SAS/STAT® 14.1

Comprehensive Statistical Capabilities

Overview

SAS/STAT 14.1 introduces two new procedures and adds capabilities to many existing analyses. This software is available with the third maintenance release for Base SAS® 9.4.

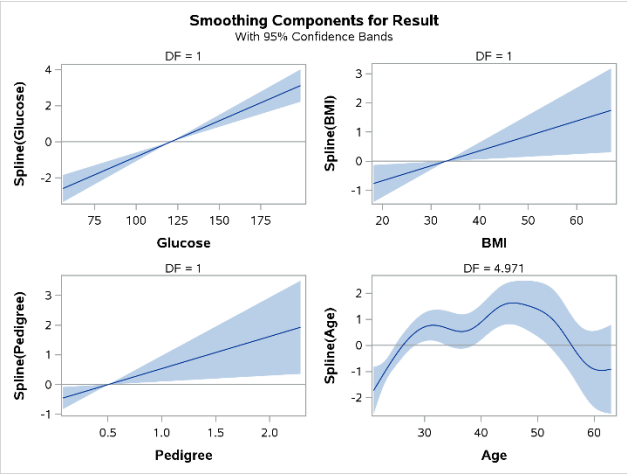
Missing Survey Data: Imputation

Nonresponse is a common problem in surveys. The resulting estimators suffer from nonresponse bias if the nonrespondents are different from the respondents. Estimators that use complete cases (only the observed units) might also be less precise. Imputation can reduce nonresponse bias and produce consistent results.

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.

Generalized Additive Models

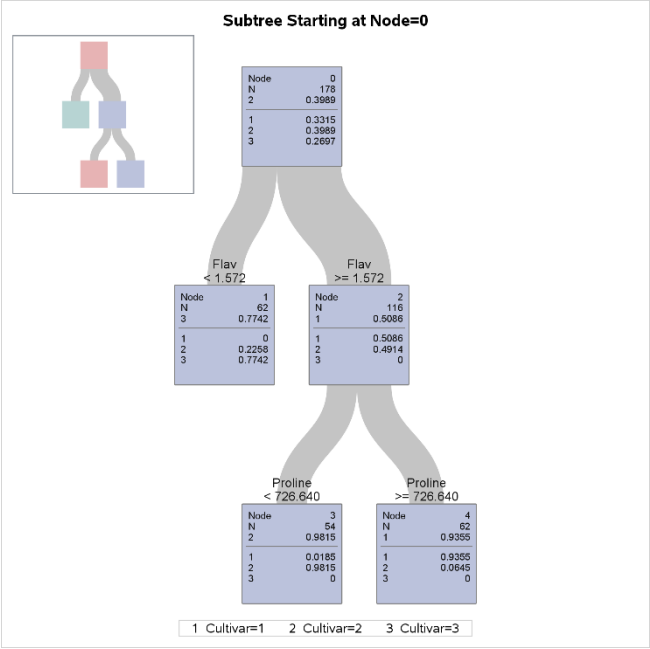
The new GAMPL procedure fits generalized additive models by penalized likelihood estimation. Based on low-rank regression splines, these models are powerful tools for nonparametric regression and smoothing. Generalized additive models are extensions of generalized linear models. They relax the linearity 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.



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 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

One of the performance improvements in SAS/STAT 14.1 is the new FASTQUAD option in the GLIMMIX procedure, which enables you to fit multilevel models that have been computationally infeasible in the past. 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 large or too slow to fit previously.

The QUANTREG procedure now supports a new alternative interior point algorithm that can be more efficient for large data. 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. The NLMIXED procedure becomes multithreaded in this release.

Bayesian Analysis

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.

Other Enhancements

- The BCHOICE procedure allows varying numbers of alternatives in choice sets for logit models.
- The FMM procedure supports the Dirichlet-multinomial distribution for a mixture component.
- The FREQ procedure provides exact mid-*p*, likelihood ratio, and Wald modified confidence limits for odds ratios.
- The GLMSELECT procedure provides the group LASSO method, and the HPGENESELECT procedure provides the LASSO method.
- The IRT procedure fits generalized partial credit models.
- The LIFETEST procedure performs nonparametric analysis of competing-risks data.
- The LOGISTIC procedure fits an adjacent-category logit model to ordinal response data.
- The NPAR1WAY procedure performs stratified rank-based analysis for two-sample data.
- The POWER procedure supports Cox proportional hazards regression models.

For More Information

SAS/STAT 14.1 is now available. For complete information see support.sas.com/statnewreleases/.