IBM SPSS Complex Samples 22

N

Before using this information and the product it supports, read the information in “Notices” on page 51.

P

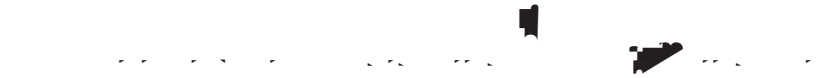
This edition applies to version 22, release 0, modification 0 of IBM SPSS Statistics and to all subsequent releases and modifications until otherwise indicated in new editions.

1. [redacted]

An inherent assumption of analytical procedures in traditional software packages is that the observations in a data file represent a simple random sample from the population of interest. This assumption is

S L . You can specify an optional string label for each stage. This is used in the output to help identify stagewise information.

N : The source variable list has the same content across steps of the Wizard. In other words, variables removed from the source list in a particular step are removed from the list in all steps. Variables returned to the source list appear in the list in all steps.



On the left side of each step in the Sampling Wizard is an outline of all the steps. You can navigate the Wizard by clicking on the name of an enabled step in the outline. Steps are enabled as long as all previous steps are valid—that is, if each previous step has been given the minimum required

This step allows you to specify the number or proportion of units to sample within the current stage. The sample size can be fixed or it can vary across strata. For the purpose of specifying sample size, clusters chosen in previous stages can be used to define strata.

U . You can specify an exact sample size or a proportion of units to sample.

- **V** . A single value is applied to all strata. If **C** is selected as the unit metric, you should enter a positive integer. If **P** is selected, you should enter a non-negative value. Unless sampling with replacement, proportion values should also be no greater than 1.
- **U** . Allows you to enter size values on a per-stratum basis via the Define Unequal Sizes dialog box.
- **R** . Allows you to select a numeric variable that contains size values for strata.

If **P** is selected, you have the option to set lower and upper bounds on the number of units sampled.

The Define Unequal Sizes dialog box allows you to enter sizes on a per-stratum basis.

S . The grid displays the cross-classifications of up to five strata or cluster variables—one stratum/cluster combination per row. Eligible grid variables include all stratification variables from the current and previous stages and all cluster variables from previous stages. Variables can be reordered within the grid or moved to the Exclude list. Enter sizes in the rightmost column. Click **L** or **V** to toggle the display of value labels and data values for stratification and cluster variables in the grid cells. Cells that contain unlabeled values always show values. Click **R** to repopulate the grid with each combination of labeled data values for variables in the grid.

E . To specify sizes for a subset of stratum/cluster combinations, move one or more variables to the

P . The estimated number of units in the population for a given stage. The rootname for the saved variable is *P a S _*.

S . The sampling rate at a given stage. The rootname for the saved variable is *Sam Ra _*.

S . The number of units drawn at a given stage. The rootname for the saved variable is *Sam S _*.

S . The inverse of the inclusion probabilities. The rootname for the saved variable is *Sam W _*.

are available during the current session but are not available in subsequent sessions unless you explicitly save them as data files. Dataset names must adhere to variable naming rules. If an external file or new dataset is specified, the sampling output variables and variables in the active dataset for the selected cases are written.

J . These options let you determine where joint probabilities are written. They are saved to an external IBM SPSS Statistics data file. Joint probabilities are produced if the PPS WOR, PPS Brewer, PPS Sampford, or PPS Murthy method is selected and WR estimation is not specified.

C . If you are constructing your sample one stage at a time, you may want to save the case selection rules to a text file. They are useful for constructing the subframe for subsequent stages.

3. [illegible] A [illegible]

The Analysis Preparation Wizard guides you through the steps for creating or modifying an analysis plan for use with the various Complex Samples analysis procedures. Before using the Wizard, you should have a sample drawn according to a complex design.

Creating a new plan is most useful when you do not have access to the sampling plan file used to draw the sample (recall that the sampling plan contains a default analysis plan). If you do have access to the sampling plan file used to draw the sample, you can use the default analysis plan contained in the sampling plan file or override the default analysis specifications and save your changes to a new file.

[illegible] A [illegible]

1. From the menus choose:
   A [illegible] > C [illegible] S [illegible] > P [illegible] A [illegible] ...
2. Select **C [illegible]**

A

At the left side of each step of the Analysis Wizard is an outline of all the steps. You can navigate the Wizard by clicking on the name of an enabled step in the outline. Steps are enabled as long as all previous steps are valid—that is, as long as each previous step has been given the minimum required specifications for that step. For more information on why a given step may be invalid, see the Help for individual steps.

A

This step allows you to specify an estimation method for the stage.

R ( ). WR estimation does not include a correction for sampling from a finite population (FPC) when estimating the variance under the complex sampling design. You can choose to include or exclude the FPC when estimating the variance under simple random sampling (SRS).

Choosing not to include the FPC for SRS variance estimation is recommended when the analysis weights have been scaled so that they do not add up to the population size. The SRS variance estimate is used in computing statistics like the design effect. WR estimation can be specified only in the final stage of a design; the Wizard will not allow you to add another stage if you select WR estimation.

E OR ( ). Equal WOR estimation includes the finite population correction and assumes that units are sampled with equal probability. Equal WOR can be specified in any stage of a design.

U OR ( ). In addition to using the finite population correction, Unequal WOR accounts for sampling units (usually clusters) selected with unequal probability. This estimation method is available only in the first stage.

A

This step is used to specify inclusion probabilities or population sizes for the current stage. Sizes can be fixed or can vary across strata. For the purpose of specifying sizes, clusters specified in previous stages can be used to define strata. Note that this step is necessary only when Equal WOR is chosen as the Estimation Method.

U . You can specify exact population sizes or the probabilities with which units were sampled.

- **V .** A single value is applied to all strata. If **P S** is selected as the unit metric, you should enter a non-negative integer. If **I P** is selected, you should enter a value between 0 and 1, inclusive.
- **U .** Allows you to enter size values on a per-stratum basis via the Define Unequal Sizes dialog box.
- **R .** Allows you to select a numeric variable that contains size values for strata.

The Define Unequal Sizes dialog box allows you to enter sizes on a per-stratum basis.

S **S** . The grid displays the cross-classifications of up to five strata or cluster variables—one stratum/cluster combination per row. Eligible grid variables include all stratification variables from the current and previous stages and all cluster variables from previous stages. Variables can be reordered within the grid or moved to the Exclude list. Enter sizes in the rightmost column. Click **L** or **V** to toggle the display of value labels and data values for stratification and cluster variables in the grid cells. Cells that contain unlabeled values always show values. Click **R** **S** to repopulate the grid with each combination of labeled data values for variables in the grid.

E . To specify sizes for a subset of stratum/cluster combinations, move one or more variables to the Exclude list. These variables are not used to define sample sizes.

A

This is the last step within each stage, providing a summary of the analysis design specifications through the current stage. From here, you can either proceed to the next stage (creating it if necessary) or save the analysis specifications.

If you cannot add another stage, it is likely because:

- No cluster variable was specified in the Design Variables step.
-

A

1. From the menus choose:
   A > **C** **S** > **P** **A** ...
2. Select **E**, and choose a plan filename to which you will save the analysis plan.
3. Click **N** to continue through the Wizard.
4. Review the analysis plan in the Plan Summary step, and then click **N** .

   Subsequent steps are largely the same as for a new design. For more information, see the Help for individual steps.
5. Navigate to the Finish step, and specify a new name for the edited plan file, or choose to overwrite the existing plan file.

Optionally, you can remove stages from the plan.

A

This step allows you to review the analysis plan and remove stages from the plan.

R **S** . You can remove stages 2 and 3 from a multistage design. Since a plan must have at least one stage, you can edit but not remove stage 1 from the design.

4.

Complex Samples analysis procedures require analysis specifications from an analysis or sample plan file in order to provide valid results.

P . Specify the path of an analysis or sample plan file.

J P . In order to use Unequal WOR estimation for clusters drawn using a PPS WOR method, you need to specify a separate file or an open dataset containing the joint probabilities. This file or dataset is created by the Sampling Wizard during sampling.

5.

The Complex Samples Frequencies procedure produces frequency tables for selected variables and displays univariate statistics. Optionally, you can request statistics by subgroups, defined by one or more categorical variables.

E . Using the Complex Samples Frequencies procedure, you can obtain univariate tabular statistics for vitamin usage among U.S. citizens, based on the results of the National Health Interview Survey (NHIS) and with an appropriate analysis plan for this public-use data.

S . The procedure produces estimates of cell population sizes and table percentages, plus standard errors, confidence intervals, coefficients of variation, design effects, square roots of design effects, cumulative values, and unweighted counts for each estimate. Additionally, chi-square and likelihood-ratio statistics are computed for the test of equal cell proportions.

Complex Samples Frequencies Data Considerations

D . Variables for which frequency tables are produced should be categorical. Subpopulation variables can be string or numeric but should be categorical.

A . The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

Obtaining Complex Samples Frequencies

1. From the menus choose:
   A > C S > F ...
2. Select a plan file. Optionally, select a custom joint probabilities file.
3. Click **C**.
4. In the Complex Samples Frequencies dialog box, select at least one frequency variable.

Optionally, you can specify variables to define subpopulations. Statistics are computed separately for each subpopulation.

C . This group allows you to request estimates of the cell population sizes and table percentages.

S . This group produces statistics associated with the population size or table percentage.

- **S .** The standard error of the estimate.
- **C .** A confidence interval for the estimate, using the specified level.
- **C .** The ratio of the standard error of the estimate to the estimate.
- **U .** The number of units used to compute the estimate.
- **D .** The ratio of the variance of the estimate to the variance obtained by assuming that the sample is a simple random sample. This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **S .** This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **C .** The cumulative estimate through each value of the variable.

T . This produces chi-square and likelihood-ratio tests of the hypothesis that the categories of a variable have equal frequencies. Separate tests are performed for each variable.

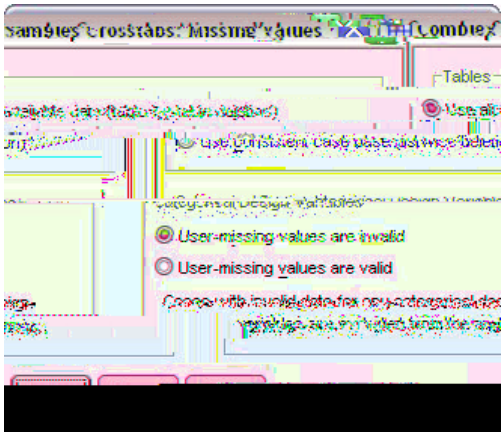


Figure 3. Missing Values dialog box

T . This group determines which cases are used in the analysis.

- **U** . Missing values are determined on a table-by-table basis. Thus, the cases used to compute statistics may vary across frequency or crosstabulation tables.
- **U** . Missing values are determined across all variables. Thus, the cases used to compute statistics are consistent across tables.

C **D** **V** . This group determines whether user-missing values are valid or invalid.

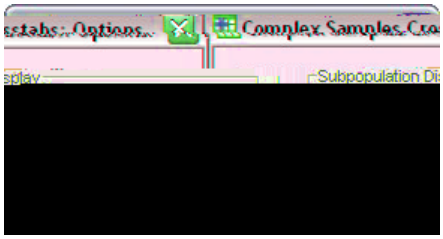


Figure 4. Options dialog box

S **D** . You can choose to have subpopulations displayed in the same table or in separate tables.

6.

The Complex Samples Descriptives procedure displays univariate summary statistics for several variables. Optionally, you can request statistics by subgroups, defined by one or more categorical variables.

E . Using the Complex Samples Descriptives procedure, you can obtain univariate descriptive statistics for the activity levels of U.S. citizens, based on the results of the National Health Interview

S **M** **V** . This group determines which cases are used in the analysis.

- **U** . Missing values are determined on a variable-by-variable basis, thus the cases used to compute statistics may vary across measure variables.
- **E** . Missing values are determined across all variables, thus the cases used to compute statistics are consistent.

C **D** **V** . This group determines whether user-missing values are valid or invalid.

O

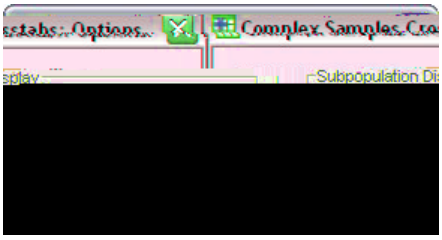


Figure 5. Options dialog box

S **D** . You can choose to have subpopulations displayed in the same table or in separate tables.

7.

The Complex Samples Crosstabs procedure produces crosstabulation tables for pairs of selected variables and displays two-way statistics. Optionally, you can request statistics by subgroups, defined by one or more categorical variables.

E . Using the Complex Samples Crosstabs procedure, you can obtain cross-classification statistics for smoking frequency by vitamin usage of U.S. citizens, based on the results of the National Health Interview Survey (NHIS) and with an appropriate analysis plan for this public-use data.

S . The procedure produces estimates of cell population sizes and row, column, and table percentages, plus standard errors, confidence intervals, coefficients of variation, expected values, design effects, square roots of design effects, residuals, adjusted residuals, and unweighted counts for each estimate. The odds ratio, relative risk, and risk difference are computed for 2-by-2 tables. Additionally, Pearson and likelihood-ratio statistics are computed for the test of independence of the row and column variables.

Complex Samples Crosstabs Data Considerations

D . Row and column variables should be categorical. Subpopulation variables can be string or numeric but should be categorical.

A . The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

Obtaining Complex Samples Crosstabs

1. From the menus choose:
   A > C S > C ...
2. Select a plan file. Optionally, select a custom joint probabilities file.
3. Click **C .**
4. In the Complex Samples Crosstabs dialog box, select at least one row variable and one column variable.

Optionally, you can specify variables to define subpopulations. Statistics are computed separately for each subpopulation.

C . This group allows you to request estimates of the cell population size and row, column, and table percentages.

S . This group produces statistics associated with the population size and row, column, and table percentages.

- **S .** The standard error of the estimate.
- **C .** A confidence interval for the estimate, using the specified level.
- **C .** The ratio of the standard error of the estimate to the estimate.
- **E .** The expected value of the estimate, under the hypothesis of independence of the row and column variable.
- **U .** The number of units used to compute the estimate.

- **D** . The ratio of the variance of the estimate to the variance obtained by assuming that the sample is a simple random sample. This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **S** . This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **R** . The expected value is the number of cases that you would expect in the cell if there were no relationship between the two variables. A positive residual indicates that there are more cases in the cell than there would be if the row and column variables were independent.
- **A** . The residual for a cell (observed minus expected value) divided by an estimate of its standard error. The resulting standardized residual is expressed in standard deviation units above or below the mean.

S **2- -2 T** . This group produces statistics for tables in which the row and column variable each have two categories. Each is a measure of the strength of the association between the presence of a factor and the occurrence of an event.

- **O** . The odds ratio can be used as an estimate of relative risk when the occurrence of the factor is rare.
- **R** . The ratio of the risk of an event in the presence of the factor to the risk of the event in the absence of the factor.
- **R** . The difference between the risk of an event in the presence of the factor and the risk of the event in the absence of the factor.

T . This produces chi-square and likelihood-ratio tests of the hypothesis that a row and column variable are independent. Separate tests are performed for each pair of variables.

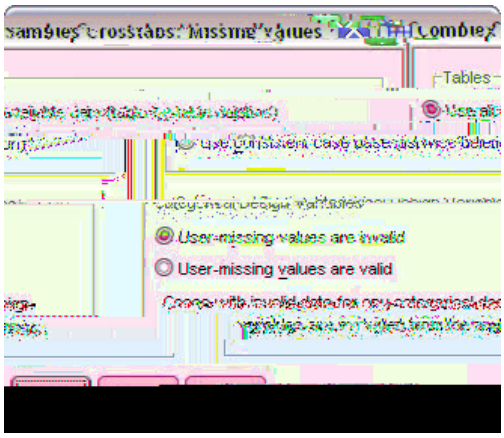


Figure 6. Missing Values dialog box

T . This group determines which cases are used in the analysis.

- **U** . Missing values are determined on a table-by-table basis. Thus, the cases used to compute statistics may vary across frequency or crosstabulation tables.
- **U** . Missing values are determined across all variables. Thus, the cases used to compute statistics are consistent across tables.

C **D** **V** . This group determines whether user-missing values are valid or invalid.

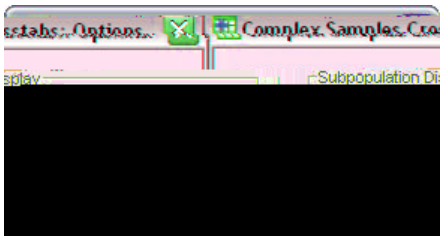


Figure 7. Options dialog box

S **D** . You can choose to have subpopulations displayed in the same table or in separate tables.

8. [illegible] [illegible] R [illegible]

The Complex Samples Ratios procedure displays univariate summary statistics for ratios of variables. Optionally, you can request statistics by subgroups, defined by one or more categorical variables.

E [illegible]. Using the Complex Samples Ratios procedure, you can obtain descriptive statistics for the ratio of current property value to last assessed value, based on the results of a statewide survey carried out according to a complex design and with an appropriate analysis plan for the data.

S [illegible]. The procedure produces ratio estimates, tests, standard errors, confidence intervals, coefficients of variation, unweighted counts, population sizes, design effects, and square roots of design effects.

Complex Samples Ratios Data Considerations

D . Numerators and denominators should be positive-valued scale variables. Subpopulation variables can be string or numeric but should be categorical.

A [illegible].

- **U** . Missing values are determined on a ratio-by-ratio basis. Thus, the cases used to compute statistics may vary across numerator-denominator pairs.
- **E** . Missing values are determined across all variables. Thus, the cases used to compute statistics are consistent.

C **D** **V** . This group determines whether user-missing values are valid or invalid.

O

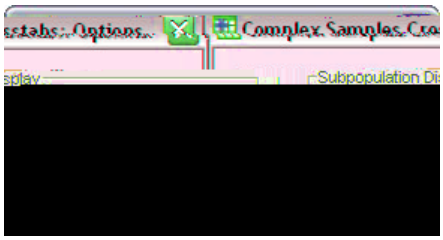


Figure 8. Options dialog box

S **D** . You can choose to have subpopulations displayed in the same table or in separate tables.

[illegible]

The Complex Samples General Linear Model (CSGLM) procedure performs linear regression analysis, as well as analysis of variance and covariance, for samples drawn by complex sampling methods. Optionally, you can request analyses for a subpopulation.

E [illegible]. A grocery store chain surveyed a set of customers concerning their purchasing habits, according to a complex design. Given the survey results and how much each customer spent in the previous month, the store wants to see if the frequency with which customers shop is related to the amount they spend in a month, controlling for the gender of the customer and incorporating the sampling design.

S [illegible]. The procedure produces estimates, standard errors, confidence intervals, tests, design effects, and square roots of design effects for model parameters, as well as the correlations and covariances between parameter estimates. Measures of model fit and descriptive statistics for the dependent and independent variables are also available. Additionally, you can request estimated marginal means for levels of model factors and factor interactions.

Complex Samples General Linear Model Data Considerations

D . The dependent variable is quantitative. Factors are categorical. Covariates are quantitative variables that are related to the dependent variable. Subpopulation variables can be string or numeric but should be categorical.

A [illegible]. The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

Obtaining a Complex Samples General Linear Model

1. From the menus choose:
   A [illegible] > C [illegible] S [illegible] > G [illegible] L [illegible] M [illegible]...
2. Select a plan file. Optionally, select a custom joint probabilities file.
3. Click **C [illegible]**.
4. In the Complex Samples General Linear Model dialog box, select a dependent variable.

Optionally, you can:

- Select variables for factors and covariates, as appropriate for your data.
- Specify a variable to define a subpopulation. The analysis is performed only for the selected category of the subpopulation variable.

[illegible]

S [illegible] M [illegible] E [illegible]. By default, the procedure builds a main-effects model using the factors and covariates specified in the main dialog box. Alternatively, you can build a custom model that includes interaction effects and nested terms.

Non-Nested Terms

For the selected factors and covariates:

I [illegible]. Creates the highest-level interaction term for all selected variables.

M [illegible]. Creates a main-effects term for each variable selected.

A 2- . Creates all possible two-way interactions of the selected variables.

A 3- . Creates all possible three-way interactions of the selected variables.

A 4- . Creates all possible four-way interactions of the selected variables.

A 5- . Creates all possible five-way interactions of the selected variables.

Nested Terms

You can build nested terms for your model in this procedure. Nested terms are useful for modeling the effect of a factor or covariate whose values do not interact with the levels of another factor. For example, a grocery store chain may follow the spending habits of its customers at several store locations. Since each customer frequents only one of these locations, the *C m* effect can be said to be the *S a* effect.

Additionally, you can include interaction effects, such as polynomial terms involving the same covariate, or add multiple levels of nesting to the nested term.

. Nested terms have the following restrictions:

- All factors within an interaction must be unique. Thus, if *A* is a factor, then specifying *A*A* is invalid.
- All factors within a nested effect must be unique. Thus, if *A* is a factor, then specifying *A(A)* is invalid.
- No effect can be nested within a covariate. Thus, if *A* is a factor and *X* is a covariate, then specifying *A(X)* is invalid.

I The intercept is usually included in the model. If you can assume the data pass through the origin, you can exclude the intercept. Even if you include the intercept in the model, you can choose to suppress statistics related to it.

M P . This group allows you to control the display of statistics related to the model parameters.

- **E** . Displays estimates of the coefficients.
- **S** . Displays the standard error for each coefficient estimate.
- **C** . Displays a confidence interval for each coefficient estimate. The confidence level for the interval is set in the Options dialog box.
- **T** . Displays a test of each coefficient estimate. The null hypothesis for each test is that the value of the coefficient is 0.
- **C** . Displays an estimate of the covariance matrix for the model coefficients.
- **C** . Displays an estimate of the correlation matrix for the model coefficients.
- **D** . The ratio of the variance of the estimate to the variance obtained by assuming that the sample is a simple random sample. This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **S** . This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.

M . Displays R^2 and root mean squared error statistics.

P . Displays summary information about the dependent variable, covariates, and factors.

S [illegible]. Displays summary information about the sample, including the unweighted count and the population size.

T [illegible] S [illegible]. This group allows you to select the type of statistic used for testing hypotheses. You can choose between *F*, adjusted *F*, chi-square, and adjusted chi-square.

S [illegible] D [illegible] F [illegible]. This group gives you control over the sampling design degrees of freedom used to compute values for all test statistics. If based on the sampling design, the value is the difference between the number of primary sampling units and the number of strata in the first stage of sampling. Alternatively, you can set a custom degrees of freedom by specifying a positive integer.

A [illegible] M [illegible] C [illegible]. When performing hypothesis tests with multiple contrasts, the overall significance level can be adjusted from the significance levels for the included contrasts. This group allows you to choose the adjustment method.

- **L [illegible]**. This method does not control the overall probability of rejecting the hypotheses that some linear contrasts are different from the null hypothesis values.
- *S a S a* . This is a sequentially step-down rejective Sidak procedure that is much less conservative in terms of rejecting individual hypotheses but maintains the same overall significance level.
- *S a B* . This is a sequentially step-down rejective Bonferroni procedure that is much less conservative in terms of rejecting individual hypotheses but maintains the same overall significance level.
- *S a* . This method provides tighter bounds than the Bonferroni approach.
- *B* . This method adjusts the observed significance level for the fact that multiple contrasts are being tested.

S V . This group allows you to save the model predicted values and residuals as new variables in the working file.

E IBM SPSS S . Writes a dataset in IBM SPSS Statistics format containing the parameter correlation or covariance matrix with parameter estimates, standard errors, significance values, and degrees of freedom. The order of variables in the matrix file is as follows.

- . Takes values (and value labels), COV (Covariances), CORR (Correlations), EST (Parameter estimates), SE (Standard errors), SIG (Significance levels), and DF (Sampling design degrees of

10. [illegible] [illegible] [illegible] R [illegible]

The Complex Samples Logistic Regression procedure performs logistic regression analysis on a binary or multinomial dependent variable for samples drawn by complex sampling methods. Optionally, you can request analyses for a subpopulation.

E [illegible]. A loan officer has collected past records of customers given loans at several different branches, according to a complex design. While incorporating the sample design, the officer wants to see if the probability with which a customer defaults is related to age, employment history, and amount of credit debt.

S [illegible]. The procedure produces estimates, exponentiated estimates, standard errors, confidence intervals, tests, design effects, and square roots of design effects for model parameters, as well as the correlations and covariances between parameter estimates. Pseudo R^2 statistics, classification tables, and descriptive statistics for the dependent and independent variables are also available.

Complex Samples Logistic Regression Data Considerations

D . The dependent variable is categorical. Factors are categorical. Covariates are quantitative variables that are related to the dependent variable. Subpopulation variables can be string or numeric but should be categorical.

A [illegible]. The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

Obtaining Complex Samples Logistic Regression

1. From the menus choose:
   A [illegible] > C [illegible] S [illegible] > L [illegible] R [illegible] ...
2. Select a plan file. Optionally, select a custom joint probabilities file.
3. Click **C [illegible]**.
4. In the Complex Samples Logistic Regression dialog box, select a dependent variable.

Optionally, you can:

- Select variables for factors and covariates, as appropriate for your data.
- Specify a variable to define a subpopulation. The analysis is performed only for the selected category of the subpopulation variable.

[illegible] [illegible] [illegible] R [illegible] R [illegible]

By default, the Complex Samples Logistic Regression procedure makes the highest-valued category the reference category. This dialog box allows you to specify the highest value, the lowest value, or a custom category as the reference category.

[illegible] [illegible] [illegible] R [illegible]

S [illegible] M [illegible] E [illegible]. By default, the procedure builds a main-effects model using the factors and covariates specified in the main dialog box. Alternatively, you can build a custom model that includes interaction effects and nested terms.

Non-Nested Terms

For the selected factors and covariates:

I . Creates the highest-level interaction term for all selected variables.

M . Creates a main-effects term for each variable selected.

A 2- . Creates all possible two-way interactions of the selected variables.

A 3- . Creates all possible three-way interactions of the selected variables.

A 4- . Creates all possible four-way interactions of the selected variables.

A 5- . Creates all possible five-way interactions of the selected variables.

Nested Terms

You can build nested terms for your model in this procedure. Nested terms are useful for modeling the effect of a factor or covariate whose values do not interact with the levels of another factor. For example, a grocery store chain may follow the spending habits of its customers at several store locations. Since each customer frequents only one of these locations, the *C m* effect can be said to be the *S a* effect.

Additionally, you can include interaction effects, such as polynomial terms involving the same covariate, or add multiple levels of nesting to the nested term.

B . Nested terms have the following restrictions:

- All factors within an interaction must be unique. Thus, if *A* is a factor, then specifying *A*A* is invalid.
- All factors within a nested effect must be unique. Thus, if *A* is a factor, then specifying *A(A)* is invalid.
- No effect can be nested within a covariate. Thus, if *A* is a factor and *X* is a covariate, then specifying *A(X)* is invalid.

I The intercept is usually included in the model. If you can assume the data pass through the origin, you can exclude the intercept. Even if you include the intercept in the model, you can choose to suppress statistics related to it.

M . Controls the display of statistics that measure the overall model performance.

- **P R-** . The R^2 statistic from linear regression does not have an exact counterpart among logistic regression models. There are, instead, multiple measures that attempt to mimic the properties of the R^2 statistic.
- **C** . Displays the tabulated cross-classifications of the observed category by the model-predicted category on the dependent variable.

P . This group allows you to control the display of statistics related to the model parameters.

- **E** . Displays estimates of the coefficients.
- **E** . Displays the base of the natural logarithm raised to the power of the estimates of the coefficients. While the estimate has nice properties for statistical testing, the exponentiated estimate, or exp(B), is easier to interpret.
- **S** . Displays the standard error for each coefficient estimate.
- **C** . Displays a confidence interval for each coefficient estimate. The confidence level for the interval is set in the Options dialog box.

C [illegible]. For each selected covariate, displays the ratio of the odds at the covariate's mean value plus the specified units of change to the odds at the mean.

When computing odds ratios for a factor or covariate, the procedure fixes all other factors at their highest levels and all other covariates at their means. If a factor or covariate interacts with other predictors in the model, then the odds ratios depend not only on the change in the specified variable but also on the values of the variables with which it interacts. If a specified covariate interacts with itself in the model (for example, *a* **a*), then the odds ratios depend on both the change in the covariate and the value of the covariate.

S [illegible] V [illegible]. This group allows you to save the model-predicted category and predicted probabilities as new variables in the active dataset.

- **D** . Displays parameter estimates and statistics at every iterations beginning with the 0th iteration (the initial estimates). If you choose to print the iteration history, the last iteration is always printed regardless of the value of .

U -M V . All design variables, as well as the dependent variable and any covariates, must have valid data. Cases with invalid data for any of these variables are deleted from the analysis. These controls allow you to decide whether user-missing values are treated as valid among the strata, cluster, subpopulation, and factor variables.

C I . This is the confidence interval level for coefficient estimates, exponentiated coefficient estimates, and odds ratios. Specify a value greater than or equal to 50 and less than 100.

O A

The command syntax language also allows you to:

- Specify custom tests of effects versus a linear combination of effects or a value (using the CUSTOM subcommand).
- Fix values of other model variables when computing odds ratios for factors and covariates (using the ODDSRATIOS subcommand).
- Specify a tolerance value for checking singularity (using the CRITERIA subcommand).
- Create user-specified names for saved variables (using the SAVE subcommand).
-

11. [illegible] [illegible] O [illegible] R [illegible]

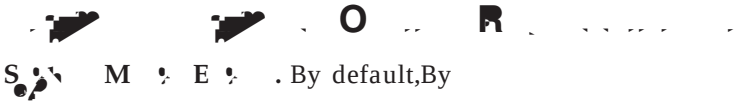
The Complex Samples Ordinal Regression procedure performs regression analysis on a binary or ordinal dependent variable for samples drawn by complex sampling methods. Optionally, you can request analyses for a subpopulation.

E [illegible]. Representatives considering a bill before the legislature are interested in whether there is public support for the bill and how support for the bill is related to voter demographics. Pollsters design and conduct interviews according to a complex sampling design. Using Complex Samples Ordinal Regression, you can fit a model for the level of support for the bill based upon voter demographics.

Complex Samples Ordinal Regression Data Considerations

D . The dependent variable is ordinal. Factors are categorical. Covariates are quantitative variables that are related to the dependent variable. Subpopulation variables can be string or numeric but should be categorical.

A [illegible] . The cases in the data file represent a sample from a complex design that should be



S M E . By default,By

- **C [illegible]**. Displays a confidence interval for each coefficient estimate. The confidence level for the interval is set in the Options dialog box.

- *B* . This method adjusts the observed significance level for the fact that multiple contrasts are being tested.

O R O R

The Odds Ratios dialog box allows you to display the model-estimated cumulative odds ratios for specified factors and covariates. This feature is only available for models using the Logit link function. A single cumulative odds ratio is computed for all categories of the dependent variable except the last; the proportional odds model postulates that they are all equal.

F .

12. [illegible] [illegible] [illegible] R [illegible]

The Complex Samples Cox Regression procedure performs survival analysis for samples drawn by complex sampling methods. Optionally, you can request analyses for a subpopulation.

E [illegible] .

Table 1. Data structure for incorporating piecewise-constant time-dependent predictors.

Pa ID	*E m*	*S a*	*P a a a*	*P m a*
1	5	Heart Attack	No	No
1	7	Hemorrhaging	Yes	No
1	8	Died	Yes	Yes
2	24	Died	No	No
3	8	Heart Attack	No	No
3	15	Died	Yes	No

A . The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

are different at different time points. In such cases, you need to specify time-dependent predictors. See the topic “Define Time-Dependent Predictor” for more information. Time-dependent predictors can be selected as factors or covariates.

The Define Time-Dependent Predictor dialog box allows you to create a predictor that is dependent upon the built-in time variable, *T_*. You can use this variable to define time-dependent covariates in two general ways:

- If you want to estimate an extended Cox regression model that allows nonproportional hazards, you can do so by defining your time-dependent predictor as a function of the time variable *T_* and the covariate in question. A common example would be the simple product of the time variable and the predictor, but more complex functions can be specified as well.
- Some variables may have different values at different time periods but aren't systematically related to time. In such cases, you need to define a , which can be done using logical expressions. Logical expressions take the value 1 if true and 0 if false. Using a series of logical expressions, you can create your time-dependent predictor from a set of measurements. For example, if you have blood pressure measured once a week for the four weeks of your study (identified as *BP1* to *BP4*), you can define your time-dependent predictor as (*T_* < 1) * *BP1* + (*T_* >= 1 & *T_* < 2) * *BP2* + (*T_* >= 2 & *T_* < 3) * *BP3* + (*T_* >= 3 & *T_* < 4) * *BP4*. Notice that exactly one of the terms in parentheses will be equal to 1 for any given case and the rest will all equal 0. In other words, this function means that if time is less than one week, use *BP1*; if it is more than one week but less than two weeks, use *BP2*; and so on.

  N : If your segmented, time-dependent predictor is constant within segments, as in the blood pressure example given above, it may be easier for you to specify the piecewise-constant, time-dependent predictor by splitting subjects across multiple cases. See the discussion on Subject Identifiers in Chapter 12, “Complex Samples Cox Regression,” on page 41 for more information.

In the Define Time-Dependent Predictor dialog box, you can use the function-building controls to build the expression for the time-dependent covariate, or you can enter it directly in the Numeric Expression text area. Note that string constants must be enclosed in quotation marks or apostrophes, and numeric constants must be typed in American format, with the dot as the decimal delimiter. The resulting variable is given the name you specify and should be included as a factor or covariate on the Predictors tab.

B S . A separate baseline hazard and survival function is computed for each value of this variable, while a single set of model coefficients is estimated across strata.

S V . Specify a variable to define a subpopulation. The analysis is performed only for the selected category of the subpopulation variable.

S M E . By default, the procedure builds a main-effects model using the factors and covariates specified in the main dialog box. Alternatively, you can build a custom model that includes interaction effects and nested terms.

Non-Nested Terms

For the selected factors and covariates:

I . Creates the highest-level interaction term for all selected variables.

M . Creates a main-effects term for each variable selected.

A 2- . Creates all possible two-way interactions of the selected variables.

A 3- . Creates all possible three-way interactions of the selected variables.

A 4- . Creates all possible four-way interactions of the selected variables.

A 5- . Creates all possible five-way interactions of the selected variables.

Nested Terms

You can build nested terms for your model in this procedure. Nested terms are useful for modeling the effect of a factor or covariate whose values do not interact with the levels of another factor. For example, a grocery store chain may follow the spending habits of its customers at several store locations. Since each customer frequents only one of these locations, the *C m* effect can be said to be the *S a* effect.

Additionally, you can include interaction effects, such as polynomial terms involving the same covariate, or add multiple levels of nesting to the nested term.

. Nested terms have the following restrictions:

- All factors within an interaction must be unique. Thus, if *A* is a factor, then specifying *A*A* is invalid.
- All factors within a nested effect must be unique. Thus, if *A* is a factor, then specifying *A(A)* is invalid.
- No effect can be nested within a covariate. Thus, if *A* is a factor and *X* is a covariate, then specifying *A(X)* is invalid.

S . Displays summary information about the sample, including the unweighted count and the population size.

E . Displays summary information about the number and percentage of censored cases.

R . Displays number of events and number at risk for each event time in each baseline stratum.

P . This group allows you to control the display of statistics related to the model parameters.

- **E** . Displays estimates of the coefficients.
- **E** . Displays the base of the natural logarithm raised to the power of the estimates of the coefficients. While the estimate has nice properties for statistical testing, the exponentiated estimate, or exp(B), is easier to interpret.
- **S** . Displays the standard error for each coefficient estimate.
- **C** . Displays a confidence interval for each coefficient estimate. The confidence level for the interval is set in the Options dialog box.
- **-** . Displays a test of each coefficient estimate. The null hypothesis for each test is that the value of the coefficient is 0.
- **C** . Displays an estimate of the covariance matrix for the model coefficients.
- **C** . Displays an estimate of the correlation matrix for the model coefficients.
- **D** . The ratio of the variance of the estimate to the variance obtained by assuming that the sample is a simple random sample. This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.

- **S** . This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.

M A . This group allows you to produce a test of the proportional hazards assumption.

- *S a B*

subject. For Schoenfeld's residual, the aggregated version is the same as that of the non-aggregated version because Schoenfeld's residual is only defined for uncensored cases. These residuals are only available when a subject identifier is specified on the Time and Event tab.

N : **S** : **V** : . Automatic name generation ensures that you keep all your work. Custom names allow you to discard/replace results from previous runs without first deleting the saved variables in the Data Editor.

E : **IBM SPSS S** . Writes a dataset in IBM SPSS Statistics format containing the parameter correlation or covariance matrix with parameter estimates, standard errors, significance values, and degrees of freedom. The order of variables in the matrix file is as follows.

- . Takes values (and value labels), COV (Covariances), CORR (Correlations), EST (Parameter estimates), SE (Standard errors), SIG (Significance levels), and DF (Sampling design degrees of freedom). There is a separate case with row type COV (or CORR) for each model parameter, plus a separate case for each of the other row types.
- : . Takes values P1, P2, ..., corresponding to an ordered list of all model parameters, for row types COV or CORR, with value labels corresponding to the parameter strings shown in the parameter estimates table. The cells are blank for other row types.
- **P1, P2, ...** These variables correspond to an ordered list of all model parameters, with variable labels corresponding to the parameter strings shown in the parameter estimates table, and take values according to the row type. For redundant parameters, all covariances are set to zero; correlations are set to the system-missing value; all parameter estimates are set at zero; and all standard errors, significance levels, and residual degrees of freedom are set to the system-missing value.

N : This file is not immediately usable for further analyses in other procedures that read a matrix file unless those procedures accept all the row types exported here.

E : **IBM SPSS S** . Writes a dataset in IBM SPSS Statistics format containing the survival function; standard error of the survival function; upper and lower bounds of the confidence interval of the survival function; and the cumulative hazards function for each failure or event time, evaluated at the baseline and at the predictor patterns specified on the Plot tab. The order of variables in the matrix file is as follows.

- **B** : . Separate survival tables are produced for each value of the strata variable.
- **S** : . The event time; a separate case is created for each unique event time.
- **S 0, LCL S 0, UCL S 0.** Baseline survival function and the upper and lower bounds of its confidence interval.
- **S R, LCL S R, UCL S R.** Survival function evaluated at the "reference" pattern (see the pattern values table in the output) and the upper and lower bounds of its confidence interval.
- **S #.#, LCL S #.#, UCL S #.#, ...** Survival function evaluated at each of the predictor patterns specified on the Plots tab and the upper and lower bounds of their confidence intervals. See the pattern values table in the output to match patterns with the number #.#.
- **H 0, LCL H 0, UCL H 0.** Baseline cumulative hazard function and the upper and lower bounds of its confidence interval.
- **H R, LCL H R, UCL H R.** Cumulative hazard function evaluated at the "reference" pattern (see the pattern values table in the output) and the upper and lower bounds of its confidence interval.
- **H #.#, LCL H #.#, UCL H #.#, ...** Cumulative hazard function evaluated at each of the predictor patterns specified on the Plots tab and the upper and lower bounds of their confidence intervals. See the pattern values table in the output to match patterns with the number #.#.

E [illegible] ML. Saves all information needed to predict the survival function, including parameter estimates and the baseline survival function, in XML (PMML) format. You can use this model file to apply the model information to other data files for scoring purposes.

O [illegible]

E [illegible]. These controls specify criteria for estimation of regression coefficients.

- **M [illegible] I [illegible].** The maximum number of iterations the algorithm will execute. Specify a non-negative integer.
- **M [illegible] S [illegible] H [illegible].** At each iteration, the step size is reduced by a factor of 0.5 until the log-likelihood increases or maximum step-halving is reached. Specify a positive integer.
- **[illegible]** When selected, the algorithm stops after an iteration in which the absolute or relative change in the parameter estimates is less than the value specified, which must be positive.
- **[illegible]** When selected, the algorithm stops after an iteration in which the absolute or relative change in the log-likelihood function is less than the value specified, which must be positive.
- **D [illegible].** Displays the iteration history for the parameter estimates and pseudo log-likelihood and prints the last evaluation of the change in parameter estimates and pseudo log-likelihood. The iteration history table prints every iterations beginning with the *0*th iteration (the initial estimates), where is the value of the increment. If the iteration history is requested, then the last iteration is always displayed regardless of .
- **T [illegible].** When there are tied observed failure times, one of these methods is used to break the ties. The Efron method is more computationally expensive.

S [illegible] F [illegible]. These controls specify criteria for computations involving the survival function.

- **M [illegible].** The **B [illegible]** (or Nelson-Aalan or empirical) method estimates the baseline cumulative hazard by a nondecreasing step function with steps at the observed failure times, then computes the baseline survival by the relation survival=exp(−cumulative hazard). The **E [illegible]** method is more computationally expensive and reduces to the Breslow method when there are no ties. The [illegible] method estimates the baseline survival by a non-increasing right continuous function; when there are no predictors in the model, this method reduces to Kaplan-Meier estimation.
- **C [illegible].** The confidence interval can be calculated in three ways: in original units, via a log transformation, or a log-minus-log transformation. Only the log-minus-log transformation guarantees that the bounds of the confidence interval will lie between 0 and 1, but the log transformation generally seems to perform "best."

U [illegible] M [illegible] V [illegible]. All variables must have valid values for a case to be included in the analysis. These controls allow you to decide whether user-missing values are treated as valid among categorical models (including factors, event, strata, and subpopulation variables) and sampling design variables.

C [illegible] (%). This is the confidence interval level used for coefficient estimates, exponentiated coefficient estimates, survival function estimates, and cumulative hazard function estimates. Specify a value greater than or equal to 0, and less than 100.

O R [illegible] A [illegible]

The command language also allows you to:

- Perform custom hypothesis tests (using the CUSTOM subcommand and /PRINT LMATRIX).
- Tolerance specification (using /CRITERIA SINGULAR).
- General estimable function table (using /PRINT GEF).
- Multiple predictor patterns (using multiple PATTERN subcommands).

- Maximum number of saved variables when a rootname is specified (using the SAVE subcommand). The dialog honors the CSCOXREG default of 25 variables.

See the *C mma S a R* for complete syntax information.

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program,

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606

Printed in USA