Chapter 1

Changes and Enhancements to SAS/STAT Software in Versions 7 and 8

Overview

This chapter summarizes the major changes and enhancements to SAS/STAT software in Versions 7 and 8. All of these changes and enhancements are incorporated into the individual procedure chapters and are described in greater detail.

With Version 7 of SAS/STAT software, the PLS, KRIGE2D, and VARIOGRAM procedures became production. These procedures were experimental in Release 6.12. With Version 8, the SURVEYSELECT, SURVEYMEANS, SURVEYREG, KDE, LOESS, TPSPLINE, and NLMIXED procedures become production. These procedures were experimental in Version 7. **V8**

Output Delivery System

All procedures now incorporate the Output Delivery System (ODS). This is a system for managing the results of a procedure. By default, the results for a procedure are directed to the SAS listing file as in previous releases, but with ODS you can create HTML or RTF files, create SAS output data sets of any table in the output, select or exclude pieces of output from a procedure, or modify the organization and style of that output. Chapter 15, "Using the Output Delivery System," describes some typical uses of ODS with SAS/STAT software and provides a description of the basic features. Refer to *The Complete Guide to the SAS Output Delivery System* for complete documention of ODS.

As part of the ODS implementation, some of the output of the SAS/STAT procedures has been reorganized to be consistent across procedures.

In Version 8, some of the table names have been changed for various reasons. However, note that the table names that were in effect for the Version 7 release are still accepted, so programs written for Version 7 will still work. **V8**

ANOVA Procedure

The NAMELEN= option enables you to specify the length of effect names to be between 20 and 200 characters.

BOXPLOT Procedure

The new BOXPLOT procedure creates side-by-side box-and-whisker plots of measurements organized in groups. A box-and-whisker plot displays the mean, quartiles, and minimum and maximum observations for a group. You can specify multiple PLOT statements and also control the layout and appearance of the plots.

CATMOD Procedure

V8

The NOPRINT option has been added to the PROC and MODEL statements. The ESTIMATE= and ALPHA= options have been added to the CONTRAST statement.

CORRESP Procedure

V8

The CORRESP procedure provides adjusted inertias with the BENZECRI and GREENACRE options.

FACTOR Procedure

The NOPRINT option is now supported in the PROC FACTOR statement. The FUZZ, ROUND, and FLAG options are no longer supported. You can duplicate this functionality by creating the appropriate data sets from PROC FACTOR using ODS and then modifying them with the DATA step.

FASTCLUS Procedure

In the FREQ statement, frequencies are no longer truncated to integers.

When the IMPUTE option is specified in the PROC FASTCLUS statement, imputed values are no longer used in computing cluster statistics. This change causes the cluster standard deviations and other statistics computed from the standard deviations to be different than in previous releases.

The new INSTAT= option reads a SAS data set previously created by the FASTCLUS procedure using the OUTSTAT= option. If you specify the INSTAT= option, no clustering iterations are performed and no output is displayed. Only cluster assignment and imputation are performed as an OUT= data set is created.

The OUTSTAT= data set also contains the cluster seeds, and observations that were previously designated _TYPE_='SCALE' are now _TYPE_='DISPERSION'.

FREQ Procedure

The FREQ procedure now includes a TEST statement that provides asymptotic tests for selected measures of association and measures of agreement. A new BINOMIAL option in the TABLES statement computes the binomial proportion for one-way tables. You can compute the confidence bounds of a one-way table and request the test that the proportion equals a specified value as well as produce exact confidence bounds for the binomial proportion and an exact p-value for the binomial proportion test. You can now request that Fleiss-Cohen scores be used to compute the weighted kappa coefficient with the AGREE(WT=FC) option. The SCOROUT option requests that the row and column scores used for computing statistics such as Cochran-Mantel-Haenzsel statistics and Peason correlation be displayed.

The PCHI option in the EXACT statement computes the exact chi-square goodness-of-fit test for one-way tables as well as the exact Pearson chi-square test for two-way tables. The MAXTIME= option in the EXACT statement specifies the maximum time that PROC FREQ uses to compute an exact p-value.

The EXACT statement also includes the MC option for computing Monte Carlo estimates of exact p-values. *V8*

The Robins, Breslow, and Greenland (1986) estimate of variance is now used to compute the confidence bounds for the odds ratio and relative risk.

GENMOD Procedure

The earlier version of PROC GENMOD used a prototype Output Delivery System. This system has been totally rewritten; as a consequence, some of the syntax associated with ODS has changed. In particular, the ODS statement now replaces the use of the MAKE statement and _PRINT_ and _DISK_ global variables. The MAKE statement continues to be supported (except for its NOPRINT option), but the ODS statement provides much greater functionality and you should convert to using it. In addition, several of the table names and associated variable names in the GENMOD procedure have changed; see the chapter on the GENMOD procedure for complete information. The OUTPUT and the TEMPLATE procedures have changed. See Chapter 15, "Using the Output Delivery System," in this book for more information about the Output Delivery System.

PROC GENMOD now includes an LSMEANS statement that provides an extension of least squares means to the generalized linear model. In addition, the ESTIMATE statement is now supported. The new DIST=NEGBIN option in the MODEL statement specifies the negative binomial distribution, and the DIST=MULT option specifies the multinomial distribution. The log function is the default link for the negative binomial distribution, and the cumulative logit is the default link function for the multinomial distribution. Note that only the ordinal model is supported for the multinomial distribution, including the links CLOGIT for cumulative logit, CPROBIT for cumulative probit, and CCLL for cumulative complementary log-log.

The GEE facilities have also been updated. Type 3 tests are now provided for model effects, and the CONTRAST statement can be used for the GEE parameter estimates. The LSMEANS and ESTIMATE statements also apply to GEE parameter estimation. The method of alternating logistic regressions (ALR) is available with the LOGOR option in the REPEATED statement, which specifies the regression structure of the log odds ratio used to model the association of the responses from subjects for binary data. You can also fit the GEE model to ordinal data now, using the independent working correlation structure.

The NAMELEN= option in the PROC GENMOD statement enables you to specify the length of effect names to be between 20 and 200 characters.

V8 The DESCENDING option in the PROC statement specifies that the levels of the response variable be sorted in reverse order. The RORDER= option defines the ordering of the levels of the response variable. The procedure now includes an ID option in the MODEL statement for the OBSTATS table, and new variables have been added to the OUTPUT= data set.

GLM Procedure

The new ALPHA= option in the PROC GLM statement specifies the level of significance for confidence intervals computed from the LSMEANS, MEANS, MODEL, and OUTPUT statements. The ALPHA= option in all of these statements overrides the ALPHA value in the PROC GLM statement. The NAMELEN= option enables you to specify the length of effect names to be between 20 and 200 characters

The ALIASING option in the MODEL statement specifies that the estimable functions should be displayed as an aliasing structure, such that each row specifies the linear combination of the parameters estimated by each estimable function. This option is very useful in fractional factorial experiments that can be analyzed without a CLASS statement. The CLPARM option in the MODEL statement produces confidence limits for the parameter estimates (when you specify the SOLUTION option) and for the results of all ESTIMATE statements.

GLMMOD Procedure

The NAMELEN= option enables you to specify the length of effect names to be between 20 and 200 characters. The PREFIX= option specifies a prefix to use in naming the columns of the design matrix in the OUTDESIGN= data set. The ZEROBASED option specifies that the numbering for the columns of the design matrix in the OUTDESIGN= data set should begin at 0.

KDE Procedure

The KDE procedure performs either univariate or bivariate kernel density estimation. Statistical density estimation involves approximating a hypothesized probability density function from observed data. Kernel density estimation is a nonparametric technique for density estimation in which a known density function (kernal) is averaged across the observed data points to create a smooth approximation. PROC KDE uses a Gaussian density as the kernel, and its assumed variance determines the smoothness of the resulting estimate. PROC KDE outputs the kernel density estimate into a SAS data set, which you can then use with other procedures for plotting or analysis.

KRIGE2D Procedure

The KRIGE2D procedure performs ordinary kriging in two dimensions. Both anisotropic and isotropic semivariogram models can be handled. Four semivariogram models are supported: the gaussian, exponential, spherical, and power models. A single nugget effect is also supported. The locations of kriging estimates can be specified in a GRID statement or read from a SAS data set. The grid specification is most suitable for a regular grid; the data set specification can handle any irregular pattern of points. PROC KRIGE2D writes the kriging estimates and associated standard errors to an output data set.

LIFETEST Procedure

The plotting facility in the LIFETEST procedure has been upgraded, and high resolution plots are now the default. The CENSOREDSYMBOL= option and the EVENTSYMBOL= option specify the symbol for the censored and event observations, respectively.

LOGISTIC Procedure

The LOGISTIC procedure includes several new MODEL statement options that provide additional control over the model-fitting process. The ABSFCONV= option specifies the absolute function convergence criterion, the FCONV= option specifies the relative function convergence criterion, the GCONV= option specifies the relative gradient convergence criterion, and the XCONV= option specifies the relative parameter convergence criterion. The RIDGING= option specifies the technique used

to improve the log-likelihood function when its value is less than that of the previous step.

PROC LOGISTIC now supports the PREDPROBS= option in the OUTPUT statement. This option requests individual, cumulative, or cross validated predicted probabilities. The LACKFIT option now enables you to specify a number n to be subtracted from the number of partitions to give the correct degrees of freedom for the Hosmer and Lemeshow test.

V8

The LOGISTIC procedure supports the CLASS statement and the specification of model effects similar to the GLM procedure. You can specify the type of parameterization to use, such as effect coding and reference coding, the ordering of the classification variables, and the reference level. Such specifications can be done globally or for individual variables. See the information on the CLASS statement for more detail.

LOESS Procedure

The LOESS procedure implements a nonparametric method for estimating regression surfaces. The LOESS procedure allows great flexibility because no assumptions about the parametric form of the regression surface are needed. The LOESS procedure is suitable when there are outliers in the data and a robust fitting method is necessary. PROC LOESS fits nonparametric models, supports the use of multidimensional data, supports both direct and interpolated fitting using kd trees, and performs statistical inference.

MDS Procedure

The DIMENSION option in the PROC MDS statement now includes a BY parameter.

MIXED Procedure

Earlier versions of PROC MIXED used a prototype Output Delivery System. This system has been totally rewritten in Version 7; as a consequence, some of the syntax associated with ODS has changed. In particular, the ODS statement now replaces the use of the MAKE statement and _PRINT_ and _DISK_ global variables. The MAKE statement continues to be supported (except for its NOPRINT option), but the ODS statement provides much greater functionality and you should convert to using it. In addition, several of the table names and associated variable names in the MIXED procedure have changed; see the chapter on the MIXED procedure for complete information. The OUTPUT procedure and the TEMPLATE procedure have changed. See Chapter 15, “Using the Output Delivery System,” in this book for more information about the Output Delivery System.

The METHOD= option in the PROC MIXED statement has three new specifications: TYPE1, TYPE2, and TYPE3. These request analysis-of-variance estimates of variance components corresponding to type 1, 2, or 3 expected mean squares, respectively. These methods apply only to variance component models with no SUBJECT= effects and no REPEATED statement. The NAMELEN= option enables you to specify the length of effect names to be between 20 and 200 characters. The NCLPRINT option suppresses the display of the "Class Level Information" table, and the NOINFO option suppresses the display of the "Model Information" and "Dimensions" tables (this option replaces the INFO option).

The ID statement specifies the variables from the input data set to be included in the new OUTP= and OUTM= data sets from the MODEL statement.

In the MODEL statement, the OUTP= and OUTPM= options specify data sets containing predicted values and predicted means, respectively. These options replace the earlier P and PM options.

The PRIOR statement includes the following new options. The ALG=INDCHAIN option specifies a new default independence chain algorithm for generating the posterior sample, and the ALG=RWCHAIN option specifies the earlier random walk chain algorithm. The BDATA= option enables you to input the base densities used by the sampling algorithm. The GRID= and GRIDT= options specify grids and transformed grids, respectively, over which to evaluate the posterior density. The OUTG= and OUTGT= options specify output data sets to be created from the grid and transformed grid evaluations. The TRANS= option specifies the particular algorithm used to determine the transformation of the covariance parameters. The NOFULLZ option in the RANDOM statement eliminates the columns in Z corresponding to the missing levels of random effects involving CLASS variables.

V8

PROC MIXED provides the Kenward-Rogers method of computing degrees of freedom with the DDFM=KENWARDROGER option in the MODEL statement.

NLMIXED Procedure

The NLMIXED procedure fits nonlinear mixed models, that is, models in which both fixed and random effects enter nonlinearly. These models have a wide variety of applications, two of the most common being pharmacokinetics and overdispersed binomial data. PROC NLMIXED enables you to specify a conditional distribution for your data (given the random effects) having either a standard form (normal, binomial, Poisson) or a general distribution that you code using SAS programming statements. PROC NLMIXED fits nonlinear mixed models by maximizing an approximation to the likelihood integrated over the random effects.

NPAR1WAY Procedure

The NPAR1WAY procedure now provides tests for scale differences: the AB, KLOTZ, MOOD, and ST options in the PROC NPAR1WAY statement request tests based on Ansari-Bradley, Klotz, Mood, and Siegel-Tukey scores, respectively. The SCORES=DATA option requests analysis with raw input data values. This option provides the flexibility of constructing any set of scores and then analyzing these scores directly with PROC NPAR1WAY. The option is available for both two-sample and multi-sample data. You can request exact p-values in the EXACT statement for all of the preceding options. Also, the EXACT statement now includes the MC option for computing Monte Carlo estimates of exact p-values and the MAXTIME= option to specify the maximum time that PROC NPAR1WAY uses to compute an exact p-value. In addition, the NPAR1WAY procedure now has a FREQ statement.

ORTHOREG Procedure

The ORTHOREG procedure now supports the CLASS statement and allows the same specification of effects in the MODEL statement as the GLM procedure does. The NOINT option is also supported.

PHREG Procedure

The PHREG procedure includes several MODEL statement options that provide additional control over the optimization process. The ABSCONV= option specifies the absolute function convergence criterion, the FCONV= option specifies the relative function convergence criterion, the GONV= option specifies the relative gradient convergence criterion, and the XCONV= option specifies the relative parameter convergence criterion. The RIDGING= option specifies the technique used to improve the log-likelihood function when its value is less than that of the previous iteration.

PROC PHREG supports the NOTRUNCATE option in the FREQ statement to allow noninteger frequency values to be used in the computations. The default value for the ORDER= option in the OUTPUT statement has been changed from SORTED to DATA.

PLAN Procedure

The PLAN procedure can now be used to produce all possible permutations of n values and all possible combinations of n values taken k at a time.

PLS Procedure

The PLS procedure fits models using any one of a number of linear predictive methods, including partial least squares (PLS). Ordinary least squares regression has the single goal of minimizing sample response prediction error, seeking linear functions of the predictors that explain as much variation in each response as possible. The techniques implemented in the PLS procedure have the additional goal of accounting for variation in the predictors, under the assumption that directions in the predictor space that are well sampled should provide better prediction for new observations when the predictors are highly correlated. All of the techniques implemented in the PLS procedure work by extracting successive linear combinations of the predictors, called *factors*, that optimally address one or both of these two goals, explaining response variation and explaining predictor variation. In particular, the method of partial least squares balances the two objectives, seeking factors that explain both response and predictor variation.

REG Procedure

The ALPHA= option in the PROC REG statement sets the significance level for the construction of confidence intervals. Plots are now high resolution graphics by default, and you must specify the LINEPRINTER option if you want lineprinter plots. The TABLEOUT option now also outputs the upper and lower confidence limits to the OUTEST= data set.

In the MODEL statement, the CLB option requests confidence limits for the parameter estimates. The ALPHA= option in the MODEL statement can be used to set the significance level for the confidence limits produced by the current MODEL statement. Otherwise, the ALPHA= option in the PROC REG statement can be used to change the α level. The MAXSTEP option specifies the maximum number of steps to take when SELECTION=STEPWISE is used, and the SINGULAR= option for tuning singularity-checking overrides the same option in the PROC REG statement.

RSREG Procedure

The RSREG procedure no longer requires the data to be sorted in order to test lack-of-fit.

SIM2D Procedure

The SIM2D procedure produces a spatial simulation for a Gaussian random field with a specified mean and covariance structure in two dimensions using an LU decomposition technique. The simulation can be conditional or unconditional. If the simulation is conditional, a set of coordinates and associated field values are read from a SAS data set. The resulting simulation will honor these data values. The mean structure can be specified as a quadratic in the coordinates. The covariance is specified by naming the form and supplying the associated parameters. The locations of simula-

tion points can be specified in a GRID statement or read from a SAS data set. The grid specification is most suitable for a regular grid; the data set specification can handle any irregular pattern of points. The SIM2D procedure writes the simulated values for each grid point to an output data set. The SIM2D procedure does not produce any displayed output.

STDIZE Procedure

The STDIZE procedure standardizes one or more numeric variables in a SAS data set by subtracting a location measure and dividing by a scale measure. A variety of location and scale measures are provided, including estimates that are resistant to outliers and clustering. You can also multiply each standardized value by a constant and add a constant. You can replace missing values by the location measure or by any specified constant; you can suppress standardization if you only want to replace missing values.

SURVEYMEANS Procedure

The SURVEYMEANS procedure produces estimates of survey population means and totals from sample survey data. The procedure also produces variance estimates, confidence limits, and other descriptive statistics. When computing these estimates, the procedure takes into account the sample design used to select the survey sample. The sample design can be a complex survey sample design with stratification, clustering, and unequal weighting.

SURVEYREG Procedure

The SURVEYREG procedure performs regression analysis for sample survey data. This procedure can handle complex survey sample designs, including designs with stratification, clustering, and unequal weighting. The procedure fits linear models for survey data and computes regression coefficients and their variance-covariance matrix.

SURVEYSELECT Procedure

The SURVEYSELECT procedure provides a variety of methods for selecting probability-based random samples. The procedure can select a simple random sample or a sample according to a complex multistage sample design that includes stratification, clustering, and unequal probabilities of selection.

TPSPLINE Procedure

The TPSPLINE procedure uses the penalized least squares method to fit a nonparametric regression model. It computes thin-plate smoothing splines to approximate smooth multivariate functions observed with noise. The TPSPLINE procedure allows great flexibility in the possible form of the regression surface. In particular, PROC TPSPLINE makes no assumptions of a parametric form for the model. The generalized cross validation (GCV) function can be used to select the amount of smoothing.

TRANSREG Procedure

The TRANSREG procedure now supports smoothing spline transformations in the MODEL statement. The SMOOTH option specifies a noniterative transform, and the SSPLINE option specifies an iterative smoothing spline transformation. You can specify the smoothing parameter with the PARAMETER= option or the new SM= option. The DESIGN option has been enhanced. Other new options in PROC TRANSREG provide control over output data set variable names and labels.

TTEST Procedure

The TTEST procedure now performs t tests for one sample, two samples, and paired observations. The ALPHA= option in the PROC TTEST statement specifies the alpha level for the confidence intervals produced. The CI= option specifies that a confidence interval be produced for the standard deviation and that the confidence interval be either an equal tailed confidence interval or an interval based on the uniformly most powerful unbiased test of $H_0{:}\sigma = \sigma_0$. The H0= option requests tests against m instead of 0.

The FREQ and WEIGHT statements are now supported. The new PAIRED statement identifies the variables to be compared in paired comparisons.

VARIOGRAM Procedure

The VARIOGRAM procedure computes sample or empirical measures of spatial continuity for two-dimensional spatial data. These continuity measures are the regular semivariogram, a robust version of the semivariogram, and the covariance. These measures are written to an output data set, allowing plotting or parameter estimation for theoretical semivariograms or covariance models. Both isotropic and anisotropic measures are available. You can then use the KRIGE2D procedure for spatial prediction.

References

Robins, J.M., Breslow, N., and Greenland, S. (1986), "Estimators of the Mantel-Haenszel Variance Consistent in Both Sparse Data and Large-Strata Limiting Models," *Biometrics*, 42, 311–323.