

Paper 265-28

SAS/STAT[®] Version 9: Progressing into the Future (Preliminary Paper)

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

Abstract

Version 9 of SAS/STAT software delivers a wealth of tools and functionality for statistical modeling and data analysis. Areas covered by new or enhanced software include multiple imputation, conditional logistic regression, robust regression, general linear models for proportional hazards, regression diagnostics, survey data analysis, power and sample size analysis, statistical distance computations, statistical graphics, and parallelization.

Introduction

This presentation overviews facilities and enhancements for statistical modeling and data analysis in Version 9 of SAS/STAT software. These developments were motivated by three factors: requests and feedback from SAS users; advances in the field of statistics; and challenges posed by large data sets and complex data.

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 software provides a vehicle for introducing new functionality.

The features and syntax of experimental software are subject to change based on user feedback, and they are documented in papers which are available for download 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.

Note: At the time of this writing, complete information concerning experimental features in Release 9.1 of SAS/STAT software was not available. This information, with more details and examples, will be provided in the final version of this paper, which is planned for

download availability at <http://www.sas.com/statistics> to coincide with SUGI 28.

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 of 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 and monotone methods. For the monotone methods, a separate imputation model can be used for each imputed variable.

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 PARM= option without the associated COVB= or XPXI= option when the PARM= 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 from a DATA= data set. 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 often 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, there may be a small number of subjects per stratum, and thus a small number of subjects relative to the number of estimated parameters. Consequently, the sample size requirements for unconditional logistic regression based on maximum likelihood estimation may not be met.

Version 9 brings conditional logistic regression to the LOGISTIC procedure via a new STRATA statement. In the past, the PHREG procedure, which is intended for proportional hazards regression analysis, was often used for conditional logistic regression by taking advantage of special computational equivalences. This workaround is no longer necessary.

Robust Regression

Modern robust regression provides powerful techniques for dealing with outliers in regression analysis. Robust regression produces stable estimates in the presence of outliers, but it is more commonly used to detect and remove outliers, so that the analysis can proceed using traditional methods.

The types of outliers which can be addressed with robust regression include problems in the response direction, problems in the covariate space (leverage points), and problems in both directions. Three gen-

eral 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. All of these methods are available in the ROBUSTREG procedure, which is production software in Release 9.1. For more details concerning the ROBUSTREG procedure, see Chen (2002).

General Linear Models for Proportional Hazards

For many years, users have requested that a CLASS statement be implemented in the PHREG procedure. In Version 9, a CLASS statement is included in the TPHREG procedure, which is 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.

The PHREG procedure has been enhanced with new model-checking techniques due to Lin *et al.* (1993, 2002), which provide assessments of the functional form of a covariate and the validity of the proportional hazards assumption. This functionality, which is experimental in Release 9.1, is described by Johnston and So (2003a).

In Release 9.1, the PHREG procedure can be used to fit a proportional means regression model for the mean cumulative number (or cost) of events up to time t in the analysis of data from recurrent events. This application is described by Johnston and So (2003b).

Regression Diagnostics for Generalized Linear Models

In Release 9.1, new model-checking methods based on the cumulative residuals approach of Lin *et al.* (1993,2002), are available in the GENMOD procedure for assessing the functional form of a covariate in the linear predictor and the form of the link function in generalized linear models and marginal models for dependent responses (GEEs). This functionality is described by Johnston and So (2003a).

Survey Data Analysis

Many researchers use sample surveys to collect their information, relying on probability-based com-

plex sample designs such as stratified selection, clustering, and unequal weighting. To make statistically valid inferences, the analysis of the data must account for the design of the study. Traditional SAS procedures, such as the MEANS and GLM procedures, are inappropriate for this purpose because they compute statistics under the assumption of simple random sampling from an infinite population.

In Version 8, SAS/STAT introduced three procedures for sample survey selection and survey data analysis. The SURVEYSELECT procedure selects a probability-based sample and produces an output data set with selected units, selection probabilities, and 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, fitting linear models and computing regression coefficients and the covariance matrix.

In Version 9, SAS/STAT provides two new procedures for the analysis of sample survey data. The SURVEYFREQ procedure produces one-way to n-way frequency and crosstabulation tables for survey data. 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 2x2 tables, the procedure computes estimates of risk differences, odds ratios, relative risks, and their confidence limits.

The experimental SURVEYLOGISTIC procedure performs logistic regression for categorical responses in sample survey data. The analysis, which incorporates design aspects such as stratification and clustering, is based on theoretical work by Binder (1981, 1983) and Roberts, Rao, and Kumar (1987). The SURVEYLOGISTIC procedure provides much of the general flexibility of the LOGISTIC procedure.

Power and Sample Size Analysis

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 for a variety of statistical tests; it determines the sample size required to get a significant result with adequate probability and characterizes the power of a study to detect a meaningful effect. Analyses covered by PROC POWER include

means, proportions, correlation, 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.

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 new DISTANCE procedure, production in Release 9.1, 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.

Statistical Graphics

Effective graphical displays are essential in modern data analysis, statistical modeling, and data mining. Traditionally, SAS users have produced statistical graphics by modifying output data sets from statistical procedures and using generic graphics procedures. This forces the user to reconstruct analytical context that is available within the procedures.

Version 9 of SAS/STAT and SAS/ETS[®] introduces an extension to the Output Delivery System (ODS) which enables procedures to automatically create graphical displays. This approach is analogous to the way in which procedures create tables with ODS. The ODS template language is being extended with statements which control the layout and appearance of displays such as scatter plots, histograms, contour plots, and box-and-whisker plots. As with tables, statistical procedures create output objects which are bound with statistical graphics templates written by the procedure developer. The displays—which are integrated with tabular output—can then be rendered in several ODS destinations, including HTML. Although the primary goal of this work is to completely automate the production of displays which are commonly needed in statistical analysis, the user can customize the dis-

plays by modifying the templates or by changing the attributes of graph style elements.

Release 9.1 represents the first major step toward comprehensive support for automated statistical graphics using ODS. The subsets of SAS/STAT and SAS/ETS procedures which will offer this functionality in Release 9.1 will be described in the final version of this paper.

Parallelization

The Threaded Kernel (TK) architecture 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 REG and GLM procedures in SAS/STAT, as well as the DMREG procedure in SAS Enterprise Miner. For more details, see Cohen (2002).

Enhancements to Existing 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, these enhancements include the following:

- CLASS statement extension in PROC GENMOD comparable to the CLASS statement in PROC LOGISTIC
- SCORE statement in PROC LOGISTIC for scoring new data, which computes posterior probabilities and fit statistics
- performance enhancements for exact logistic computations in PROC LOGISTIC
- improved confidence intervals for the survivor function in PROC LIFETEST, which are based on transformations
- exact p -values for multivariate tests in PROC GLM

- exact confidence intervals for 2×2 tables in PROC FREQ
- exact confidence limits for the common odds ratio in PROC FREQ
- PARAM=REFERENCE option and improved syntax in PROC CATMOD
- sparse method of computing degrees of freedom in PROC LOESS
- stratified k-sample test in PROC LIFETEST using the GROUP= option in the STRATA statement
- new tests in PROC LIFETEST (Tarone-Ware, Peto-Peto, Fleming-Harrington G - ρ in addition to logrank and Wilcoxon tests)
- trend tests in PROC LIFETEST for detecting ordered alternatives of hazard rates

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.
- 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.
- Johnston, G. and So, Y. (2003a), "Let the Data Speak: New Regression Diagnostics Based on Cumulative Residuals", *Proceedings of the Twenty-Eighth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.
- Johnston, G. and So, Y. (2003b), "Analysis of Data from Recurrent Events", *Proceedings of the Twenty-Eighth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.
- Lin, D. Y., Wei, L. J., and Ying, Z. (1993), "Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals," *Biometrika*, 80, 557-572.

Lin, D. Y., Wei, L. J., and Ying, Z. (2002), "Model-Checking Techniques Based on Cumulative Residuals," *Biometrics*, 58, 1-12.

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. Springer, New York.

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

Contact Information

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

SAS, SAS/STAT, and SAS/ETS are registered trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.