Enhancements to SAS/STAT Software in SAS 9

The SAS 9 family, of which the latest release is SAS 9.1, brings new procedures and enhancements to SAS/STAT software. The following are the highlights:

Statistical Graphics Using ODS

A number of SAS/STAT procedures use an experimental extension to the Output Delivery System (ODS) that enables them to create statistical graphics. The facility is invoked when you include an ODS GRAPHICS statement prior to your procedure statements. Graphics are then created automatically, or when you specify procedure options for graphics. Current output destinations include HTML, LATEX, PRINTER, and RTF. ODS styles control the appearance of the graphs, which are customizable. Over a dozen SAS/STAT procedures now take advantage of ODS graphics, including such mainstays as GENMOD, GLM, LIFETEST, MIXED, PHREG, and REG.

Survey Data Analysis

New survey data analysis capabilities are now included in the SAS System. The SURVEYFREQ procedure produces one-way to n-way frequency and crosstabulation tables for survey data. These tables include estimates of totals and proportions (overall, row percentages, and column percentages) and the corresponding standard errors. Like the other survey procedures, PROC SURVEYFREQ computes variance estimates based on the sample design used to obtain the survey data, which can include stratification, clustering, and unequal weighting. PROC SURVEYFREQ also provides design-based tests of association between variables.

The SURVEYLOGISTIC procedure performs logistic regression for data that arise from a survey sampling scheme. PROC SURVEYLOGISTIC incorporates complex survey sample designs in its estimation process. Variances of the regression parameters and odds ratios are computed using a Taylor expansion approximation. The SURVEYLOGISTIC procedure is similar in syntax to the LOGISTIC procedure, and it can fit link functions such as the logit, cumulative logit, generalized logit, probit, and complementary log-log functions.

Robust Regression

The ROBUSTREG procedure provides resistant (stable) results in the presence of outliers by limiting the influence of outliers. The procedure can be used for outlier detection and robust regression, and it provides M estimation, LTS estimation, S estimation, and MM estimation. With these four methods, the ROBUSTREG procedure acts as an integrated tool for outlier detection and robust regression with various contaminated data. The ROBUSTREG procedure is scalable so that it can be used for applications in data cleansing and data mining.
Mixed Models

The MIXED procedure now supports geometrically anisotropic covariance structures and covariance models in the Matérn class. The LCOMPONENTS option in the MODEL statement produces one degree of freedom tests for fixed effects that correspond to individual estimable functions for Type I, II, and III effects. The experimental RESIDUAL option of the MODEL statement computes Pearson-type and (internally) studentized residuals. The experimental INFLUENCE option in the MODEL statement computes influence diagnostics by noniterative or iterative methods. Experimental ODS graphics display the results for both of these options.

Power and Sample Size Computations

SAS 9 contains the POWER and GLMPOWER procedures, which perform power and sample size computations, and PSS, a web application that provides an interface to these computations. The POWER procedure covers a variety of statistical analyses such as t tests, equivalence tests, and confidence intervals for means; exact binomial, chi-square, Fisher’s exact, and McNemar tests for proportions; multiple regression and correlation; one-way analysis of variance; and rank tests for comparing survival curves. The GLMPOWER procedure performs prospective analyses for linear models. You specify the design and the cell means using an exemplary data set and specify the model and contrasts using MODEL and CONTRAST statements like those in the GLM procedure. The point-and-click PSS application includes tasks for determining sample size and power for a variety of statistical analyses. It provides multiple input parameter options, stores results in a project format, displays power curves, and produce appropriate narratives for the results.

<table>
<thead>
<tr>
<th>Index</th>
<th>N Total</th>
<th>Mean Diff</th>
<th>Std Dev</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>6</td>
<td>7.5</td>
<td>0.395069</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>6</td>
<td>7.5</td>
<td>0.977279</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>8</td>
<td>7.5</td>
<td>0.616660</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>8</td>
<td>7.5</td>
<td>0.999951</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>6</td>
<td>9.5</td>
<td>0.267288</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>6</td>
<td>9.5</td>
<td>0.878362</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>8</td>
<td>9.5</td>
<td>0.429651</td>
</tr>
<tr>
<td>8</td>
<td>100</td>
<td>8</td>
<td>9.5</td>
<td>0.986411</td>
</tr>
</tbody>
</table>

**Proc Power:**

TwoSampleMeans
  NTot = 20 100
  Power = .
  MeanDiff = 6 8
  StdDev = 7.5 9.5;
  Plot;
run;

Distance Measures

The new DISTANCE procedure computes various measures of distance, dissimilarity, or similarity between the observations (rows) of a SAS data set. These proximity measures are stored as a lower triangular matrix or a square matrix in an output data set that can then be used as input to the CLUSTER, MDS, and MODECLUS procedures. The input data set may contain numeric or character variables, or both, depending on which proximity measure is used. PROC DISTANCE also provides various nonparametric and parametric methods for standardizing variables. Distance matrices are used frequently in data mining, genomics, marketing, financial analysis, management science, education, chemistry, psychology, biology, and various other fields.

Multiple Imputation
The MI and MIANALYZE procedures create and analyze multiply imputed data sets for incomplete multivariate data. The MI procedure creates multiply imputed data sets for incomplete p-dimensional multivariate data. It uses methods that incorporate appropriate variability across m imputations. Once the m complete data sets are analyzed using standard SAS/STAT procedures, PROC MIANALYZE can be used to generate valid statistical inferences by combining the results. The MI procedure provides three methods for imputing missing values. For monotone missing data patterns, either a parametric regression method that assumes multivariate normality or a nonparametric method that uses propensity scores is available. For an arbitrary missing data pattern, a Markov chain Monte Carlo (MCMC) method that assumes multivariate normality can be used.

New features in PROC MI include the PMM and REGPMM options in the MCMC and MONOTONE statements, respectively, which request predictive mean matching method to impute an observed value that is closest to the predicted value. Other additions include an experimental CLASS statement to specify categorical variables. Experimental options LOGISTIC and DISCRIM in the MONOTONE statement impute missing categorical variables by logistic and discriminant methods, respectively. New features in PROC MI include the enhancement of the DATA= option to input data sets that contains both parameter estimates and their associated standard errors in each observation. Other additions to PROC MIANLYZE include a TEST statement and an experimental CLASS statement.

### Conditional Logistic Regression

Conditional logistic regression is an appropriate technique when you have data that are highly stratified, such as clinical trials data where you have a large number of clinics and a small number of subjects in each clinic. Basing the logistic regression on a conditional likelihood allows you to eliminate the strata parameters, or nuisance parameters, and focus on estimating the effects of interest. SAS users can now use the LOGISTIC procedure to perform conditional logistic regression with the new STRATA statement. For an exact conditional analysis, specifying the STRATA statement produces an efficient stratified analysis.

Other enhancements to the LOGISTIC procedure include the new SCORE statement that enables you to score new data sets and compute fit statistics and ROC curves without refitting the model. Several new CLASS parameterizations are also available: ordinal, bathtub, orthogonal effect, orthogonal reference, orthogonal ordinal, and orthogonal bathtub. Also, you can now output the design matrix using the new OUTDESIGN= option.

### Proportional Hazards Regression

The TPHREG procedure adds the CLASS statement to the PHREG procedure. Model effects such as covariates, main effects, interactions, and nested effects, can be specified in the same way as in the GLM procedure. Similar to the LOGISTIC procedure, PROC TPHREG supports the less than full-rank parameterization of the GLM procedure, and it also supports various full-rank parameterization schemes such as reference coding and effect coding. These enhancements will be incorporated into the PHREG procedure in a future release.

Meanwhile, the PHREG procedure has been updated in numerous ways. The new WEIGHT statement enables you to specify case weights when you are using the BESLOW or EFRON method for handling ties. Robust sandwich variance estimators (Binder 1992) are now computed for the estimated regression parameters. You can specify the NORMALIZE option to normalize the weights so that they add up the actual sample size.

The TEST statement now includes the AVERAGE option, which enables you to compute a combined estimate of all the effects in the statement. The new E option in the TEST statement specifies that the linear coefficients and constants be printed. The recurrence algorithm of Gail et al (1981) for computing the exact discrete partial likelihood and its partial derivatives has been modified to use the logarithmic scale, which allows the procedure to handle a greater number of times with this method.

### Other Key Enhancements
**FREQ**

The BDT option includes Tarone's adjustment in the Breslow-Day test for homogeneity of odds ratios. The ZEROS option in the WEIGHT statement includes zero-weight observations in the analysis. With the ZEROS option, PROC FREQ displays zero-weight levels in crosstabulation and frequency tables. For one-way tables, the ZEROS option includes zero-weight levels in chi-square tests and binomial statistics. For multi-way tables, the ZEROS option includes zero-weight levels in kappa statistics. The FREQ procedure now produces exact confidence limits for the common odds ratio and related tests.

**GENMOD**

The GENMOD procedure now offers several new full-rank CLASS variable parameterizations: polynomial, orthogonal polynomial, effect, orthogonal effect, reference, orthogonal reference, ordinal, orthogonal ordinal, bathtub, and orthogonal bathtub. The default parameterization remains the same less-than-full-rank parameterization used in previous releases. Other new CLASS statement options, like the ability to request a specific reference level, are also available.

**GLM**

The GLM procedure now compute exact p-values for three of the four multivariate tests (Wilks' Lambda, the Hotelling-Lawley Trace, and Roy's Greatest Root) and an improved F-approximation for the fourth (Pillai's Trace). Specifying MSTAT=EXACT computes exact p-values for three of the four tests (Wilks' Lambda, the Hotelling-Lawley Trace, and Roy's Greatest Root) and an improved F-approximation for the fourth (Pillai's Trace).

**LIFETEST**

The SURVIVAL statement enables you to create confidence bands (also known as simultaneous confidence intervals) for the survivor function $S(t)$ and to specify a transformation for computing the confidence bands and the pointwise confidence intervals. Additional tests are available for comparing two or more samples of survival data, including the Tarone-Ware test, the Peto-Peto test, the modified Peto-Peto test, and the Fleming-Harrington family of tests. Trend tests for ordered alternatives can be requested. Stratified tests compare survival functions while adjusting for prognostic factors that affect the event rates.

**NPAR1WAY**

The D option generates the one-sided D+ and D- statistics for the asymptotic two-sample Kolmogorov-Smirnov test, in addition to the two-sided D statistic given by the EDF option. The KS option in the EXACT statement gives exact tests for the Kolmogorov-Smirnov D+, D-, and D for two-sample problems.

**TRANSREG**

The TRANSREG procedure has new transformation options for centering and standardizing variables, CENTER and Z, before the transformations. The new algorithm option INDIVIDUAL with METHOD=MORALS fits each model for each dependent variable individually and independently of the other dependent variables.

**Obtaining SAS 9.1**

SAS 9.1 is currently available. To obtain more information, ask your organization’s SAS representative to contact the SAS Customer Interaction Center at 1.800.727.0025.