The following are highlights of the five updates on SAS 9.4, which is also when the SAS® High Performance Statistics procedures were included for single machine use.

### 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 that the data are missing at random (MAR). It implements observation-specific and subject-specific weighted estimating equations.

### 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, fractional 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 survey analysis procedures.

### Causal Inference

The PSMATCH procedure provides tools for propensity score analysis. It provides inverse probability and ATT weighting, stratification by propensity scores, and matching. PROC PSMATCH also provides methods for assessing the balance of baseline covariates and other variables in the treated and control groups after matching, weighting, or stratification.

The CAUSALTRT procedure estimates the average causal effect of a binary treatment, $T$, on a continuous or discrete outcome, $Y$. Depending on the application, the binary treatment variable $T$ can represent an intervention (such as smoking cessation versus control), an exposure to a condition (such as attending a private versus public school), or an existing characteristic of subjects (such as high versus low socioeconomic status). The CAUSALTRT procedure can estimate two types of causal effects: the average treatment effect (ATE) and the average treatment effect for the treated (ATT).

![Diagram of genetic variant, lung cancer, and smoking](image)

The CAUSALMED procedure estimates causal mediation effects from observational data. In causal mediation analysis, you have an outcome $Y$, a treatment $T$, and a mediator variable $M$ hypothesized to be causally affected by $T$ with a direct effect on $Y$. In addition, you have confounding variables $C$. The goal of the analysis is to estimate the direct causal effect of $T$ on $Y$ and the indirect causal effect of $T$ on $Y$ through $M$.

PROC CAUSALMED fits generalized linear models that have binary, negative binomial, Poisson, or normal distributions for the outcome and binary or normal distributions for the mediator. PROC CAUSALMED implements the regression approach of Valeri and VanderWeele (2013) and VanderWeele (2014).
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 evaluate the equality of survival functions. PROC ICLIFETEST offers multiple imputation and a bootstrap method to compute the standard errors of the survival estimates, and 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

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.

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. PROC PHREG also provides the alternative approach of cause-specific proportional hazards analysis. This involves applying the Cox model to the cause-specific hazard for each event type separately.

In addition, the LIFETEST procedure performs nonparametric analysis of competing-risks data.

Time-Dependent ROC Analysis

The PHREG procedure now provides concordance statistics and time-dependent ROC curves for assessing predictive accuracy. The ROC statement specifies a model to be used in computing concordance statistics or ROC curves. You can generate these ROC curves for the model in the MODEL statement and for a model specified in the ROC statement. Besides displaying the area under the ROC curve, you can display the difference in the AUC curves between models. You can also specify the time points for the computations.

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. PROC HPGENSELECT provides the LASSO method.

The GLMSELECT procedure now provides the group LASSO method, and you can specify a safe screening method or a sure independence screening method to reduce a large number of regressors to a smaller subset.

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, now multithreaded, has been updated with new sampling algorithms for continuous parameters: the Hamiltonian Monte Carlo (HMC) and the No-U-Turn Sampler (NUTS). These algorithms can lead to dramatic improvements in sampling efficiency in many cases.

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.

In addition, the new NORMALCAR option in the RANDOM statement specifies a spatial conditional autoregressive (CAR) prior that can be used to model spatial correlations among sites and neighbors.

**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 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 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
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.

The IRT procedure 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:

- 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.

Other Enhancements

- The GENMOD and GAMPL procedures now support the Tweedie distribution.
- The FASTQUAD option in the GLIMMIX procedure enables you to fit multilevel models that have been computationally infeasible.
- The QUANTREG procedure now supports a new alternative interior point algorithm that can be more efficient for large data.
- 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 provides score confidence limits for the odds ratio and also provides common risk differences and confidence limits.
- The CALIS and FACTOR procedures produce path diagrams.
- The NLMIXED and MCMC procedures provide a CMPTMODEL statement to specify compartment models for pharmacokinetic analysis.
- Bootstrap variance estimation is now available in the survey data analysis procedures.

For more information, see support.sas.com/statdoc/