Paper 257-27
A Preview of SAS/STAT® Version 9: Moving in New Directions and Building on Old Favorites

Maura Stokes, Robert Rodriguez, Randy Tobias
SAS Institute Inc.
Cary, North Carolina, USA

Abstract

Version 9 of SAS/STAT software brings you a variety of new tools for your statistical computing needs. The Power and Sample Size Application (PSS) provides sample size and power computations for a variety of analyses through a web interface. Experimental software in Version 9 moves SAS/STAT in new directions, including robust regression, which is supported by the ROBUSTREG procedure, and logistic regression for sample survey data, available with the SURVEYLOGISTIC procedure. In response to user requests, a number of procedures have been enhanced significantly. Conditional logistic regression is available in the LOGISTIC procedure through the new STRATA statement, and scoring of data sets is available through the new SCORE statement. New features in several other procedures are also discussed.

Introduction

Version 9 of SAS/STAT software delivers a variety of new procedures and enhancements, which are motivated by methodological advances in statistics, technological improvements in the SAS® System, and user requests for extensions to existing software. In order to respond to this spectrum of requirements, the development for each new release of SAS/STAT is balanced across a combination of experimental procedures, procedures which are achieving production status for the first time, and incorporation of new features in standard procedures. Experimental procedures provide a vehicle for introducing new methodology; their feature sets and syntax are subject to change based on user feedback, and they are documented in papers and preliminary documentation which are available for downloading at http://www.sas.com/statistics/. Typically, experimental procedures attain production status in the following release through additional development and testing, as well as standard documentation.

The purpose of this paper is to describe the combination of "old" and "new" work which supports the goals for Version 9 of SAS/STAT software.

Multiple Imputation

Multiple imputation provides a useful strategy for dealing with missing values. Instead of filling in a single value for each missing value, Rubin's (1987) multiple imputation procedure replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. These multiply-imputed data sets are then analyzed by using standard procedures for complete data and combining the results from these analyses. No matter which complete-data analysis is used, the process of combining results from different imputed data sets is essentially the same. This results in statistically valid inferences that properly reflect the uncertainty due to missing values.

Version 8 SAS/STAT software introduced the experimental MI and MIANALYZE procedures for creating and analyzing 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 about these parameters by combining the results.

The MI procedure provides three methods for imputing missing values and the method of choice depends on the type of missing data pattern. For monotone missing data patterns, you can use a parametric regression method that assumes multivariate normality or a nonparametric method based on propensity scores. For an arbitrary missing data pattern, you can use a Markov chain Monte Carlo (MCMC) method that assumes multivariate normality.

In Version 9, the MI procedure includes predictive mean matching for the MCMC method, as well as the monotone methods. For the monotone methods, a separate imputation model can be used for each imputed variable. The following statements illustrate the syntax for specifying different models.

```
proc mi;
   monotone propensity( y2 )
            reg( y3 y4 = y1 y2 y1*y2 );
   var y1 y2 y3 y4;
run;
```

In addition, classification variables can be used either as covariates or as imputed variables for the monotone methods. The logistic and discrimination methods can be used to impute classification variables.

MIANALYZE procedure updates for Version 9 include a simplification of the input data sets. The procedure allows the use of the PARMS= option without the associated COVB= or XPXI= option when the PARMS= data set contains parameter estimates and associated standard errors computed from imputed data sets. The procedure can also read the parameter estimates and associated standard errors in a DATA= data set when the specified standard errors are enclosed in parentheses and are corresponding to parameter estimates in the order in which they appear in the VAR statement, as illustrated in the following statements.

```
proc mianalyze;
   var y1-y3 (sy1-sy3);
run;
```

In addition, the updates also include a TEST statement for assessing the significance of linear combinations of the parameters.

Conditional Logistic Regression

Conditional logistic regression has long been used in epidemiology where a retrospective study matches subjects, or cases, having an event of interest with similar subjects, or controls, who do not have the event. More recently, conditional logistic regression has also been applied to highly stratified data and crossover studies. With highly stratified data, you may have a small number of subjects per stratum, and thus a small number of subjects relative to the number of parameters you are estimating because stratification effects must be estimated. Consequently, the sample size requirements for the usual maximum likelihood approach to unconditional logistic regression may not be met.

In the past, SAS users resorted to using the PHREG procedure for conditional logistic regression. While this procedure is designed for proportional hazards regression analysis, computational equivalences make it appropriate for conditional logistic regression. However, this is more awkward than using a procedure that is designed for logistic regression. Version 9 brings conditional logistic regression to the LOGISTIC procedure via the new STRATA statement.

Stokes et al. (1999) include an example of a clinical trial in which researchers studied the effects of a new treatment for a skin condition. A pair of patients participated from each of 79 clinics. One person received the treatment and another person received the placebo. Age, sex, and an initial score for the skin condition (ranging from 1 to 4 for mild to severe) were recorded. The response was whether the skin condition improved. Note that because there are only two observations per clinic, it would not be possible to estimate properly a clinic effect.

The data have the following form, where each line consists of two observations.

```
data trial;
   input center treat $ sex $ age improve
         initial @@;
   datalines;
   1  t f 27 0 1  1 p f 32 0 2
   2  t f 41 1 3  2 p f 47 0 1
   3  t m 19 1 4  3 p m 31 0 4
   4  t m 55 1 1  4 p m 24 1 3
   . . .
```

The following PROC LOGISTIC statements produce the desired analysis. Putting the variable CENTER in the STRATA statement produces strata based on the values of CENTER. Note the use of the REF= option in the CLASS statement to specify the reference level to be used in creating effects for SEX and TREAT. The EVENT= option in the MODEL statement provides a useful way to determine the value of the response variable IMPROVE on which the analysis is based.

```
proc logistic data=trial2;
   class sex(ref='f') treat(ref='p') / param=ref;
   strata center;
   model improve(event='1') = initial age sex treat;
run;
```

Figure 1 displays the Type 3 information. The initial score is highly significant in this model, and treatment is marginally significant. The variables AGE and SEX appear not to have influence in this model. Depending on whether there are research reasons to maintain SEX and AGE in the model, subsequent models might not include these variables.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
initial	1	10.6106	0.0011
age	1	1.2253	0.2683
sex	1	0.9176	0.3381
treat	1	3.8053	0.0511

Figure 1. Type 3 Tests for Clinical Trial Data

Figure 2 displays the parameter estimates.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
initial		1	1.0915	0.3351	10.6106	0.0011
age		1	0.0248	0.0224	1.2253	0.2683
sex	m	1	0.5312	0.5545	0.9176	0.3381
treat	t	1	0.7025	0.3601	3.8053	0.0511

Figure 2. Parameter Estimates for Clinical Trial Data

Robust Regression

Robust regression is useful for dealing with outliers in regression analysis. It can be used to detect outliers and to produce stable estimates in their presence by reducing their influence. The types of outliers addressed with robust regression include problems in the response direction, problems in the covariate space (leverage points), and problems in both directions.

Three general methods of robust regression are commonly employed. Huber M-estimation (Huber 1973) is the simplest approach and is appropriate when you can assume that the outliers are mainly in the response direction; it is not robust with respect to leverage points. Least Trimmed Squares (LTS) is a high breakdown method introduced by Rousseeuw (1984). Rousseeuw and Yohai (1984) introduced another high breakdown method that can be more efficient than LTS estimation. Finally, MM-estimation, introduced by Yohai (1987), combines both high breakdown estimation and M-estimation.

The ROBUSTREG procedure, experimental in Version 9, brings robust regression to SAS/STAT software, and it provides all of the above-mentioned methods. Until now, the LTS and FAST-LTS methods have been available in SAS only through the use of the LTS call in SAS/IML software.

To illustrate the benefits of PROC ROBUSTREG, consider the classic stackloss data first presented by Brownlee (1965). The data concern the oxidation of ammonia to nitric acid and consists of 21 observations. Y is the stackloss, and the covariates are X1, the rate of operation, X2, the cooling water inlet temperature, and X3, the acid concentration.

```
data stack;
   input x1 x2 x3 y @@;
   datalines;
 80  27  89  42  58  18  82  11
 80  27  88  37  58  19  93  12
 75  25  90  37  50  18  89   8
 62  24  87  28  50  18  86   7
 62  22  87  18  50  19  72   8
 62  23  87  18  50  19  79   8
 62  24  93  19  50  20  80   9
 62  24  93  20  56  20  82  15
 58  23  87  15  70  20  91  15
 58  18  80  14
 58  18  89  14
 58  17  88  13
 ;
```

The PROC ROBUSTREG invocation fits a robust regression. The DIAGNOSTICS option in the MODEL statement requests a table for outlier statistics, and the LEVERAGE option requests the addition of leverage point diagnostics to this table. Specifying X1 in the ID statement means that the value of X1 will be used to identify the observations in the diagnostics table. The TEST statement produces a test of significance for covariate X3.

```
proc robustreg data=stack;
   model y= x1 x2 x3 / diagnostics leverage;
   id x1;
   test x3;
run;
```

Figure 3 displays the summary statistics for the model. The MAD statistic provides a robust estimate of the univariate scale, computed as the corrected *median absolute deviation*.

Summary Statistics

Variable	Q1	Median	Q3	Mean	Standard Deviation	MAD
x1	53	58	62	60.42857	9.168268	5.930409
x2	18	20	24	21.09524	3.160771	2.965204
x3	82	87	89.5	86.28571	5.358571	4.447807
y	10	15	19.5	17.52381	10.17162	5.930409

Figure 3. Summary Statistics

Figure 4 displays the parameter estimates. The Scale is a point estimate of the scale parameter in the linear regression model.

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-42.2854	9.5045	-60.9138	-23.6569	19.79	<.0001
x1	1	0.9276	0.1077	0.7164	1.1387	74.11	<.0001
x2	1	0.6507	0.2940	0.0744	1.2270	4.90	0.0269
x3	1	-0.1123	0.1249	-0.3571	0.1324	0.81	0.3683
Scale	1	2.2819					

Figure 4. Parameter Estimates for Stackloss Data

Figure 5 displays the model diagnostics. Four observations are found to have high leverages.

Diagnostics

Obs	x1	Mahalanobis Distance	Robust MCD Distance	Leverage	Robust Residual	Outlier
1	80.00	2.2536	5.5284	*	1.0995	
2	80.00	2.3247	5.6374	*	-1.1409	
3	75.00	1.5937	4.1972	*	1.5604	
4	62.00	1.2719	1.5887		3.0381	*
21	70.00	2.1768	3.6573	*	-4.5733	*

Figure 5. Diagnostics

Figure 6 displays significance tests for variable X3. These results indicate that variable X3 is not influential.

Robust Linear Tests

Test	Test Statistic	Lambda	DF	Chi-Square	Pr > ChiSq
Rho-test	0.9378	0.7977	1	1.18	0.2782
Rn2-test	0.8092	_	1	0.81	0.3683

Figure 6. Test of Significance

Further analysis might include sharper outlier detection, which can be accomplished by specifying a smaller tuning constant for the bisquare weight function. This would lower the asymptotic efficiency but sharpen the default M-estimator.

For more details concerning the ROBUSTREG procedure, see Chen (2002).

CLASS Statement in PROC PHREG

For many years, users have requested that a CLASS statement be implemented in the PHREG procedure. In Version 9, this is accomplished with the TPHREG procedure, a test version of the PHREG procedure. This means that you can specify interaction terms for your model as in the GLM procedure. In addition, the CLASS statement supports various nonsingular parameterizations as in the LOGISTIC procedure.

New Procedures for Survey Data Analysis

Many researchers use sample surveys to collect their information, relying on probability-based complex sample designs such as stratified selection, clustering, and unequal weighting. This is done to select samples at lowest possible cost that can produce estimates that are precise enough for the purposes of the study. To make statistically valid inferences, the study design must be taken into account in the data analysis. Traditional SAS procedures such as the MEANS and GLM procedures are inappropriate for analyzing survey data because they compute statistics under the assumption that the sample is drawn from an infinite population with simple random sampling.

In Version 8, SAS/STAT introduced three procedures for sample survey selection and survey data analysis. PROC SURVEYSELECT enables you to select a probability-based sample and produces an output data set that contains the selected units, their selection probabilities, and the sampling weights. The SURVEYMEANS procedure computes estimates of survey population totals and means, estimates of their variances, confidence limits, and other descriptive statistics. The SURVEYREG procedure performs regression analysis for sample survey data.

In Version 9, SAS/STAT provides two experimental procedures for the analysis of sample survey data. 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, 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. The design can be a complex sample survey design with stratification, clustering, and unequal weighting. PROC SURVEYFREQ also provides design-based tests of association between variables. And for 2×2 tables, the procedure computes estimates of risk differences, odds ratios, relative risks, and their confidence limits.

The SURVEYLOGISTIC procedure performs logistic regression for sample survey data. Categorical response data are often collected in large scale surveys, and logistic regression is an obvious choice for analyzing such data. Since the survey collection schemes usually include such factors as stratification and clustering, the analysis must contend with the sample design. Theoretical work in this area was done in the 1980s by Binder (1981, 1983) and Roberts, Rao, and Kumar (1987). The SURVEYLOGISTIC procedure provides much of the capability of the LOGISTIC procedure.

Power and Sample Size Computations

Version 9 brings comprehensive facilities for power and sample size computation to the SAS System in the form of two new procedures in SAS/STAT software and a web application. The POWER procedure performs power analysis, including determining the sample size required to get a significant result with adequate probability and characterizing the power of a study to detect a meaningful effect. Analyses covered by PROC POWER include means, proportions, corre-

lation, regression, ANOVA, and survival analysis. The GLMPOWER procedure provides similar functionality for linear models.

The Power and Sample Size Application (PSS) is a web application that provides power and sample size computations via a point-and-click interface. A variety of statistical tasks are covered, including t-tests, ANOVA, confidence intervals, proportions, equivalence testing, linear models, and survival analysis. The application provides multiple input parameter options, stores results in a project format, displays power curves, and produces appropriate narratives for the results. PSS can be run locally or from a server.

The POWER and GLMPOWER procedures and the PSS Application are experimental in Version 9.

Parallel Computing

The Threaded Kernel (TK) architecture being introduced in Version 9 enables SAS to incorporate high performance parallel computing enhancements. SAS procedures have traditionally been single-threaded, meaning that computational steps are processed strictly sequentially and one at a time. In contrast, TK-enabled SAS can run in multiple threads, allowing different pieces of code to run simultaneously, or in parallel. With several threads executing concurrently, a single program can divide its work between several processors, and thus run faster.

Replacing single-threaded computational algorithms with multi-threaded algorithms requires subtle resource management and complex task coordination. However, if this is done well, then multi-threading can deliver dramatic performance improvements. Among the critical procedures that take advantage of TK in Version 9 are the SAS/STAT workhorse procedures, REG and GLM, as well as the DMREG procedure in Enterprise Miner™. The parts of the procedures that have been successfully parallelized include accumulating the sufficient statistics for the analysis, inverting the information matrix, and calculating ANOVA sums of squares. For more details, see Cohen (2002).

Statistical Distances

Distance matrices are used frequently in data mining, genomics, marketing, financial analysis, management science, education, chemistry, psychology, biology, and various other fields.

The DISTANCE procedure, experimental in Version 9, 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 (depending on the SHAPE= option) 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.

Enhancements to SAS/STAT Procedures

Many of the updates in each new release of SAS/STAT software are enhancements to existing procedures in response to customer suggestions and feedback. In Version 9, some of these enhancements include the following:

- CLASS statement update in PROC GENMOD
- SCORE statement in PROC LOGISTIC
- transform confidence intervals in PROC LIFETEST
- exact p-values for multivariate tests in PROC GLM

Extended CLASS statement in PROC GENMOD

The CLASS statement in the GENMOD procedure has been extended in Version 9 and is now comparable to the CLASS statement in the LOGISTIC procedure. The user can now specify the desired style of parameterization, including reference cell coding, effect coding, orthogonal coding, and even polynomial coding. Parameterization can be specified as a whole or individually for specific effects. In addition, the reference levels for effects can be specified, instead of being fixed as the last ordered level of the effect variable.

Scoring in PROC LOGISTIC

You can use the new SCORE statement in PROC LOGISTIC to score new data (compute posterior probabilities) and compute their fit statistics. If the response variable is binary, you can also compute the ROC curve for the new data. Not only can you score a data set right after fitting a model, you can also score a data set based on previously saved model information.

The scored values and optionally the corresponding confidence limits are output to a SAS data set. You can request fit statistics for the data, which is especially useful for test or validation data. For binary response data, you can also create a SAS data set containing the receiver operating characteristics (ROC) curve. You can specify multiple SCORE statements in the same invocation of PROC LOGISTIC.

PROC LOGISTIC provides the capability of scoring new data based on information of a previously fitted model without having to refit the model. You use the

OUTMODEL= option in the PROC statement to save the model information in the specified SAS data set. To score new data without having to refit the model, you use the INMODEL= option to specify the SAS data set that contains the model information.

Transform Confidence Limits in PROC LIFETEST

In PROC LIFETEST, you specify the OUTSURV= option to request an output SAS data set that contains the point estimates of the survivor function and the corresponding pointwise confidence intervals. However, the linear pointwise confidence intervals are often out of the [0,1] range for extreme values of the time variable. By applying the asymptotic normality to a transformation of the survivor function, you can obtain better confidence intervals. The new CONFTYPE= option allows you to pick a transformation for the confidence intervals. The choices are the arcsine-square root, complementary log-log, log, and logit transformations.

Exact p-Values for Multivariate Tests in PROC GLM

Exact p-values are now available for the multivariate tests in PROC GLM with the specification of the MSTAT=EXACT option in the MANOVA and REPEATED statements. The approximate p-values corresponding to F statistics will still be printed by default; however, research has shown that some of these approximations may not be entirely satisfactory.

Conclusion

Version 9 brings another round of updates and enhancements to SAS/STAT software. A major new direction, multiple imputation, becomes mature with the production release of the MI and MIANALYZE procedures. Work will be ongoing to add even more multiple imputation methods to these and possibly other procedures. A longstanding SAS user request for more power and sample size tools is realized with the release of both the PSS Interface and the POWER and GLMPOWER procedures. The recent SAS direction in providing tools for survey selection and survey data analysis is further developed with procedures for survey logistic regression and contingency table analysis. The longtime SAS commitment to providing comprehensive software for statistical modeling is continued with the new procedure for robust regression, and now the modeling can often be done more quickly with the introduction of parallelization in several procedures. And finally, numerous existing procedures have been updated, often in response to user feedback, so that they will be even more useful to statisticians doing data analysis.

References

Binder, D. A. (1981), “On the Variances of Asymptotically Normal Estimators from Complex Surveys,” *Survey Methodology*, 7, 157–170.

Binder, D. A. (1983), “On the Variances of Asymptotically Normal Estimators from Complex Surveys,” *International Statistical Review*, 51, 279–292.

Brownlee, K. A. (1965), *Statistical Theory and Methodology in Science and Engineering, Second Edition*, New York: John Wiley & Sons

Chen, C. (2002), “Robust Regression and Outlier Detection with the ROBUSTREG Procedure”, *Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.

Cohen, R. A. (2002), “SAS Meets Big Iron: High Performance Computing in SAS Analytic Procedures”, *Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.

Huber, P. J. (1973), “Robust regression: Asymptotics, conjectures, and Monte Carlo”, *Ann. Stat.*, 1. 799–821.

Roberts, G., Rao, J. N. K., and Kumar, S (1987), “Surveylogistic Regression Analysis of Sample Survey Data”, *Biometrika*, 74, 1–12.

Rousseeuw, P. J. (1984), “Least Median of Squares Regression”, *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P. J. and Yohai, V. J. (1984). “Robust regression by means of S-estimators. In Robust and Nonlinear Time Series”, J. Franke, W. Hardle and D. Martin, eds.)*Lecture Notes in Statistics*, 26 256–272, New York: Springer.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons, Inc.

Stokes, M.E., Davis, C.S., and Koch, G.G. (2000), *Categorical Data Analysis Using the SAS System, Second Edition*, Cary, NC: SAS Institute Inc.

Contact Information

Maura Stokes, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513.