurdle model, two variables are considered, a count variable and a hurdle variable. The count variable takes on values for individuals who are able to assume values of one and above following the truncated negative binomial model. The hurdle variable is a binary latent variable with one denoting that an individual is unable to assume any value except zero. The hurdle variable is referred to by adding to the name of the count variable the number sign (#) followed by the number 1.

OPTIONS FOR DATA ANALYSIS

The CLASSES, AUXILIARY, and SURVIVAL options are used only in the analysis. These options are described below.

CLASSES

The CLASSES option is used to assign names to the categorical latent variables in the model and to specify the number of latent classes in the model for each categorical latent variable. This option is required for TYPE=MIXTURE. Between-level categorical latent variables must be identified as between-level variables using the BETWEEN option. The CLASSES option is specified as follows:

CLASSES = c1 (2) c2 (2) c3 (3);

where *c1*, *c2*, and *c3* are the names of the three categorical latent variables in the analysis model. The numbers in parentheses specify the number of classes that will be used for each categorical latent variable in the analysis. The categorical latent variable *c1* has two classes, *c2* has two classes, and *c3* has three classes.

AUXILIARY

Auxiliary variables are variables that are not part of the analysis model. With `TYPE=MIXTURE`, the `AUXILIARY` option is used to automatically carry out the 3-step approach. There are eight settings of the `AUXILIARY` option that automatically carry out the 3-step approach. Two of these settings are used to identify a set of variables not used in the first step of the analysis that are possible covariates in a multinomial logistic regression for a categorical latent variable. The multinomial logistic regression uses all covariates at the same time. Six of the settings are used to identify a set of variables not used in the first step of the analysis for which the equality of means across latent classes will be tested. The equality of means is tested one variable at a time. Only one of these eight settings can be used in an analysis at a time. Only one categorical latent variable is allowed with the 3-step approach. The manual 3-step approach is described in Asparouhov and Muthén (2014a, b).

The two settings that are used to identify a set of variables not used in the first step of the analysis that are possible covariates in a multinomial logistic regression for a categorical latent variable are `R3STEP` (Vermunt, 2010; Asparouhov & Muthén, 2012b) and `R` (Wang et al., 2005). `R3STEP` is preferred. `R` is superseded by `R3STEP` and should be used only for methods research.

The six settings that are used to identify a set of variables not used in the first step of the analysis for which the equality of means across latent classes will be tested are `BCH` (Bakk & Vermunt, 2014), `DU3STEP` (Asparouhov & Muthén, 2012b), `DCAT` (Lanza et al., 2013), `DE3STEP` (Asparouhov & Muthén, 2012b), `DCON` (Lanza et al., 2013), and `E` (Asparouhov, 2007). `BCH` is preferred for continuous distal outcomes. `DU3STEP` should be used only when there are no class changes between the first and last steps. `DCAT` is for categorical distal outcomes. The following settings for continuous distal outcomes, `DE3STEP`, `DCON`, and `E`, should be used only for methods research.

All of the settings are specified in the same way. The setting in parentheses is placed behind the variables on the AUXILIARY statement that will be used as covariates in the multinomial logistic regression or for which the equality of means will be tested. Following is an example of how to specify the R3STEP setting:

```
AUXILIARY = race (R3STEP) ses (R3STEP) x1-x5 (R3STEP);
```

where race, ses, x1, x2, x3, x4, and x5 will be used as covariates in a multinomial logistic regression in a mixture model.

An alternative specification for the eight settings that is convenient when there are several variables that cannot be specified using the list function is:

```
AUXILIARY = (R3STEP) x1 x3 x5 x7 x9;
```

where all variables after R3STEP) will be used as covariates in a multinomial logistic regression in a mixture model.

Following is an example of how to specify more than one setting in the same AUXILIARY statement:

```
AUXILIARY = gender age (BCH) educ ses (BCH) x1-x5 (BCH);
```

where all of the variables on the AUXILIARY statement will be saved if the SAVEDATA command is used, will be available for plots if the PLOT command is used, and tests of equality of means across latent classes will be carried out for the variables age, ses, x1, x2, x3, x4, and x5.

SURVIVAL

The SURVIVAL option is used to identify the variables that contain information about time to event and to provide information about the number and lengths of the time intervals in the baseline hazard function to be used in the analysis. The SURVIVAL option must be used in conjunction with the TIMECENSORED option. The SURVIVAL option can be specified in five ways: the default baseline hazard function, a non-parametric baseline hazard function, a semi-parametric

baseline hazard function, a parametric baseline hazard function, and a constant baseline hazard function.

The `SURVIVAL` option is specified as follows when using the default baseline hazard function:

```
SURVIVAL = t;
```

where `t` is the variable that contains time-to-event information. The default is either a semi-parametric baseline hazard function with ten time intervals or a non-parametric baseline hazard function. The default is a semi-parametric baseline hazard function with ten time intervals for models where `t` is regressed on a continuous latent variable, for multilevel models, and for models that require Monte Carlo numerical integration. In this case, the lengths of the time intervals are selected internally in a non-parametric fashion. For all other models, the default is a non-parametric baseline hazard function as in Cox regression where the number and lengths of the time intervals are taken from the data and the baseline hazard function is saturated.

The `SURVIVAL` option is specified as follows when using a non-parametric baseline hazard function as in Cox regression:

```
SURVIVAL = t (ALL);
```

where `t` is the variable that contains time-to-event information and `ALL` is a keyword that specifies that the number and lengths of the time intervals are taken from the data and the baseline hazard is saturated. It is not recommended to use the keyword `ALL` when the `BASEHAZARD` option of the `ANALYSIS` command is `ON` because it results in a large number of baseline hazard parameters.

The `SURVIVAL` option is specified as follows when using a semi-parametric baseline hazard:

```
SURVIVAL = t (10);
```

where `t` is the variable that contains time-to-event information. The number in parentheses specifies that 10 intervals are used in the analysis where the lengths of the time intervals are selected internally in a non-parametric fashion.

The SURVIVAL option is specified as follows when using a parametric baseline hazard function:

```
SURVIVAL = t (4*5 1*10);
```

where t is the variable that contains time-to-event information. The numbers in parentheses specify that four time intervals of length five and one time interval of length ten are used in the analysis.

The SURVIVAL option is specified as follows when using a constant baseline hazard function:

```
SURVIVAL = t (CONSTANT);
```

where t is the variable that contains time-to-event information and CONSTANT is the keyword that specifies a constant baseline hazard function.

VARIABLES WITH SPECIAL FUNCTIONS FOR DATA GENERATION AND ANALYSIS

The TSCORES, WITHIN, and BETWEEN options are used for both data generation and in the analysis. These options are described below.

TSCORES

The TSCORES option is used in conjunction with TYPE=RANDOM to name and define the variables to be generated that contain information about individually-varying times of observation for the outcome in a longitudinal study. Variables listed in the TSCORES statement can be used only in AT statements in the MODEL and MODEL POPULATION commands to define a growth model. They cannot be used with other statements in the MODEL command. The TSCORES option is specified as follows:

```
TSCORES ARE a1 (0 0) a2 (1 .1) a3 (2 .2) a4 (3 .3);
```

where a1, a2, a3, and a4 are variables to be generated that contain the individually-varying times of observation for an outcome at four time points. The first number in parentheses is the mean of the variable. The second number in parentheses is the standard deviation of the variable.

Each variable is generated using a univariate normal distribution using the mean and standard deviation specified in the TSCORES statement.

WITHIN

The **WITHIN** option is used with **TYPE=TWOLEVEL**, **TYPE=THREELEVEL**, and **TYPE=CROSSCLASSIFIED** to identify the variables in the data set that are measured on the individual level and to specify the levels on which they are modeled. All variables on the **WITHIN** list must be measured on the individual level. An individual-level variable can be modeled on all or some levels.

For **TYPE=TWOLEVEL**, an individual-level variable can be modeled on only the within level or on both the within and between levels. If a variable measured on the individual level is mentioned on the **WITHIN** list, it is modeled on only the within level. It has no variance in the between part of the model. If it is not mentioned on the **WITHIN** list, it is modeled on both the within and between levels. The **WITHIN** option is specified as follows:

WITHIN = y1 y2 x1;

where y1, y2, and x1 are variables measured on the individual level and modeled on only the within level.

For **TYPE=THREELEVEL**, an individual-level variable can be modeled on only level 1, on levels 1 and 2, levels 1 and 3, or on all levels. Consider a model where students are nested in classrooms and classrooms are nested in schools. Level 1 is student; level 2 is classroom; and level 3 is school. If a variable measured on the individual level is mentioned on the **WITHIN** list without a label, it is modeled on only level 1. It has no variance on levels 2 and 3. If it is mentioned on the **WITHIN** list with a level 2 cluster label, it is modeled on levels 1 and 2. It has no variance on level 3. If it is mentioned on the **WITHIN** list with a level 3 cluster label, it is modeled on levels 1 and 3. It has no variance on level 2. If it is not mentioned on the **WITHIN** list, it is modeled on all levels.

Following is an example of how to specify the **WITHIN** option for **TYPE=THREELEVEL**:

WITHIN = y1-y3 (level2) y4-y6 (level3) y7-y9;

In the example, y1, y2, and y3 are variables measured on the individual level and modeled on only level 1. Variables modeled on only level 1 must precede variables modeled on the other levels. Y4, y5, and y6 are variables measured on the individual level and modeled on levels 1 and 2. Y7, y8, and y9 are variables measured on the individual level and modeled on levels 1 and 3.

An alternative specification of the WITHIN option above reverses the order of the level 2 and level 3 variables:

WITHIN = y1-y3 (level3) y7-y9 (level2) y4-y6;

Variables modeled on only level 1 must precede variables modeled on the other levels. Another alternative specification is:

WITHIN = y1-y3;
 WITHIN = (level2) y4-y6;
 WITHIN = (level3) y7-y9;

In this specification, the WITHIN statement for variables modeled on only level 1 must precede the other WITHIN statements. The order of the other WITHIN statements does not matter.

For TYPE=CROSSCLASSIFIED, an individual-level variable can be modeled on only level 1, on levels 1 and 2a, levels 1 and 2b, or on all levels. Consider a model where students are nested in schools crossed with neighborhoods. Level 1 is student; level 2a is school; and level 2b is neighborhood. If a variable measured on the individual level is mentioned on the WITHIN list without a label, it is modeled on only level 1. It has no variance on levels 2a and 2b. If it is mentioned on the WITHIN list with a level 2a cluster label, it is modeled on levels 1 and 2a. It has no variance on level 2b. If it is mentioned on the WITHIN list with a level 2b cluster label, it is modeled on levels 1 and 2b. It has no variance on level 2a. If it is not mentioned on the WITHIN list, it is modeled on all levels.

Following is an example of how to specify the WITHIN option for TYPE=CROSSCLASSIFIED:

WITHIN = y1-y3 (level2a) y4-y6 (level2b) y7-y9;

In the example, y1, y2, and y3 are variables measured on the individual level and modeled on only level 1. Variables modeled on only level 1 must precede variables modeled on the other levels. Y4, y5, and y6 are variables measured on the individual level and modeled on levels 1 and 2a. Y7, y8, and y9 are variables measured on the individual level and modeled on levels 1 and 2b.

BETWEEN

The **BETWEEN** option is used with **TYPE=TWOLEVEL**, **TYPE=THREELEVEL**, and **TYPE=CROSSCLASSIFIED** to identify the variables in the data set that are measured on the cluster level(s) and to specify the level(s) on which they are modeled. All variables on the **BETWEEN** list must be measured on a cluster level. A cluster-level variable can be modeled on all or some cluster levels.

For **TYPE=TWOLEVEL**, a cluster-level variable can be modeled on only the between level. The **BETWEEN** option is specified as follows:

BETWEEN = z1 z2 x1;

where z1, z2, and x1 are variables measured on the cluster level and modeled on the between level. The **BETWEEN** option is also used to identify between-level categorical latent variables with **TYPE=TWOLEVEL MIXTURE**.

For **TYPE=THREELEVEL**, a variable measured on level 2 can be modeled on only level 2 or on levels 2 and 3. A variable measured on level 3 can be modeled on only level 3. Consider a model where students are nested in classrooms and classrooms are nested in schools. Level 1 is student; level 2 is classroom; and level 3 is school. If a variable measured on level 2 is mentioned on the **BETWEEN** list without a label, it is modeled on levels 2 and 3. If a variable measured on level 2 is mentioned on the **BETWEEN** list with a level 2 cluster label, it is modeled on only level 2. It has no variance on level 3. A variable measured on level 3 must be mentioned on the **BETWEEN** list with a level 3 cluster label. Following is an example of how to specify the **BETWEEN** option for **TYPE=THREELEVEL**:

```
BETWEEN = y1-y3 (level2) y4-y6 (level3) y7-y9;
```

In this example, y1, y2, and y3 are cluster-level variables measured on level 2 and modeled on both levels 2 and 3. Variables modeled on both levels 2 and 3 must precede variables modeled on only level 2 or level 3. Y4, y5, and y6 are cluster-level variables measured on level 2 and modeled on level 2. Y7, y8, and y9 are cluster-level variables measured on level 3 and modeled on level 3.

An alternative specification of the BETWEEN option above reverses the order of the level 2 and level 3 variables:

```
BETWEEN = y1-y3 (level3) y7-y9 (level2) y4-y6;
```

Variables modeled on both levels 2 and 3 must precede variables modeled on only level 2 or level 3. Another alternative specification is:

```
BETWEEN = y1-y3;
BETWEEN = (level2) y4-y6;
BETWEEN = (level3) y7-y9;
```

In this specification, the BETWEEN statement for variables modeled on both levels 2 and 3 must precede the other BETWEEN statements. The order of the other BETWEEN statements does not matter.

For TYPE=CROSSCLASSIFIED, a variable measured on level 2a must be mentioned on the BETWEEN list with a level 2a cluster label. It can be modeled on only level 2a. A variable measured on level 2b must be mentioned on the BETWEEN list with a level 2b cluster label. It can be modeled on only level 2b. Consider a model where students are nested in schools crossed with neighborhoods. Level 1 is student; level 2a is school; and level 2b is neighborhood. Following is an example of how to specify the BETWEEN option for TYPE=CROSSCLASSIFIED:

```
BETWEEN = (school) y1-y3 (neighbor) y4-y6;
```

In this example, y1, y2, and y3 are cluster-level variables measured on level 2a, school, and modeled on only level 2a. Y4, y5, and y6 are cluster-level variables measured on level 2b, neighborhood, and modeled on only level 2b.

POPULATION, COVERAGE, AND STARTING VALUES

The POPULATION, COVERAGE, and STARTING options are used as population parameter values for data generation; population parameter values for computing parameter coverage that are printed in the first column of the output labeled Population; and as starting values for the analysis. These values are the parameter estimates obtained from a previous analysis where the parameter estimates are saved using the ESTIMATES option of the SAVEDATA command. These options are described below.

POPULATION

The POPULATION option is used to name the data set that contains the population parameter values to be used in data generation. Following is an example of how the POPULATION option is specified:

```
POPULATION = estimates.dat;
```

where estimates.dat is a file that contains the parameter estimates from a previous analysis of the model that is specified in the MODEL POPULATION command.

COVERAGE

The COVERAGE option is used to name the data set that contains the population parameter values to be used for computing parameter coverage in the Monte Carlo summary. They are printed in the first column of the output labeled Population. Following is an example of how the COVERAGE option is specified:

```
COVERAGE = estimates.dat;
```

where estimates.dat is a file that contains the parameter estimates from a previous analysis of the model that is specified in the MODEL command.

STARTING

The **STARTING** option is used to name the data set that contains the values to be used as starting values for the analysis. Following is an example of how the **STARTING** option is specified:

```
STARTING = estimates.dat;
```

where `estimates.dat` is a file that contains the parameter estimates from a previous analysis of the model that is specified in the **MODEL** command.

SAVING DATA AND RESULTS

The **REPSAVE**, **SAVE**, **RESULTS**, and **BPARAMETERS** options are used to save data and results. These options are described below.

REPSAVE

The **REPSAVE** option is used in conjunction with the **SAVE** option to save some or all of the data sets generated in a Monte Carlo study. The **REPSAVE** option specifies the numbers of the replications for which the data are saved. The keyword **ALL** can be used to save the data from all of the replications. The list function is also available with **REPSAVE**. To save the data from specific replications, **REPSAVE** is specified as follows:

```
REPSAVE = 1 10-15 100;
```

which results in the data from replications 1, 10, 11, 12, 13, 14, 15, and 100 being saved. To save the data from all replications, **REPSAVE** is specified as follows:

```
REPSAVE = ALL;
```

SAVE

The **SAVE** option is used to save data from the first replication for future analysis. It is specified as follows:

```
SAVE = rep1.dat;
```

CHAPTER 19

where rep1.dat is the name of the file in which data from the first replication is saved. The data are saved using a free format.

The SAVE option can be used in conjunction with the REPSAVE option to save data from any or all replications. When the SAVE option is used with the REPSAVE option, it is specified as follows:

```
SAVE = rep*.dat;
```

where the asterisk (*) is replaced by the replication number. For example, if replications 10 and 30 are saved, the data are stored in the files rep10.dat and rep30.dat. A file is also produced that contains the names of all of the data sets. To name this file, the asterisk (*) is replaced by the word list. The file, in this case replist.dat, contains the names of the generated data sets. The variables are not always saved in the order that they appear in the NAMES statement.

RESULTS

The RESULTS option is used to save the analysis results for each replication of the Monte Carlo study in an ASCII file. The results saved include the replication number, parameter estimates, standard errors, and a set of fit statistics. The parameter estimates and standard errors are saved in the order shown in the TECH1 output in free format delimited by a space. The values are saved as E15.8. The RESULTS option is specified as follows:

```
RESULTS = results.sav;
```

where results.sav is the name of the file in which the analysis results for each replication will be saved.

BPARAMETERS

The BPARAMETERS option is used in Bayesian analysis to specify the name of the ASCII file in which the Bayesian posterior parameter values for all iterations are saved. Following is an example of how this option is specified:

```
BPARAMETERS = bayes.dat;
```

where bayes.dat is the name of the file in which the Bayesian posterior parameter values for all iterations will be saved.

LAGGED

The LAGGED option is used in time series analysis to specify the maximum lag to use for an observed variable during model estimation. Following is an example of how to specify the LAGGED option:

```
LAGGED = y (1);
```

where y is the variable in a time series analysis and the number 1 in parentheses is the maximum lag that will be used in model estimation. The lagged variable is referred to in the MODEL command by adding to the name of the variable an ampersand (&) and the number of the lag. The variable y at lag one is referred to as y&1.

Following is an example of how to specify a maximum lag of 2 for a set of variables:

```
LAGGED = y1-y3 (2);
```

where y1, y2, and y3 are variables in a time series analysis and the number 2 in parentheses is the maximum lag that will be used in model estimation. The lagged variables are referred to in the MODEL command by adding to the name of the variable an ampersand (&) and the number of the lag. The variable y1 at lag one is referred to as y1&1. The variable y1 at lag two is referred to as y1&2.

CHAPTER 19