SAS/STAT in SAS 9.4

Catching up to SAS/STAT 14.1

Moving to the SAS 9.4 platform from earlier releases to SAS 9.4 gives you access to a rich new set of additional SAS/STAT functionality. The enhancements of the 13.1, 13.2, and 14.1 releases are summarized below.

### Missing Data Analysis

Managing missing data properly has been a major direction for SAS. Recent releases have introduced key methodologies for missing data analysis.

#### Sensitivity Analysis

Evaluate how departures from the missing at random (MAR) assumption affect your inferences by using the new MNAR statement in the MI procedure, which imputes missing values by taking the pattern-mixture approach and assuming MNAR (missing not at random). By comparing inferential results for the latter values to results imputed under MAR, you can assess the sensitivity of your analysis to the MAR assumption.

#### Weighted Generalized Estimating Equations

The GEE procedure fits models to longitudinal data by using the generalized estimating equations (GEE) method of Liang and Zeger (1986). It also provides the weighted estimated equation method for handling missing data, which assumes the data are missing at random (MAR). The GEE procedure relies on syntax similar to that provided by the GENMOD procedure for fitting GEE models. PROC GEE implements observation-specific and subject-specific weighted estimating equations. Both weighted estimators provide unbiased and consistent estimates when data are missing at random.

#### Missing Survey Data: Imputation

The SURVEYIMPUTE procedure imputes missing values of an item in a sample survey by replacing them with observed values from the same item. Imputation methods include single and multiple hot-deck imputation and fully efficient fractional imputation (FEFI). Donor selection techniques include simple random selection with or without replacement, probability proportional to weights selection, and approximate Bayesian bootstrap selection. When you use FEFI, the procedure also produces imputation-adjusted replicate weights that can be used with any survey analysis procedure in SAS/STAT to estimate both the sampling variability and the imputation variability.

### Modern Survival Analysis

#### Interval Censored Data

The ICLIFETEST procedure performs nonparametric survival analysis for interval-censored data. You can use the ICLIFETEST procedure to compute nonparametric estimates of survival functions and to examine the equality of survival functions via statistical tests. PROC ICLIFETEST uses either a multiple imputation method or a bootstrap method to compute the standard errors of the survival estimates. It supports several transformation-based confidence intervals and produces survival plots. It provides:

- weighted generalized log-rank test
- weight functions for testing early or late differences
- stratified test for survival differences within predefined populations
- trend test for ordered alternatives
- multiple comparisons

The ICPHREG procedure fits proportional hazards regression models to interval-censored data. You can fit models that have a variety of configurations with respect to the baseline hazard function,
including the piecewise constant model and the cubic spline model. PROC ICPHREG maximizes the full likelihood instead of the Cox partial likelihood to estimate the regression coefficients. Standard errors of the estimates are obtained by inverting the observed information matrix that is derived from the full likelihood.

**Competing Risk Models**

The competing-risks model of Fine and Gray is now available in the PHREG procedure. Competing risks arise in the analysis of time-to-event data when the event of interest can be impeded by a prior event of a different type. In the presence of competing risks, the Kaplan-Meier method of estimating the survivor function is biased because you can no longer assume that a subject will experience the event of interest if the follow-up period is long enough. The proportional hazards model of Fine and Gray focuses on modeling the cumulative incidence of the event of interest.

In addition, the LIFETEST procedure performs nonparametric analysis of competing-risks data and the POWER procedure supports Cox proportional hazards regression models.

**Modern Statistical Models**

**Model Selection**

The new HPGENSELECT procedure is a high-performance procedure that provides model building for generalized linear models. It provides forward, backward, and stepwise variable selection and optionally chooses the best model based on the AIC, AICC, or SBC criterion. PROC HPGENSELECT fits models for standard distributions in the exponential family, such as the normal, Poisson, and Tweedie distributions. In addition, it fits multinomial models for ordinal and nominal responses, and it fits zero-inflated Poisson and negative binomial models for count data. The HPGENSELECT procedure provides the LASSO method.

The GLMSELECT procedure now provides the group LASSO method and enables you to apply either a safe screening method or a sure independence screening method to reduce a large number of regressors to a smaller subset from which model selection is performed.

**Generalized Additive Models**

The new GAMPL procedure fits generalized additive models by penalized likelihood estimation. Based on low-rank regression splines, these models are powerful tools for nonparametric regression and smoothing. Generalized additive models are extensions of generalized linear models. They relax the linearity assumption in generalized linear models by allowing spline terms in order to characterize nonlinear dependency structures.

With PROC GAMPL, each spline term is constructed by the thin-plate regression spline technique. A roughness penalty is applied to each spline term by a smoothing parameter that controls the balance between goodness of fit and the roughness of the spline curve. PROC GAMPL fits models for standard distributions in the exponential family, such as normal, Poisson, and gamma distributions.
Spatial Point Pattern Analysis

The new SPP procedure analyzes spatial point patterns. The goal of spatial point pattern analysis is to describe the occurrence of events (observations) that compose the pattern. PROC SPP enables you to specify the study area as a window, or you can rely on the input data coordinates to automatically compute a suitable study area by using the Ripley-Rasson window estimator. You can perform exploratory analysis of spatial point patterns by using the F, G, J, K, L, and PCF distance functions, which compare the empirical function distributions to the theoretical homogeneous Poisson process. PROC SPP provides:

- nonparametric intensity estimation by using different types of kernels
- adaptive kernel estimation
- parametric inhomogeneous Poisson process models to perform model validation

Bayesian Analysis

Updates to the MCMC Procedure

The MCMC procedure has been updated with new sampling algorithms for continuous parameters: the Hamiltonian Monte Carlo (HMC) and the No-U-Turn Sampler (NUTS). These algorithms use Hamiltonian dynamics to enable distant proposal in the sampling, making them efficient in many scenarios. These algorithms can lead to drastic improvements in sampling efficiency in many cases, resulting in fewer draws needed to achieve the same accuracy.

PROC MCMC now supports models that require lagging and leading variables, enabling you to easily fit models such as autoregressive models, dynamic linear models, and state space models. An ordinary differential equation solver and a general integration function have also been added, which enable the procedure to fit models that contain differential equations (for example, pharmacokinetic models) or models that require integration (for example, marginal likelihood models). And the PREDDIST statement in PROC MCMC now supports prediction from a marginalized random-effects model, which enables more realistic and useful prediction from many models.

Bayesian Discrete Choice Models

The new BCHOICE procedure performs Bayesian analysis for discrete choice models. Discrete choice models are used in marketing research to model decision makers’ choices among alternative products and services. The decision makers might be people, households, companies, and so on, and the alternatives might be products, services, actions, or any other options or items about which choices must be made. The collection of alternatives is known as a choice set, and when individuals are asked to choose, they usually assign a utility to each alternative.

The BCHOICE procedure provides Bayesian discrete choice models such as the multinomial logit, multinomial logit with random effects, and the nested logit. Varying numbers of alternatives in choice sets is allowed for logit models. The probit response function is also available. PROC BCHOICE obtains samples from the corresponding posterior distributions, produces summary and diagnostic statistics, and saves the posterior samples to an output data set that can be used for further analysis.

Item Response Theory Models

The new IRT procedure fits item response theory models. These models are widely used in education to calibrate and evaluate items in tests, questionnaires, and other instruments; they are used to score subjects on their abilities, attitudes, and other latent traits. In recent years, IRT models have also become increasingly popular in health behavior, quality of life, and clinical research.

PROC IRT supports several response models for binary and ordinal responses, including Rasch models; one-, two-, three-, and four-parameter
models; and graded response models with a logistic or probit link. PROC IRT also does the following:

- enables different items to have different response models
- performs multidimensional exploratory and confirmatory analysis
- performs multiple-group analysis
- estimates factor scores

### Classification and Regression Trees

Classification and regression trees are techniques used both in data mining and in standard statistical practice. Classification trees predict a categorical response, and regression trees predict a continuous response. Tree models partition the data into segments called nodes by applying splitting rules, which assign an observation to a node based on the value of one of the predictors. The partitioning is done recursively, starting with the root node that contains all the data, continuing down to the terminal nodes, which are called leaves. The resulting tree model typically fits the training data well, but might not necessarily fit new data well. To prevent overfitting, a pruning method can be applied to find a smaller subtree that balances the goals of fitting both the training data and new data.

The new HPSPLIT procedure creates classification and regression tree models. It provides choices of algorithms for both growth and pruning, a variety of options for handling missing values, whole and partial tree plots, cross validation plots, ROC curves, and partial tree plots.

### Performance Enhancements

The MCMC procedure is now multithreaded.

The new FASTQUAD option in the GLIMMIX procedure enables you to fit multilevel models that have been computationally infeasible. This option implements the multilevel quadrature algorithm of Pinheiro and Chao (2006), which means you can now fit multilevel models that would have been too difficult to fit previously.

The QUANTREG procedure now supports a new alternative interior point algorithm that can be more efficient for large data. Also, the PHREG procedure adds a FAST option that can speed up fitting of the Breslow and Efron partial likelihoods for the counting process style of input.

### Other Enhancements

Examples of the many other enhancements include:

- The GENMOD procedure now supports the Tweedie distribution.
- The NLIN procedure generates both bootstrap estimates of confidence intervals for the parameters and bootstrap estimates of the covariance matrix and correlation matrix of the parameter estimates.
- The NLMIXED procedure enables you to specify more than one RANDOM statement in order to fit hierarchical nonlinear mixed models.
- The FREQ procedure now provides score confidence limits for the odds ratio and the relative risk.
- The FREQ procedure provides exact mid-p, likelihood ratio, and Wald modified confidence limits for odds ratios.
- The NPAR1WAY procedure performs stratified rank-based analysis for two-sample data.
- The CALIS and FACTOR procedure produces path diagrams.

### For More Information

For complete information about all SAS/STAT releases, see the documentation available at support.sas.com/statdoc/