Paper 185-31

You Can't Stop Statistics: SAS/STAT® Software Keeps Rolling Along

Maura Stokes, Robert Rodriguez, and R & D Staff
SAS Institute Inc.
Cary, North Carolina, USA

REALLY IMPORTANT CAVEAT: At the time of writing, the software was still under development. Syntax, results, and graphs contained in this paper are seriously likely to change before the production release of SAS 9.2.

INTRODUCTION

Release 9.2 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. SAS 9.2 got a head start with the release of additional SAS/STAT procedures via Web download. These procedures, GLIMMIX, QUANTREG, and GLMSELECT, were made available as experimental procedures only for SAS 9.1 for Windows, and the GLIMMIX procedure was made production in a subsequent release.

This paper describes the statistical development that will surface in SAS 9.2. First, the "post" SAS 9.1 download procedures will be reviewed, with examples from the GLIMMIX and GLMSELECT procedures. Attention turns next to a preview of SAS Stat Studio, the successor to SAS/INSIGHT software. ODS Graphics becomes production with SAS 9.2, and this paper reviews its many enhancements and describes the use of the new effect plot in PROC LOGISTIC.

One of the most exciting enhancements to SAS/STAT sofware is the first installment of facilities for Bayesian analyses in several procedures. These new capabilities are illustrated with examples from the PHREG and GENMOD procedures. The TTEST procedure emerges as one of the more updated procedures in SAS 9.2, and this paper includes an example of using PROC TTEST to perform a crossover analysis. New regressions diagnostics were introduced in the MIXED procedure in SAS 9.1, and this trend continues with the addition of diagnostics for GEE analysis in the GENMOD procedure. The usefulness of this new feature is illustrated with an example. Finally, the myriad of other SAS/STAT changes are highlighted. Please see the Papers section of www.sas.com/statistics for an updated version of this paper once SAS 9.2 has been released.

GENERALIZED LINEAR MIXED MODELS

The new GLIMMIX procedure fits generalized linear mixed models (GLMMs). Like linear mixed models fit by the MIXED procedure, GLMMs assume normal random effects. Conditional on these random effects, the data have a distribution in the exponential family. For example, the binary, binomial, Poisson, and negative binomial distributions are discrete members of this family. The normal, beta, gamma, and chi-square distributions are representatives of the continuous distributions in this family. In the absence of random effects, the GLIMMIX procedure fits generalized linear models, like those fit by the GENMOD procedure.

This version of the GLIMMIX procedure fits generalized linear mixed models based on linearization. The default estimation method in PROC GLIMMIX for models containing random effects is a technique known as restricted pseudo-likelihood (RPL) estimation (Wolfinger and O'Connell 1993). This technique employs an expansion around the current estimate of the best linear unbiased predictors of the random effects (METHOD=RSPL). Other methods implemented in the procedure enable you to change from a restricted (residual) pseudolikelihood to a maximum pseudolikelihood criterion and to change the locus of the expansion.

The GLIMMIX procedure offers a number of features, some of which are new to SAS mixed modeling tools:

- flexible covariance structures for random effects
- bias-adjusted empirical covariance estimators
- joint modeling for multivariate data
- Analysis of Means
- low-rank smoothing based on mixed models
- custom hypotheses about LS-means
- multiplicity adjustments
- SAS programming statements for computing model effects, or defining link and variance functions
- ODS graphics for LS-means display and comparison

The GLIMMIX procedure generalizes the MIXED and GENMOD procedures in two important ways. First, the response can have a nonnormal distribution. The MIXED procedure assumes that the response is normally distributed. Second, the GLIMMIX procedure incorporates random effects in the model and so allows for subject-specific (conditional) and population-averaged (marginal) inference. The GENMOD procedure only allows for marginal inference. The NLMIXED procedure also fits generalized linear mixed models, but the class of models it can accommodate is narrower. The NLMIXED procedure relies on approximating the marginal log likelihood by integral approximation through Gaussian quadrature. The GLIMMIX procedure, on the other hand, determines the marginal log likelihood as that of an approximate linear mixed model. This allows multiple random effects, nested and crossed random effects, multiple cluster types, and R-side random components. The MIXED procedure implements one method of optimization, a ridge-stabilized Newton Raphson algorithm. With the GLIMMIX procedure, you can choose from a variety of optimization methods. The default method for GLMMs is a Quasi-Newton algorithm. A ridge stabilized Newton-Raphson algorithm, akin to the method in the MIXED procedure, is also available.

The GLIMMIX procedure models all random components of the model through the RANDOM statement; it does not require a REPEATED statement to produce analyses similar to those produced by the REPEATED statement in the MIXED procedure.

EXAMPLE OF GENERATING ODDS RATIOS WITH PROC GLIMMIX

The GLIMMIX procedure was first released as an experimental download and then made production in the spring of 2005. An updated release was made available during the fall of 2005, and, besides maintenance, this release included a new facility for producing odds ratios for the appropriate models, including odds ratios in the presence of interactions.

The ODDSRATIO option in the MODEL statement produces estimates of odds ratios and their confidence limits when the link function is the logit, cumulative logit, or generalized logit. The procedure produces odds ratios for any classification main effects. In addition, it produces odds ratios for continuous effects in the model, producing them for fixed levels of a classification effect if the model contains an interaction between the continuous variable and the classification variable. PROC GLIMMIX also produces an odds ratio for continuous variables in the model that interact with other continuous variables in the model.

Odd ratios for effects involving CLASS variables are obtained by taking differences of least-squares means of the effect in question. If a continuous covariate is involved in the model, then that covariate is set to its average. When odds ratios are computed for a continuous covariate with an interaction with a classification variable, least-squares means are computed for that covariate average, say $\bar{x}$, and also for $\bar{x} + 1$. Differences are then computed for each level of the classification variable and exponentiated to produce the odds ratio. If no interaction is involved, a single odds ratio is produced that represents the odds of a one-unit shift in the covariate from $\bar{x}$.

For classification variables, you can control the level against which comparisons are made by using the DIFF= suboption; you can also change the reference value for the continuous variables with the AT suboption and specify the units of change with the UNIT= suboption. Note that you can also compute odds and odds ratios with the ODDSRATIO option of the LSMEANS statement and the EXP option of the ESTIMATE and LSMESTIMATE statements. You might need to revert to these statements to generate more customized odds ratio estimates.

The following example illustrates the use of this facility. These data are taken from the PROC LOGISTIC documentation and concern a study of the analgesic effects of treatments on elderly patients with neuralgia. The response is whether pain was reported, and the treatments are A, B, and Placebo. Patient age is recorded, as well as the duration of the pain episode. The following DATA step enters these data in SAS.

```
data Neuralgia;
      input Treatment $ Sex $ Age Duration Pain $ @@;
      datalines;
P  F  68   1  No   B  M  74  16  No  P  F  67  30  No
P  M  66  26  Yes  B  F  67  28  No  B  F  77  16  No
A  F  71  12  No   B  F  72  50  No  B  F  76   9  Yes
A  M  71  17  Yes  A  F  63  27  No  A  F  69  18  Yes
B  F  66  12  No   A  M  62  42  No  P  F  64   1  Yes
A  F  64  17  No   P  M  74   4  No  A  F  72  25  No
P  M  70   1  Yes  B  M  66  19  No  B  M  59  29  No
A  F  64  30  No   A  M  70  28  No  A  M  69   1  No
B  F  78   1  No   P  M  83   1  Yes B  F  69  42  No
B  M  75  30  Yes  P  M  77  29  Yes P  F  79  20  Yes
A  M  70  12  No   A  F  69  12  No  B  F  65  14  No
B  M  70   1  No   B  M  67  23  No  A  M  76  25  Yes
P  M  78  12  Yes  B  M  77   1  Yes B  F  69  24  No
P  M  66   4  Yes  P  F  65  29  No  P  M  60  26  Yes
A  M  78  15  Yes  B  M  75  21  Yes A  F  67  11  No
P  F  72  27  No   P  F  70  13  Yes A  M  75   6  Yes
B  F  65   7  No   P  F  68  27  Yes P  M  68  11  Yes
P  M  67  17  Yes  B  M  70  22  No  A  M  65  15  No
P  F  67   1  Yes  A  M  67  10  No  P  F  72  11  Yes
A  F  74   1  No   B  M  80  21  Yes A  F  69   3  No
;
run;
```

The following statements request a logistic regression analysis. Specifying ODDSRATIO in the MODEL statement of PROC GLIMMIX requests odds ratios. The SOLUTION option is used to display the parameter estimates.

```
proc glimmix data=Neuralgia;
  class Treatment Sex;
  model Pain= Treatment Age Treatment*Age /
              solution oddsratio;
  ods select ParameterEstimates OddsRatios;
run;
```

Figure 1 contains the resulting parameter estimates. Treatment appears to be important, as does the age*treatment interaction.

The GLIMMIX Procedure

Parameter Estimates

Effect	Pain	Treatment	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	No		1.5662	6.7769	54	0.23	0.8181
Treatment		A	36.2610	18.7283	54	1.94	0.0581
Treatment		B	37.1456	20.4310	54	1.82	0.0746
Treatment		P	0	.	.	.	.
Age			-0.03820	0.09728	54	-0.39	0.6961
Age*Treatment		A	-0.4804	0.2626	54	-1.83	0.0729
Age*Treatment		B	-0.4733	0.2741	54	-1.73	0.0899
Age*Treatment		P	0	.	.	.	.

Figure 1. Parameter Estimates

Figure 2 contains the odds ratio estimates. The columns "Treatment" and "_Treatment" tell you the levels of comparison for the treatment level, and the columns "Age" and "_Age" do the same for age. For example, the value 13.617 is the odds ratio for Treatment A compared to Treatment P. It means that those patients receiving Treatment A are 13.617 times more likely to obtain relief than those patients receiving the placebo. Since there is an age*treatment interaction, there are three odds ratios displayed for age, one for each level of treatment. For example, the value 0.595 means that a one-unit shift in age for those patients receiving Treatment A resulted in 0.595 greater odds of relief (or 1.68 greater odds of not obtaining relief). Increasing age appears to have a marginally negative effect on treatment.

Odds Ratio Estimates

Treatment	Age	_Treatment	_Age	Estimate	DF	95% Confidence Limits	
A	70.05	P	70.05	13.617	54	2.070	89.578
B	70.05	P	70.05	54.150	54	2.284	>999.999
A	71.05	A	70.05	0.595	54	0.365	0.971
B	71.05	B	70.05	0.600	54	0.359	1.002
P	71.05	P	70.05	0.963	54	0.792	1.170

Effects of continuous variables are assessed as one unit offsets from the mean. The AT suboption modifies the reference value and the UNIT suboption modifies the offsets.

Figure 2. Parameter Estimates

UPCOMING FEATURES IN PROC GLIMMIX

The GLIMMIX procedure is still in active development. In SAS 9.2, PROC GLIMMIX includes maximum likelihood estimation by Laplace approximation, stepdown multiple comparison tests, the COVTEST statement for likelihood-based testing and confidence intervals for covariance parameters, TYPE=LIN and other important covariance structures from the MIXED procedure, and mixed model penalized B-splines, also known as P-Splines.

For more information on PROC GLIMMIX, see the documentation and the 2004 SUGI paper "Introducing the GLIMMIX Procedure for Generalized Linear Models" by Oliver Schabenberger. To download this procedure for use with your SAS 9.1 Windows release, see www.sas.com/statistics.

QUANTILE REGRESSION

The QUANTREG procedure models the effects of covariates on the conditional quantiles of a response variable by means of quantile regression. Ordinary least-squares regression models the relationship between one or more covariates X and the conditional mean of the response variable Y given X=x. Quantile regression extends the regression model to conditional quantiles of the response variable, such as the 90th percentile. Quantile regression is particularly useful when the rate of change in the conditional quantile, expressed by the regression coefficients, depends on the quantile.

The main advantage of quantile regression over least-squares regression is its flexibility for modeling data with heterogeneous conditional distributions. Data of this type occur in many fields, including biomedicine, econometrics, survival analysis, and ecology. Quantile regression provides a complete picture of the covariate effect when a set of percentiles is modeled, and it makes no distributional assumption about the error term in the model. Consequently, quantile regression offers considerable model robustness. In addition, quantile regression, which includes median regression, is robust to extreme points in the response direction.

Quantile regression is also flexible in the sense that it does not involve a link function that relates the variance and the mean of the response variable. Generalized linear models, which you can fit with the GENMOD procedure, require both a link function and a distributional assumption such as the normal or Poisson distribution. The goal of generalized linear models is inference about the regression parameters in the linear predictor for the mean of the population. In contrast, the goal of quantile regression is inference on regression coefficients for the conditional quantiles of a response variable that is usually assumed to be continuous.

The QUANTREG procedure:

- Implements the simplex, interior point, and smoothing algorithms for estimation
- Provides three methods to compute confidence intervals for the regression quantile parameter: sparsity, rank, and resampling.
- Provides two methods to compute the covariance and correlation matrices of the estimated parameters: an asymptotic method and a bootstrap method
- Provides two tests for the regression parameter estimates: the Wald test and a likelihood ratio test
- Uses robust multivariate location and scale estimates for leverage point detection
- Provides multithreaded processing when multiple processors are available

The QUANTREG procedure becomes production with the SAS 9.2 release. For more information, see the documentation and the 2005 SUGI paper “An Introduction to Quantile Regression and the QUANTREG Procedure” by Lin Chen. To download this procedure for use with your SAS 9.1 Windows release, see www.sas.com/statistics.

MODEL SELECTION

In recent years, the increasing amount of information being captured and stored means that statistical model building can involve thousands of potential predictors. Reducing the predictors to a reasonable subset makes sense, especially for models intended to be used for prediction purposes. Statistical model selection attempts to provide the subsets of variables that work, even though it relies on various definitions of the “best” model and heuristic procedures for approximating the true but computationally infeasible solution.

The GLMSELECT procedure performs effect selection in the framework of general linear models. A variety of model selection methods are available, including the LASSO method of Tibshirani (1996) and the related LAR method of Efron et al. (2004). The GLMSELECT procedure offers extensive capabilities for

customizing the selection with a wide variety of selection and stopping criteria. Methods include traditional and computationally efficient significance-level-based criteria to more computationally intensive validation-based criteria. The procedure also provides graphical summaries of the selection search.

Note that while the model selection question seems reasonable, trying to answer it for real data can lead to various issues such as the influence of outliers and the fact that any particular method might not be finding the "best" model according to another method's criterion. To some degree, these pitfalls are intrinsic, and some statisticians eschew model selection processes altogether. However, certain features of GLMSELECT, such as its numerous capabilities for customizing the selection, can help to minimize these pitfalls. See Cohen (2006) for further discussion of these issues.

PROC GLMSELECT focuses on the standard general linear model for univariate responses and offers great flexibility for and insight into the model selection algorithm. It handles both continuous and classification model effects. You can then take the model it selects and perform a more detailed analysis with either the REG or GLM procedure. Note that the GLMSELECT procedure can perform selection from a very large number of effects (tens of thousands).

Highlights of the GLMSELECT procedure include:

- multiple effect selection methods based on a variety of selection criteria
- stopping rules based on a variety of model evaluation criteria
- leave-one-out and k-fold cross-validation
- support for crossed and nested effects including hierarchy among effects
- internal partitioning of data into training, validation, and testing roles
- graphical representation of the selection process
- parallel processing of BY-groups

PROC GLMSELECT EXAMPLE

The following data set, from the GLMSELECT documentation, contains salary and performance information for Major League baseball players who played at least one game in both the 1986 and 1987 seasons, excluding pitchers. The salaries (*Sports Illustrated*, April 20, 1987) are for the 1987 season and the performance measures are from 1986 (*Collier Books, The 1987 Baseball Encyclopedia Update*).

```
data baseball;
   length name $ 18;
   length team $ 12;
   input name $ 1-18 nAtBat nHits nHome nRuns nRBI nBB
      yrMajor crAtBat crHits crHome crRuns crRbi crBB
      league $ division $ team $ position $ nOuts nAssts
      nError salary;
   datalines;
Allanson, Andy       293    66     1    30    29    14
                       1   293    66     1    30    29    14
                    American East Cleveland C 446 33 20 .
Ashby, Alan          315    81     7    24    38    39
                      14  3449   835    69   321   414   375
                    National West Houston C 632 43 10 475
Davis, Alan          479   130    18    66    72    76
                       3  1624   457    63   224   266   263
                    American West Seattle 1B 880 82 14 480
Dawson, Andre        496   141    20    65    78    37
                      11  5628  1575   225   828   838   354
```

```
                    National East Montreal RF 200 11 3 500
...
run;
```

Suppose you want to investigate whether you can model player salaries for the 1987 season based on performance measures for the previous season. The aim is to find a succinct model that does not overfit these particular data, so it can be useful for prediction. Since salary variation is greater for the higher salaries, it makes sense to do a log transform on salary. This is performed in the previous DATA step.

The following statements request a standard stepwise effect selection. Eighteen candidate explanatory variables are listed in the MODEL statement, including classification variables league and division. In order to show an analysis similar to one you could perform with PROC REG, the SELECT=SL suboption is supplied to the SELECTION=STEPWISE option so that effects are entered according to an SLE < 0.15 and an SLS > 0.15. The final selected model is one in which no more effects can enter or leave the model according to these criteria. The PLOTS=ALL option in the PROC statement requests all available graphics. The DETAILS=ALL option requests entry and removal information, ANOVA table and parameter estimates, and model candidate information for each step, in addition to selection summary information. The STATS=ALL option requests that all model fit criteria be included in fit summary tables and fit statistics tables.

```
ods graphics on;
proc glmselect data=baseball plots=all;
   class league division;
   model logSalary = nAtBat nHits nHome nRuns nRBI nBB
      yrMajor crAtBat crHits crHome crRuns crRbi
      crBB league division nOuts nAssts nError
      / details=all selection=stepwise(select=sl)stats=all;
run;
ods graphics off;
```

Figure 3 contains information about the selection method, the selection and stopping criteria, and the entry and exit significance levels.

```
                    The GLMSELECT Procedure

Data Set                                  WORK.BASEBALL
Dependent Variable                            logSalary
Selection Method                               Stepwise
Select Criterion                     Significance Level
Stop Criterion                       Significance Level
Entry Significance Level (SLE)                     0.15
Stay Significance Level (SLS)                      0.15
Effect Hierarchy Enforced                          None

      Number of Observations Read         322
      Number of Observations Used         263

             Class Level Information

      Class         Levels    Values

      league             2    American National
      division           2    East West
```

Figure 3. Model Selection Process Summary

The first variable entered into the model is crRuns. Figure 4 contains the ANOVA table for the model at Step 1. It also displays various fit statistics.

The GLMSELECT Procedure
Stepwise Selection: Step 1
Effect Entered: crRuns

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	1	86.73997	86.73997	188.01
Error	261	120.41377	0.46136	
Corrected Total	262	207.15373		

Root MSE	0.67923
Dependent Mean	5.92722
R-Square	0.4187
Adj R-Sq	0.4165
AIC	-201.46090
AICC	0.24195
BIC	-200.78721
C(p)	111.23152
PRESS	123.91950
SBC	-194.31659
ASE	0.45785

Figure 4. Analysis of Variance

Figure 5 shows the parameter estimates.

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	5.294720	0.062306	84.98
crRuns	1	0.001718	0.000125	13.71

Figure 5. Parameter Estimates at Step 1

The GLMSELECT procedure provides a variety of graphics that help you visualize the selection process. Figure 6 displays the Needle Plot, a visual panorama of the Entry values (log p-value) for the model candidates at Step 2. A table of candidates from the remaining effects is also displayed at each step of the selection process. Variable Hits would appear to be the next variable selected for the model.

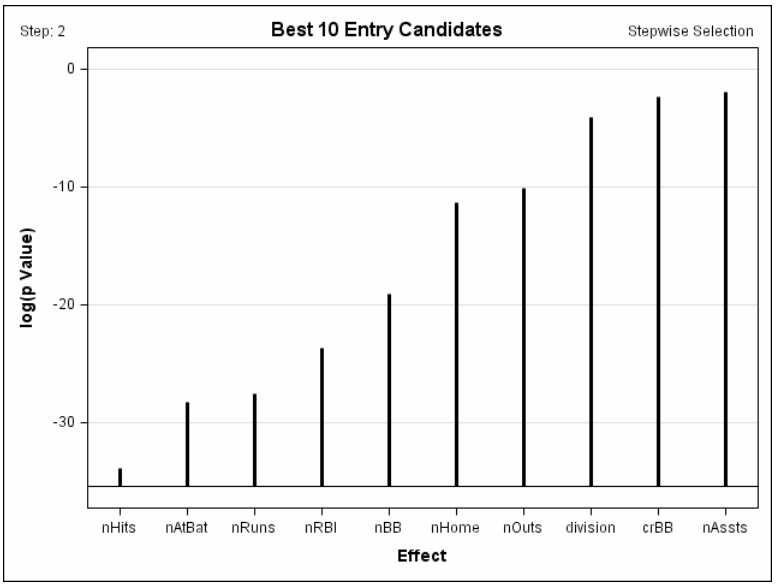

Figure 6. Needle Plot of Entry Candidates at Step 2

This selection process continues, and, in this case, leads to a final model with 8 explanatory variables in 10 steps. Figure 7 contains the final model parameter estimates table.

Parameter Estimates

Parameter		DF	Estimate	Standard Error	t Value
Intercept		1	4.159467	0.150425	27.65
nAtBat		1	−0.001563	0.000945	−1.65
nHits		1	0.010392	0.003004	3.46
nBB		1	0.010191	0.002799	3.64
yrMajor		1	0.074428	0.019046	3.91
crHits		1	0.000434	0.000184	2.36
crBB		1	−0.000613	0.000380	−1.61
division	East	1	0.152959	0.071016	2.15
division	West	0	0	.	.
nOuts		1	0.000227	0.000133	1.70

Figure 7. Final SL Model Parameter Estimates

However, you might feel that another model fit criterion is more appropriate. One of the plots offered by the GLMSELECT procedure is a criterion panel, which displays all model fit statistics available for assessing the sequence of models in the selection process. This plot is generated as part of the graphs produced with the PLOTS=ALL option in the PROC statement, or it can be requested specifically with PLOTS=CRITERIONPANEL option.

Figure 8 displays the Criterion Panel for this analysis.

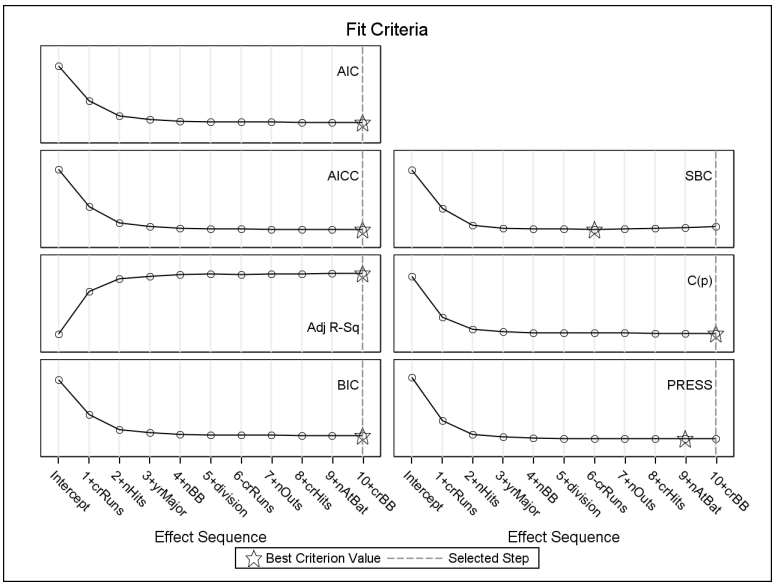

Figure 8. Criterion Panel

Note that the stepwise selection process would end earlier if either the SBC (Schwarz Bayesian Information Criterion) or PRESS (Predicted Residual Sum of Squares) criterion was employed to determine the final model. (Note that the SBC criterion is the default for PROC GLMSELECT if no SELECTION= method is specifiied.)

The following statements request a stepwise selection that uses the PRESS statistic for determining the final model. The suboption CHOOSE=PRESS of the SELECTION= option specifies the PRESS statistic for that purpose.

```
proc glmselect data=baseball plots=all;
   class league division;
   model logSalary = nAtBat nHits nHome nRuns nRBI nBB
                yrMajor crAtBat crHits crHome crRuns crRbi
                crBB league division nOuts nAssts nError
               / selection=stepwise(select=sl choose=press) stats=all;
run;
```

Figure 9 and Figure 10 display the ANOVA table and the parameter estimates for the model chosen at Step 9.

Selected Model

The selected model, based on PRESS, is the model at Step 9.

Effects: Intercept nAtBat nHits nBB yrMajor crHits division nOuts

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	7	124.67715	17.81102	55.07
Error	255	82.47658	0.32344	
Corrected Total	262	207.15373		

Root MSE	0.56872
Dependent Mean	5.92722
R-Square	0.6019
Adj R-Sq	0.5909
AIC	-288.98522
AICC	-0.08849
BIC	-286.39410
C(p)	6.58753
PRESS	88.55275
SBC	-260.40799
ASE	0.31360

Figure 9. Analysis of Variance at Step 9 for PRESS Statistic Criterion Model

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	4.176133	0.150539	27.74
nAtBat	1	-0.001468	0.000946	-1.55
nHits	1	0.011078	0.002983	3.71
nBB	1	0.007226	0.002115	3.42
yrMajor	1	0.070056	0.018911	3.70
crHits	1	0.000247	0.000143	1.72
division East	1	0.143082	0.070972	2.02
division West	0	0	.	.
nOuts	1	0.000241	0.000134	1.81

Figure 10. Parameter Estimates at Step 9 for PRESS Statistic Criterion Model

Model selection is often performed to find a model that will then be used for prediction purposes. One problem in model selection is that you might be overfitting the data, which could lead to poor predictive model performance. The GLMSELECT procedure provides several methods for overcoming these problems. One such technique is cross validation, where you subdivide the data into k parts and obtain models for each of k subsets by omitting one part at a time. The selection criterion used is called CV PRESS, computed by summing the squares of the residuals when each of the submodels is scored on the data omitted in fitting the submodel.

When you have sufficient data, you can also reserve part of the data for testing your model. The GLMSELECT procedure also has provides techniques for partitioning your data into disjoint subsets for training, validation, and testing roles.

For more information on PROC GLMSELECT, see the paper in these proceedings, “Introducing the

GLMSELECT Procedure for Model Selection," by Robert Cohen (Cohen, 2006), as well as the documentation. To download this procedure for use with your SAS 9.1 Windows release, see www.sas.com/statistics.

ANNOUNCING NEW SOFTWARE: SAS STAT STUDIO

Many users are familiar with the dynamic graphics and interactive modeling available with the SAS/INSIGHT product. SAS Stat Studio is the successor to SAS/INSIGHT. It combines the power of SAS/STAT for statistical modeling with the interactivity of SAS/INSIGHT. SAS Stat Studio takes advantage of the SAS/IML matrix programming language and new extensions so that you can implement new analysis methods yourself and have immediate access to dynamic graphics for displaying the results. Figure 11 displays its dynamically-linked data table and graphics, program editor, and statistical results from SAS/STAT.

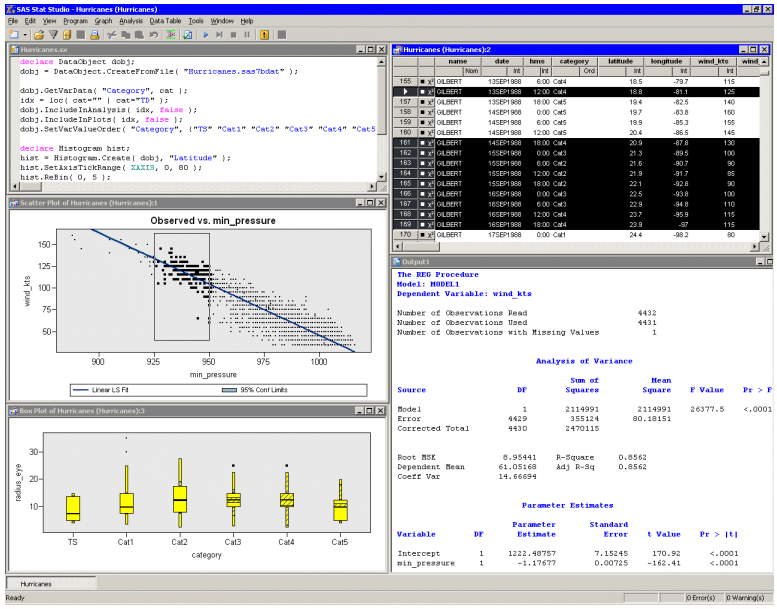

Figure 11. Upcoming SAS Stat Studio

The SAS Stat Studio client runs on a Windows desktop. The client can connect with one or more servers, and data can reside on the client or the server. Stat Studio point-and-click capabilities for graphical exploration and modeling, as well as its programming interface allows you to:

- identify structure and features in multivariate data
- experiment with standard modeling techniques
- implement new analysis methods
- transform data and pass output across multiple methods
- display SAS/STAT results with dynamically linked interactive graphics
- validate models and link outliers to the original data

SAS Stat Studio is an exciting addition to the data analysis arsenal for both existing SAS/INSIGHT users who want a more modern interface, as well as SAS/STAT users who want to extend and customize the statistical techniques in SAS/STAT.

ODS STATISTICAL GRAPHICS TAKES CENTER STAGE

SAS 9.1 introduced an experimental extension to the Output Delivery System (ODS), which enabled over two dozen SAS/STAT and SAS/ETS procedures to create statistical graphics as automatically as they create

tables. This extension, referred to as ODS Graphics, requires minimal additional syntax, and it provides displays commonly needed for data analysis and statistical modeling. With the production release of ODS Graphics in SAS 9.2, over 50 procedures in SAS/STAT and SAS/ETS now offer ODS Graphics. Support for ODS Graphics is also being added to two Base SAS statistical procedures, FREQ and UNIVARIATE, and is planned for SAS/QC software.

Just a few of the new graphics include:

- effect plots in PROC LOGISTIC
- contour and ridge plots in PROC RSREG
- plot of odds ratios over strata in PROC FREQ
- residual, predicted value, and regression diagnostic plots in PROC GENMOD
- means comparisons and analysis of means plots for LS-means in PROC GLM and PROC GLIMMIX
- display of the number of subjects at risk in the plot of Kaplan-Meier curves in PROC LIFETEST
- graph of smoothed hazard functions ungrouped data in PROC LIFETEST
- effects plots for binary and polytomous responses in PROC LOGISTIC
- ROC plots for comparing models in PROC LOGISTIC
- histograms, box plots, plots for paired and crossover data in PROC TTEST

New SAS/GRAPH procedures targeted at statistical graphics also take advantage of ODS Graphics functionality. Additional ODS styles designed for statistical work are being added, and a new interactive graphics editor allows you to make immediate changes to graphs, such as modifying titles and annotating points. See "Creating Statistical Graphics in SAS 9.2: What Every Statistical User Should Know" (Rodriguez and Balan, 2006) in these proceedings for more information.

EFFECT PLOT IN PROC LOGISTIC

This example illustrates the use of one of the new plots in SAS/STAT software in SAS 9.2. The LOGISTIC procedure now provides an effects plot, which allows you to visually compare the influence of each covariate by examining visual representations of predicted probabilities for each combination of covariate values. The data are taken from Stokes, Davis, and Koch (2000), Section 8.4. Patients with either a complicated or uncomplicated diagnosis of urinary tract infections were given one of three treatments.

The following statements create SAS data set one.

```
data one;
   length Diagnosis $ 13;
   input Diagnosis $ Treatment $ Cured Notcured;
   N= Cured+Notcured;
   datalines;
Complicated A 78 28
Complicated B 101 11
Complicated C 68 46
Uncomplicated A 40 5
Uncomplicated B 54 5
Uncomplicated C 34 6
;
```

The following SAS program fits a main effects model with treatment and diagnosis as the main effects. The EFFECT suboption is specified to produce an effect plot. In this example, the x-axis will display all combinations of treatment and diagnosis, and the y-axis will display the predicted probabilities. The CLBAND suboption requests the 95% confidence limits that, for these discrete axes, are displayed as the extents of a box plot. The YVIEW suboption defines the range of the y-axis.

```
ods graphics on;
proc logistic data=one  plots=(effect(clband yview=(.5,1)));
   class Treatment Diagnosis / param=ref;
   model Cured/N= Diagnosis Treatment;
   ods select effectplot;
run;
ods graphics off;
```

Figure 12 displays the predicted probabilities for each combination of treatment and diagnosis. You can see from this graphic that, controlling for treatment, an uncomplicated diagnosis has a higher probability of being cured. Also, for both types of diagnoses, patients on Treatment B seem to have a higher expected cure-rate than patients on Treatment A, and patients on Treatment C have the worst probability of being cured—this is more apparent in Figure 13. Note that you still need to perform tests to assess the significance of these differences.

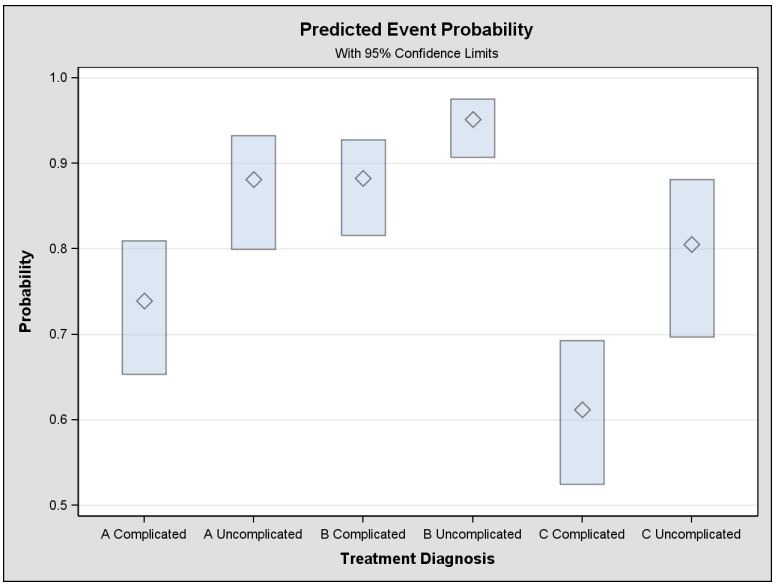

Figure 12. Binary Logistic Predicted Probabilities Plot Created with PROC LOGISTIC

You can create a different view of the predicted probabilities to make it easier to see these results. In the following program, the SLICEBY=TREATMENT option groups the predicted probabilities by treatment, while the x-axis will display each of the diagnosis levels, and the CLBAR option displays the confidence limits of the predicted probabilities as error bars. The CONNECT option draws a line connecting the predicted probabilities for each diagnosis by treatment combination. The results are displayed in Figure 13.

```
ods graphics on;
proc logistic data=one plots=(effect(sliceby=Treatment clbar connect yview=(.5,1)));
   class Treatment Diagnosis / param=ref;
   model Cured/N= Diagnosis Treatment;
   ods select effectplot;
run;
ods graphics off;
```

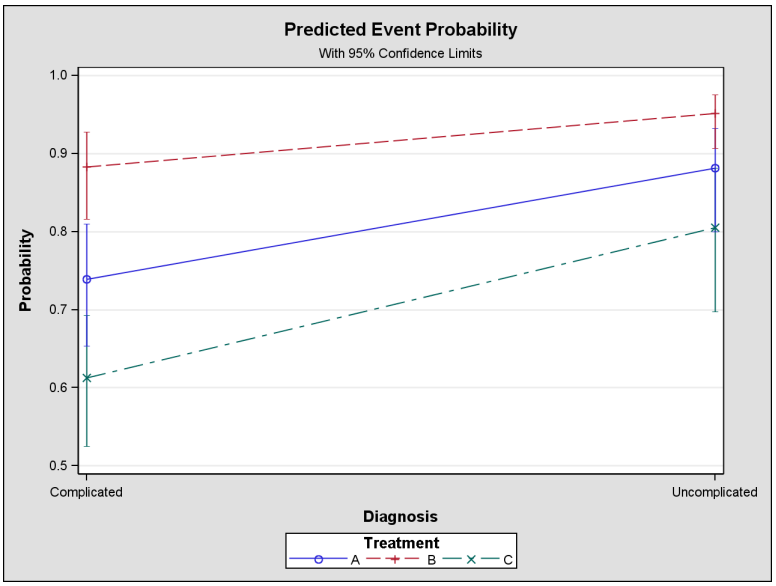

Figure 13. Binary Logistic Predicted Probabilities with SLICEBY=Treatment

BAYESIAN ANALYSIS IN SAS/STAT SOFTWARE

SAS 9.2 delivers the first installment of Bayesian software development to SAS/STAT users. The use of Bayesian techniques is growing rapidly among statisticians, largely due to computational advances that have taken place in the last 15 years. SAS 9.2 introduces modern Bayesian analysis to the PHREG, LIFEREG, and GENMOD procedures.

The Bayesian approach to statistical inference includes the incorporation of prior knowledge into making inferences on unknown quantities in a model. Often, these unknown quantities are parameters of interest. In a frequentist analysis, a parameter θ in a statistical model is considered to be a fixed constant. In a Bayesian analysis, θ is considered to be random; uncertainty concerning the parameter is expressed through probability statements, such as distributions.

Oversimplified, the following steps are involved in performing a Bayesian analysis:

1. The *prior* distribution for θ, $p(\theta)$ is specified. This expresses belief about the parameter before the data are examined.
2. Given observed data, say $\mathbf{x}$, you choose a statistical model $p(\mathbf{x}|\theta)$ that describes the distribution of $\mathbf{x}$ given θ. You can think of $p(\mathbf{x}|\theta)$ as a likelihood function.
3. You update your beliefs about θ by combining information from the prior distribution and the data through calculation of the *posterior* distribution, $p(\theta|\mathbf{x})$.

This last step is carried out by using Bayes' theorem; hence the term Bayesian analysis. Bayesian inference about θ is based on the posterior distribution of θ. You can summarize this distribution by finding point estimates, testing hypotheses, or constructing probability statements through *credible intervals*. In most cases, $p(\theta|\mathbf{x})$ doesn't have a closed form, making it difficult to obtain the posteriors, especially with arbitrary prior distributions and likelihood functions. Simulation is thus an important aspect of Bayesian analysis, since it provides a way for direct estimation of the posterior density. Markov Chain Monte Carlo (MCMC) methods are widely used.

Obviously, there is much more to Bayesian statistics than can be outlined in this paper. Fortunately, many textbooks now address this topic. See Berry (1996) and Lee (2004) for introductory treatments, and see Box and Tiao (1992) for an intermediate treatment. Gelman et al. (2004) provide extensive statistical

applications, and Spiegelhalter, et al. (2004) discuss clinical trials data analysis. Ibrahim et al. (2001) discuss Bayesian survival analysis.

The following examples provide a first look at these new facilities in the PHREG and GENMOD procedures.

BAYESIAN ANALYSIS FOR THE PIECEWISE EXPONENTIAL MODEL IN PROC PHREG

Consider the results of a small randomized trial on rats. Suppose you randomize 40 rats that have been exposed to a carcinogen into two treatment groups (Drug X and Placebo). The event of interest is death from cancer induced by the carcinogen. The response is the time from randomization to death. Four rats died of other causes; their survival times are regarded as censored observations. Interest lies in evaluating the effects of treatment and gender on survival.

The data set Exposed contains four variables: Days (survival time in days from treatment to death), Status (censoring indicator variable: 0 if censored and 1 if not censored), Treatment (treatment indicator: 0 if Placebo and 1 if Drug X), and Sex (gender: 0 if female and 1 if male).

```
data Exposed;
   input Days  Status Treatment Sex @@;
   datalines;
179   1   1   0   378   0   1   1
256   1   1   0   355   1   1   1
262   1   1   1   319   1   1   1
256   1   1   0   256   1   1   1
255   1   1   1   171   1   1   0
224   0   1   0   325   1   1   1
225   1   1   0   325   1   1   1
287   1   1   1   217   1   1   0
319   1   1   1   255   1   1   0
264   1   1   1   256   1   1   0
237   0   0   0   291   1   0   1
156   1   0   0   323   1   0   1
270   1   0   1   253   1   0   1
257   1   0   1   206   1   0   0
242   1   0   1   206   1   0   0
157   1   0   0   237   1   0   1
249   1   0   1   211   1   0   0
180   1   0   0   229   1   0   0
226   1   0   0   234   1   0   0
268   0   0   1   209   1   0   0
;
```

The BAYES statement in PROC PHREG invokes a Bayesian analysis, and the PIECEWISE= option in the BAYES statement stipulates the piecewise exponential model. Several diagnostic plots are produced by default.

```
ods graphics on;
proc phreg data=Exposed;
   model Days*Status(0)=Treatment Sex;
   bayes piecewise=loghazard;
run;
ods graphics off;
```

Associated with each interval of constant hazard is a parameter representing the constant hazard or the log of the constant hazard. The latter is specified by PIECEWISE=LOGHAZARD, which is used in this example.

Numerically it is more stable to use the log-hazard parameters than the hazard parameters. By default, PROC PHREG partitions the time axis into 8 intervals with approximately equal number of uncensored observations per interval (Figure 14). Note that the log-hazard parameters are named Alpha1,Alpha2, . . ., Alpha8.

Figure 14 displays the hazard time intervals.

The PHREG Procedure

Bayesian Analysis

Constant Hazard Time Intervals

Interval [Lower,	Upper)	N	Event	Log Hazard Parameter
0	193	5	5	Alpha1
193	221	5	5	Alpha2
221	239.5	7	5	Alpha3
239.5	255.5	5	5	Alpha4
255.5	256.5	4	4	Alpha5
256.5	278.5	5	4	Alpha6
278.5	321	4	4	Alpha7
321	Infty	5	4	Alpha8

Figure 14. Intervals with Constant Baseline Hazards

There are 10 parameters in this piecewise exponential model, 8 log-hazard parameters and 2 regression coefficients, one for Treatment and the other for Sex. Figure 15 shows the results of the maximum likelihood estimation.

Bayesian Analysis

Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits	
Alpha1	1	-6.2975	0.4575	-7.1942	-5.4008
Alpha2	1	-3.9290	0.4646	-4.8397	-3.0183
Alpha3	1	-2.8045	0.4890	-3.7630	-1.8461
Alpha4	1	-1.8072	0.6044	-2.9919	-0.6225
Alpha5	1	1.3639	0.6729	0.0451	2.6828
Alpha6	1	-0.7888	0.7889	-2.3350	0.7575
Alpha7	1	-0.8918	0.8182	-2.4954	0.7119
Alpha8	1	0.5455	0.8932	-1.2051	2.2961
Treatment	1	-1.2164	0.3967	-1.9939	-0.4389
Sex	1	-2.6039	0.5418	-3.6658	-1.5420

Figure 15. Maximum Likelihood Estimates of the Log-Hazard Parameters and the Regression Coefficients

By default, PROC PHREG uses the noninformative prior distributions for the log-hazard parameters and the regression coefficients, but you can also choose a normal prior for these parameters instead. A chain of 12,000 iterations is generated using the Gibbs sampler, with the first 2,000 iterations as burn-in. You can alter the number of burn-in and the number of posterior samples by specifying the NBI= and NMC= options, respectively. Summary statistics of the posterior samples are shown in Figure 16 and Figure 17. The results are quite comparable to those based on maximum likelihood (Figure 15).

Bayesian Analysis

Descriptive Statistics of the Posterior Samples

Parameter	N	Mean	Standard Deviation	Quantiles 25%	50%	75%
Alpha1	10000	−6.4163	0.4859	−6.7131	−6.3824	−6.0765
Alpha2	10000	−4.0558	0.4899	−4.3627	−4.0268	−3.7058
Alpha3	10000	−2.9280	0.5043	−3.2467	−2.8941	−2.5776
Alpha4	10000	−1.9046	0.6245	−2.3128	−1.8833	−1.4739
Alpha5	10000	1.2445	0.6963	0.8068	1.2656	1.7137
Alpha6	10000	−0.8493	0.8281	−1.3860	−0.8540	−0.3014
Alpha7	10000	−0.9455	0.8508	−1.5003	−0.9394	−0.3804
Alpha8	10000	0.5008	0.9343	−0.1189	0.4998	1.1362
Treatment	10000	−1.2376	0.4010	−1.5063	−1.2399	−0.9641
Sex	10000	−2.6808	0.5655	−3.0405	−2.6625	−2.2928

Figure 16. Mean, Standard Deviation, and Quantiles of the Posterior Samples

Bayesian Analysis

Interval Statistics of the Posterior Samples

Parameter	Alpha	Credible Interval		HPD Interval	
Alpha1	0.050	−7.4898	−5.5620	−7.4192	−5.5177
Alpha2	0.050	−5.1057	−3.1993	−5.0080	−3.1390
Alpha3	0.050	−3.9992	−2.0368	−3.9341	−1.9814
Alpha4	0.050	−3.1862	−0.7362	−3.1873	−0.7411
Alpha5	0.050	−0.2158	2.5341	−0.0694	2.6606
Alpha6	0.050	−2.5030	0.7610	−2.4881	0.7730
Alpha7	0.050	−2.6620	0.7168	−2.6001	0.7613
Alpha8	0.050	−1.3323	2.3489	−1.3608	2.3042
Treatment	0.050	−2.0214	−0.4616	−1.9975	−0.4415
Sex	0.050	−3.8675	−1.6250	−3.8191	−1.5935

Figure 17. 95% Credible and Highest Probability Density Intervals (HPD) of the Posterior Samples

Convergence diagnostics are critical when you are making inference from Markov chain based simulations. If the Markoc chain hasn't converged, which means that it hasn't explored the target chain fully, then inferences based on the samples will be misleading.

You can request specific convergence diagnostics with the PHREG procedure by specifying the DIAG= option. By default, a table containing lag1, log5, lag10, and lag50 autocorrelations for each parameter is displayed (Figure 18), as well as the the Gelman-Rubin diagnostics (Figure 19), and the Geweke diagnostics, (Figure 20). The diagnostics do not detect any failure of convergence; sample autocorrelations diminish rapidly over time and the Geweke diagnostics, which test the convergence of the means, are all nonsignificant. The Gelman-Rubin diagnostics are all close to 1.

With ODS Graphics enabled, a panel consisting of the trace plot, the autocorrelation function plot and the kernel density plot is displayed for each parameter. The panel of diagnostic plots for the Treatment parameter is showed in Figure 21 and does not reveal any sign of nonconvergence. The trace plot is dense, indicating good mixing of the Markov Chain. The autocorrelation plot suggests that correlations among the posterior samples are not high. Finally, the density plot indicates that the posterior distribution for treatment has a single mode and a normal distribution shape.

Bayesian Analysis

Autocorrelations of the Posterior Samples

Parameter	Lag1	Lag5	Lag10	Lag50
Alpha1	0.0425	0.0095	0.0029	-0.0118
Alpha2	0.0814	0.0399	0.0295	-0.0018
Alpha3	0.1421	0.0706	0.0489	0.0193
Alpha4	0.4326	0.2077	0.1048	-0.0118
Alpha5	0.4146	0.2211	0.1150	-0.0060
Alpha6	0.5921	0.3114	0.1508	0.0043
Alpha7	0.6214	0.3409	0.1696	0.0023
Alpha8	0.6791	0.3750	0.1796	0.0043
Treatment	0.6422	0.1951	0.1078	-0.0106
Sex	0.8222	0.4201	0.1918	0.0079

Figure 18. Autocorrelation of the Posterior Samples

Gelman-Rubin Diagnostics

Parameter	Estimate	97.5% Bound
Alpha1	1.0001	1.0004
Alpha2	1.0001	1.0005
Alpha3	1.0007	1.0028
Alpha4	1.0009	1.0035
Alpha5	1.0007	1.0026
Alpha6	1.0012	1.0035
Alpha7	1.0017	1.0057
Alpha8	1.0016	1.0054
Treatment	1.0010	1.0034
Sex	1.0020	1.0065

Figure 19. Gelman-Rubin Diagnostics

Bayesian Analysis

Geweke Diagnostics

Parameter	z	Pr > |z|
Alpha1	0.5576	0.5771
Alpha2	-1.6153	0.1063
Alpha3	0.6377	0.5237
Alpha4	0.4928	0.6222
Alpha5	0.2568	0.7973
Alpha6	0.2332	0.8156
Alpha7	0.5362	0.5918
Alpha8	0.2475	0.8045
Treatment	-0.6964	0.4862
Sex	-0.1986	0.8426

Figure 20. Geweke Diagnostics of the Chain

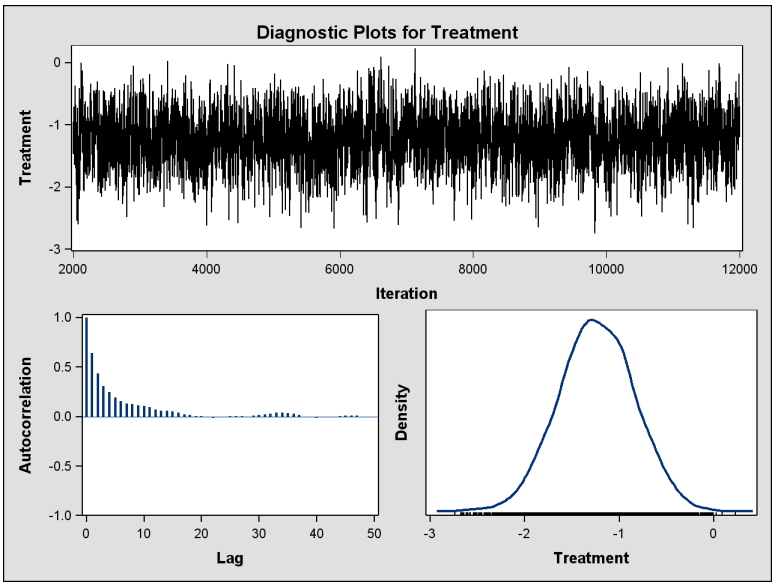

Figure 21. Trace, Autocorrelation Function, and Density Plots for Treatment)

BAYESIAN ANALYSIS FOR THE POISSON REGRESSION MODEL IN PROC GENMOD

Consider the following data on patients from clinical trials. The data are a subset of the data described in Ibrahim, J. G., Chen, Ming-Hui, and Lipsitz, S. R. (1999). Primary interest lies in how the number of cancerous liver nodes is predicted by six other baseline characteristics. The number of nodes is modeled by a Poisson regression model with the six baseline characteristics as covariates. PROC GENMOD performs an analysis to obtain Bayesian estimates of the regression coefficients.

The following DATA step creates the SAS data set liver:

```
data liver;
   input x1-x6 y;
   label
   y  = "Number of Cancerous Liver Nodes"
   x1 = "Body Mass Index"
   x2 = "Age in Years"
   x3 = "Time Since Diagnosis of Disease in Weeks"
   x4 = "Two Biochemical Markers (each classified as normal=1 or abnormal=0)"
   x5 = "Anti Hepatitis B Antigen"
   x6 = "Associated Jaundice (yes=1, no=0)";
datalines;
19.1358    50.0110     51.000      0        0       1        3
23.5970    18.4959      3.429      0        0       1        9
20.0474    56.7699      3.429      1        1       0        6
28.0277    59.7836      4.000      0        0       1        6
28.6851    74.1589      5.714      1        0       1        1
18.8092    31.0630      2.286      0        1       1       61
28.7201    52.9178     37.286      1        0       1        6
21.3669    61.6603     54.143      0        1       1        6
23.7332    42.2904      0.571      1        0       1       21
....
```

The following program requests a Poisson regression for response variable y, based on all six explanatory

variables. The standard maximum likelihood analysis will be performed. In addition, the BAYES statement requests a Bayesian analysis.

```
ods graphics on;
proc genmod data=liver;
   model y = x1-x6 / dist=poisson;
   bayes;
run;
ods graphics off;
```

Uniform noninformative prior distributions for the regression parameters are chosen for the analysis by default. Normal and Jeffrey's prior distributions for the regression parameters are also available.

Summary statistics from the posterior distribution samples are computed to identify the significant parameters and provide pointwise estimates. By default, 10,000 sample values are generated after a burn-in period of 2000 samples. The number of samples and burn-in period can be specified to be other values.

Figure 22 shows the maximum likelihood estimates for the model.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	
Intercept	1	2.4508	0.2284	2.0032	2.8984
x1	1	-0.0044	0.0080	-0.0201	0.0114
x2	1	-0.0135	0.0024	-0.0181	-0.0088
x3	1	-0.0029	0.0022	-0.0072	0.0014
x4	1	-0.2715	0.0795	-0.4272	-0.1157
x5	1	0.3215	0.0832	0.1585	0.4845
x6	1	0.2077	0.0827	0.0456	0.3698
Scale	0	1.0000	0.0000	1.0000	1.0000

NOTE: The scale parameter was held fixed.

Figure 22. ML Parameter Estimation

Next come the Bayesian results. Figure 23 and Figure 24 displays the summary statistics and interval statistics for the posterior samples, respectively. The results are similar to those based on maximum likelihood.

Bayesian Analysis

Descriptive Statistics of the Posterior Samples

Parameter	N	Mean	Standard Deviation	Quantiles 25%	50%	75%
Intercept	10000	2.4517	0.2318	2.2946	2.4522	2.6107
x1	10000	-0.00461	0.00823	-0.00989	-0.00445	0.00101
x2	10000	-0.0135	0.00239	-0.0150	-0.0135	-0.0119
x3	10000	-0.00291	0.00214	-0.00436	-0.00291	-0.00140
x4	10000	-0.2711	0.0803	-0.3246	-0.2705	-0.2179
x5	10000	0.3199	0.0826	0.2648	0.3210	0.3766
x6	10000	0.2075	0.0826	0.1505	0.2082	0.2623

Figure 23. Descriptive Statistics for Posterior Samples

Interval Statistics of the Posterior Samples

Parameter	Alpha	Credible Interval		HPD Interval	
Intercept	0.050	2.0029	2.8989	1.9987	2.8923
x1	0.050	−0.0216	0.0111	−0.0204	0.0120
x2	0.050	−0.0181	−0.00872	−0.0181	−0.00871
x3	0.050	−0.00713	0.00116	−0.00695	0.00129
x4	0.050	−0.4278	−0.1112	−0.4344	−0.1196
x5	0.050	0.1540	0.4777	0.1598	0.4822
x6	0.050	0.0464	0.3698	0.0402	0.3632

Figure 24. Interval Statistics of the Posterior Samples

Diagnostic results are displayed in Figure 25, Figure 26, and Figure 27.

Bayesian Analysis

Autocorrelations of the Posterior Samples

Parameter	Lag1	Lag5	Lag10	Lag50
Intercept	0.4913	0.1855	0.0818	0.0037
x1	0.8684	0.4846	0.2077	0.0171
x2	0.2053	0.0468	0.0296	0.0086
x3	0.8789	0.5141	0.2337	0.0059
x4	0.0772	−0.0071	0.0126	−0.0012
x5	0.1487	0.0078	−0.0097	−0.0199
x6	0.1567	0.0148	−0.0013	−0.0070

Figure 25. Autocorrelations of the Posterior Samples

Gelman-Rubin Diagnostics

Parameter	Estimate	97.5% Bound
Intercept	0.9999	1.0000
x1	1.0004	1.0010
x2	1.0003	1.0011
x3	1.0036	1.0115
x4	1.0000	1.0001
x5	1.0001	1.0003
x6	1.0002	1.0007

Figure 26. Gelman-Rubin Diagnostics

Geweke Diagnostics

Parameter	z	Pr > |z|
Intercept	-0.2259	0.8212
x1	-0.7153	0.4745
x2	2.3670	0.0179
x3	-1.5812	0.1138
x4	0.1587	0.8739
x5	1.2745	0.2025
x6	2.2139	0.0268

Figure 27. Geweke Diagnostics

None of the diagnostics indicate any failure of convergence.

Trace, autocorrelation, and kernel density plots from the posterior samples are displayed in Figure 28 for variable Body Mass Index.

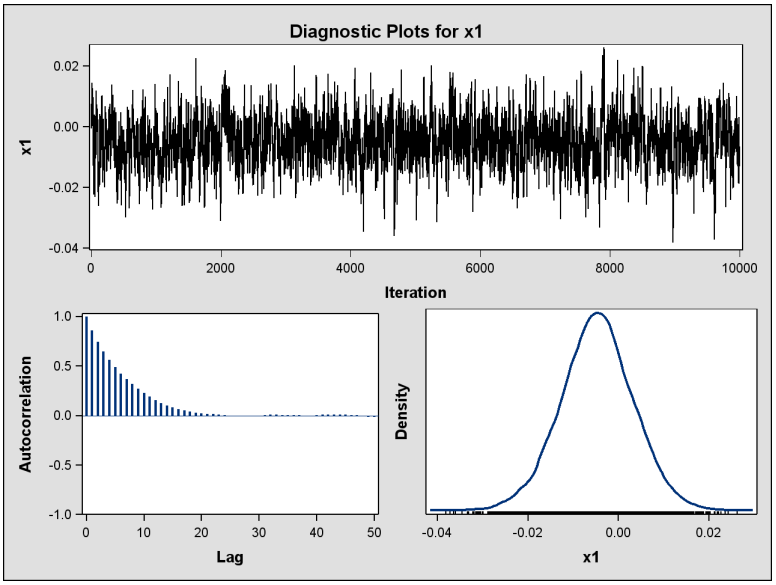

Figure 28. Bayes Diagnostic Plots for Body Mass Index

This plot reveals no problems.

THINGS GO BETTER WITH PROC TTEST

While you might think that a procedure included in the first release of SAS software should be finished by now, that's not the case. One of the most improved procedures in SAS 9.2 is the TTEST procedure. SAS 9.2 enhances PROC TTEST in several ways, most notably by including ODS graphics to supplement the numerical results with visual displays. Univariate and comparative versions of histograms, box plots, and densities are automatically generated, and graphics are customizable as well. The new SIDES= option enables one-sided versions of tests and confidence intervals, useful for achieving higher statistical power when the directionality can be assumed in advance.

In addition, more analyses are now available with the TTEST procedure. The new TOST= option produces equivalence tests and intervals, popular in such situations as bioequivalence, where the goal is to establish whether the effects of two drugs are significantly *similar* rather than different. Standard p-values from *t* tests are not appropriate for assessing similarity since insignificant results may stem from low power rather than

the lack of a meaningful effect. PROC TTEST also supports the analysis of crossover designs (AB/BA) with the CROSSOVER= option. These designs, the favorite in many different fields of practice, reduce variability and increase the power of treatment comparisons by administering both of two treatments to each subject.

The following example illustrates the analysis of a simple crossover design study with the TTEST procedure.

AB/BA CROSSOVER DESIGNS

Crossover trials are experiments in which each subject is given a sequence of different treatments. They are especially common in clinical trials. The reduction in variability from taking multiple measurements on a subject allows for more precise treatment comparisons. The simplest such design is the AB/BA crossover, in which each subject receives each of two treatments in a randomized order.

Senn (1993, Chapter 3) discusses a study comparing the effectiveness of two bronchodilators, formoterol (F) and salbutamol (S), in the treatment of childhood asthma. A total of 13 children are recruited for an AB/BA crossover design. A random sample of seven of the children are assigned to the sequence F/S, receiving a dose of formoterol upon an initial visit (Period 1) and then a dose of salbutamol upon a later visit (Period 2). The other six children are assigned to the sequence S/F, receiving the treatments in the reverse order. Periods 1 and 2 are sufficiently spaced so that no carryover effects are anticipated. After a child inhales a dose of a bronchodilator, peak expiratory flow (PEF) is measured. Higher PEF indicates greater effectiveness.

The SAS data set asthma is generated with the following statements:

```
data asthma;
   input Drug1 $ Drug2 $ PEF1 PEF2 @@;
   datalines;
F S 310 270   F S 310 260  F S 370 300
F S 410 390   F S 250 210  F S 380 350
F S 330 365
S F 370 385   S F 310 400  S F 380 410
S F 290 320   S F 260 340  S F 90  220
;
run;
```

Given the assumption of no carryover effect, three primary tests of interest emerge:

- test of treatment effect assuming negligible period effect
- test of treatment effect allowing for possible period effect
- test of period effect

All of these tests reduce to two-sample or paired *t* tests on simple transformations of the data (hence the choice of PROC TTEST as a new home for these analyses). The data is split into two groups according to the two sequences; a two-sample *t* test on halved period differences between the sequences assesses the treatment effect, and a two-sample *t* test on halved crossover differences (defined as period differences but with those for the second sequence negated) between sequences assesses the period effect. If the period effect is assumed to be negligible, then a paired *t* test comparing the A and B measurements (regardless of sequence) for each subject assesses the treatment effect.

Although it is possible to analyze an AB/BA crossover design by using earlier versions of PROC TTEST after computing the appropriate transformations in the DATA step—or by using linear models procedures such as MIXED, GLM, or ANOVA—explicit coverage of this design in PROC TTEST allows for more specialized output and graphs that have been developed specifically for crossover designs.

You can perform an analysis of the treatment and period effects (assuming a model including terms for both effects) by using the new CROSSOVER option in the VAR statement. By default, histograms, box plots, and QQ plots are produced; a subject profile plot is produced when the CROSSOVER= option is specified.

```
ods graphics on;
proc ttest data=asthma;
   var PEF1 PEF2 / crossover= (Drug1 Drug2);
run;
ods graphics off;
```

The variables PEF1 and PEF2 represent the responses in Period 1 and Period 2. The variables Drug1 and Drug2 similarly represent the treatments applied in each period.

Figure 29 displays a panel of histograms for the responses at Period 1 and Period 2 for each sequence of treatments, and Figure 30 displays subject profiles. These profiles suggest that formoterol is more effective that salbutamol.

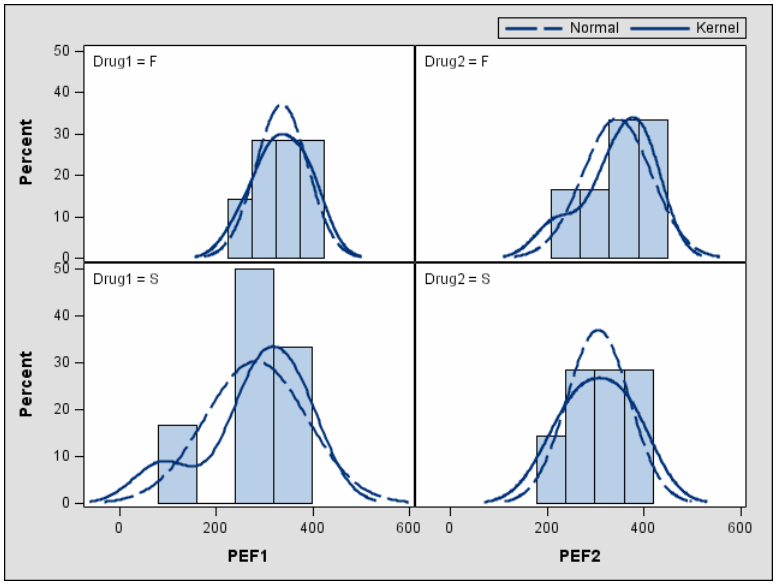

Figure 29. Histograms from the TTEST Procedure

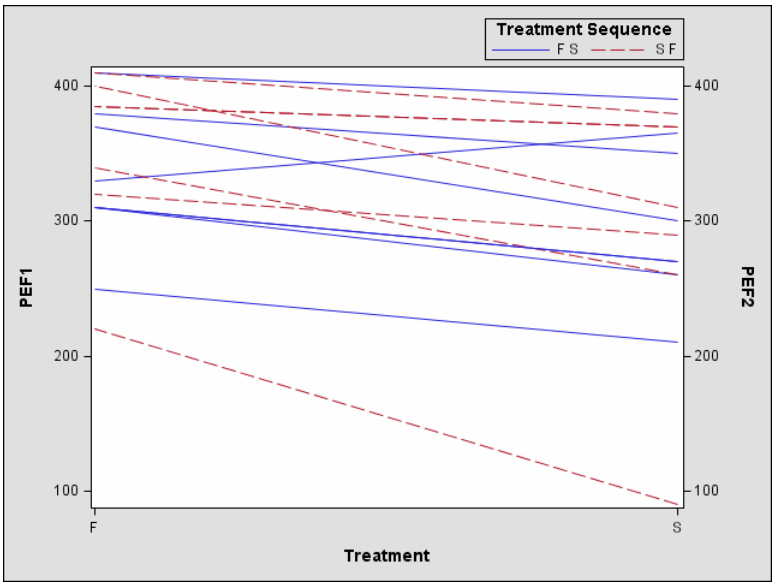

Figure 30. Subject Profiles from the TTEST Procedure

Figure 31 and Figure 32 display the results from the crossover analysis.

The TTEST Procedure

Response Variables: PEF1 PEF2

Crossover Variable Information

Period	Response	Treatment
1	PEF1	Drug1
2	PEF2	Drug2

Treatment	Period	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	1	7	337.1	53.7631	20.3206	250.0	410.0
F	2	6	345.8	70.8814	28.9372	220.0	410.0
S	1	6	283.3	105.4	43.0245	90.0000	380.0
S	2	7	306.4	64.7247	24.4636	210.0	390.0
Diff (F-S)			46.6071	19.3702	10.7766		
	Diff (1-2)		-15.8929	19.3702	10.7766		

Treatment	Period	Method	Mean	95% CL Mean		Std Dev
F	1		337.1	287.4	386.9	53.7631
F	2		345.8	271.4	420.2	70.8814
S	1		283.3	172.7	393.9	105.4
S	2		306.4	246.6	366.3	64.7247
Diff (F-S)		Pooled	46.6071	22.8881	70.3262	19.3702
Diff (F-S)		Satterthwaite	46.6071	21.6585	71.5558	
	Diff (1-2)	Pooled	-15.8929	-39.6119	7.8262	19.3702
	Diff (1-2)	Satterthwaite	-15.8929	-40.8415	9.0558	

Figure 31. Crossover Output

Treatment	Period	Method	95% CL	Std Dev
F	1		34.6446	118.4
F	2		44.2447	173.8
S	1		65.7841	258.5
S	2		41.7082	142.5
Diff (F-S)		Pooled	13.7217	32.8882
Diff (F-S)		Satterthwaite		
	Diff (1-2)	Pooled	13.7217	32.8882
	Diff (1-2)	Satterthwaite		

Treatment	Period	Method	Variances	DF	t Value	Pr > |t|
Diff (F-S)		Pooled	Equal	11	4.32	0.0012
Diff (F-S)		Satterthwaite	Unequal	9.1017	4.22	0.0022
	Diff (1-2)	Pooled	Equal	11	-1.47	0.1683
	Diff (1-2)	Satterthwaite	Unequal	9.1017	-1.44	0.1838

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	5	6	1.84	0.4797

Figure 32. Crossover Output

The estimated mean treatment difference is 46.6071 with 95% confidence interval [22.8881, 70.3262]. The treatment effect is highly significant with p=0.0012 using the usual *t* test and p=0.0022 using the Satterthwaite *t* test. The period effect is not significant at the α=0.05 level, with p=0.1683 using the usual *t* test and p=0.1838 using the Satterthwaite *t* test.

The question of equal variances (which determines the choice of usual vs. Satterthwaite *t* test and characterizes the Equality of Variances test) is the same for both the treatment and period effect tests, as both of these tests are two-sample *t* tests using the treatment sequence as the group.

To ignore the period effect (assume it is zero) and gain one more degree of freedom in testing the treatment difference, use the IGNOREPERIOD option in the VAR statement:

```
proc ttest data=asthma;
   var PEF1 PEF2 / crossover= (Drug1 Drug2) ignoreperiod;
run;
```

The tests of treatment difference ignoring period (not shown here) are similar to those in Figure 31. The p-values for the treatment effect becomes p=0.0017.

REGRESSION DIAGNOSTICS FOR GEE MODELS IN THE GENMOD PROCEDURE

One of the major enhancements to the MIXED procedure in SAS 9.1 was the introduction of regression diagnostics and influence statistics. SAS 9.2 brings regression diagnostics to generalized estimating equations(GEE) analysis in the GENMOD procedure. One of these measures is an extended version of the Cook's distance statistic (Preisser and Qaqish 1999). The following example illustrates the use of these statistics.

Consider the following subset of data from a controlled randomized trial directed at assessing the impact of urinary incontinence guideline adoption by primary case providers on patient outcomes. The data were originally discussed in Preisser and Qaqish (1999), and they have also been analyzed by Fay and Graubard (2001). The binary response is bothered, which measures whether a patient is bothered by urinary incontinence. Researchers recorded values for a number of explantory variables. Patients are clustered according

to physician practice, so that a repeated measures model fit by GEEs) is appropriate. The GEE version of the Cook's distance statistic is computed to identify physician practices that are influential on the overall fit. Influential individual observations (patients) are also identified by observation-level Cook's statistics for GEEs.

```
data preqaq99;
   label pat_id="Patient ID";
   input pract_id pat_id bothered female ageyrs age dayacc severe toilet;
   datalines;
8        1      1        1      77     0.1  7      3      8
8        2      0        1      82     0.6  1      1      3
8        3      1        1      78     0.2  7      3      6
24       4      0        1      87     1.1  0.286  2      6
24       5      0        1      78     0.2  2      2      4
27       6      0        1      79     0.3  1      2      4
182      89     0        0      85     0.9  5.571  2      6
182      90     1        0      77     0.1  4      2      6
182      91     0        1      78     0.2  2      2      6
182      92     0        1      85     0.9  0.143  1      5
182      93     1        1      76     0    4      2      7
182      94     0        1      83     0.7  0.286  2      4
...
232      132    1        0      87     1.1  7      3      7
232      133    1        0      79     0.3  4      2      5
235      134    0        1      85     0.9  0.143  3      3
235      135    1        1      84     0.8  5      1      3
235      136    0        1      76     0    0.143  2      10
235      137    1        1      76     0    5      2      6
;
run;
```

The following program fits the Poisson regression model, and the REPEATED statement requests a GEE analysis. Specifying PLOTS=(COOKSD and CLUSTERCOOKSD) requests plots of Cook's distance versus cluster and Cook's distance versus observation number.

```
ods graphics on;
proc genmod data = preqaq99 descending plots=(cooksd clustercooksd);
   class pract_id ;
   model bothered = female age dayacc severe toilet/d=bin itprint ;
   repeated sub=pract_id/corr=exch modelse;
run;
ods graphics off;
```

Plots of the Cook's statistics help to identify Practice 12 as the most influential cluster and Patient 44 as the most influential observation.

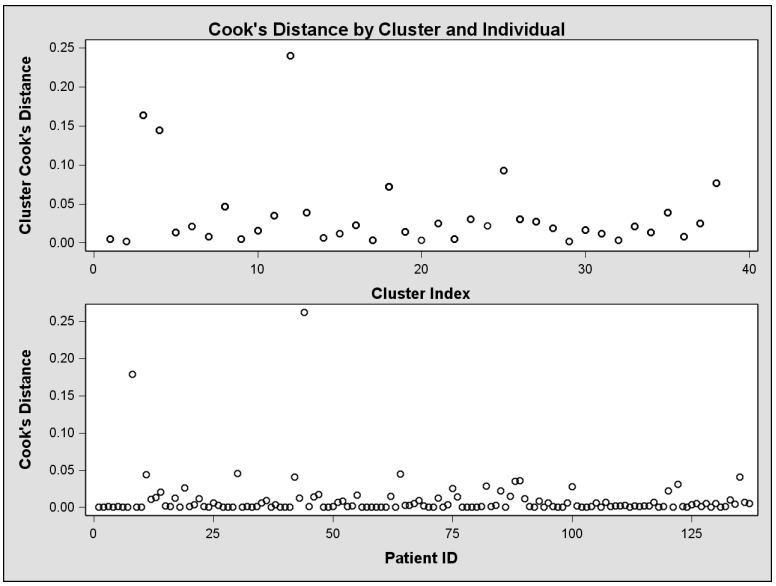

Figure 33. Plot of Cook's Distance by Cluster and Individual Produced by PROC GENMOD

POWER AND SAMPLE SIZE APPLICATION

The SAS SAS Power and Sample Size Application has been rewritten as a desktop client that doesn't require a web browser or a web server. You can now use either SAS/IntegrationTechnologies or SAS/CONNECT to connect to a remote SAS server, and of course, you can still use it locally on your desktop or laptop. You can continue to use your existing projects that you created with PSS 2.0.

New analyses include a Wilcoxon-Mann-Whitney test for two distributions, a confidence interval for a proportion, and logistic regression for a binary response. You can now request fractional sample sizes as input or output for all analyses. Alternate input parameterizations have been added to the Two Correlated and Two Independent Proportions analyses.

Figure 34 displays the task for the Wilcoxon test in the new client.

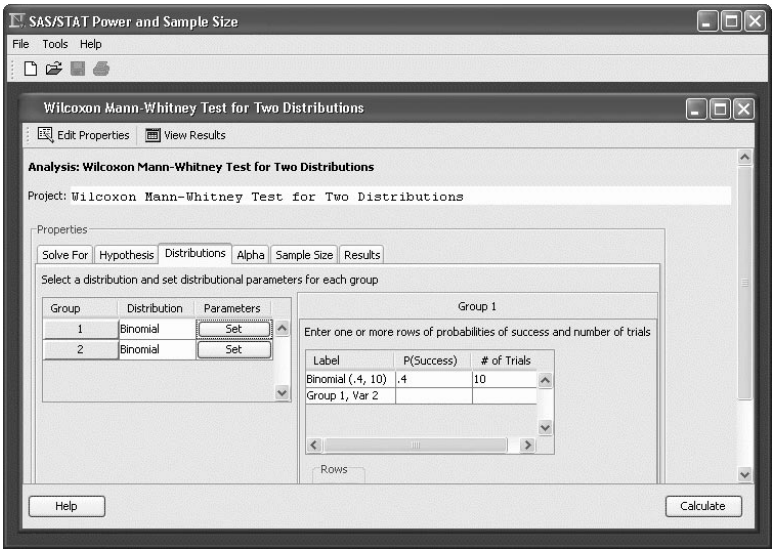

Figure 34. Wilcoxon Task

Figure 35 displays Wilcoxon results, produced by the POWER procedure.

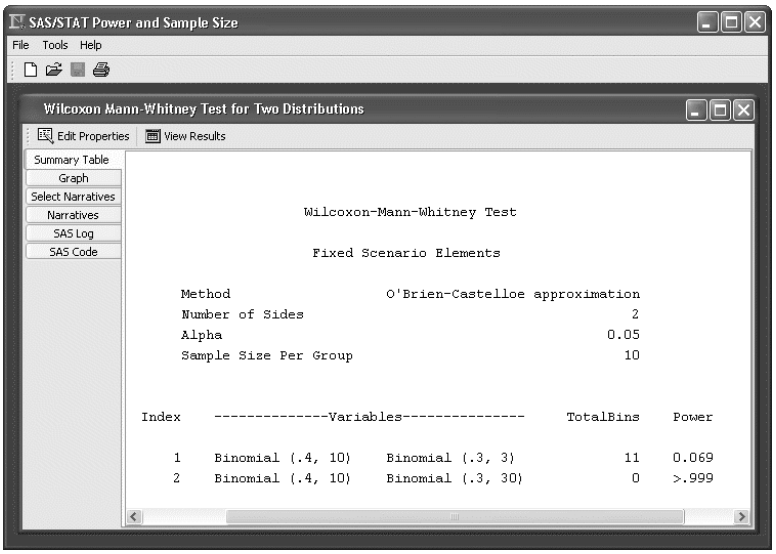

Figure 35. Wilcoxon Results

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 SAS 9.2, over 200 updates have been made, including many new graphics, as mentioned above. Some of the other enhancements include:

- new methods for binomial confidence intervals in PROC FREQ
- Zelen's test for stratified/multiway tables in PROC FREQ
- AIC and QIC statistics added to PROC GENMOD results
- METHOD=LAPLACE added as an estimation option in PROC MIXED
- Conover test added to PROC NPAR1WAY
- Hodges-Lehmann confidence interval for differences in medians in PROC NPAR1WAY
- Firth's method for handling monotone likelihoods added to PROC PHREG
- heteroscedasticity consistent (White) standard errors now produced by PROC REG
- DOMAIN statement added to the SURVEYLOGISTIC and SURVEYMEANS procedure
- JACKKNIFE and BRR variance estimation methods added to the survey analysis procedures
- allocation capabilities added to the SURVEYSELECT procedure

In addition, many features introduced in SAS 9.1 as experimental features become production in SAS 9.2, including the following:

- CLASS statement is included in PROC PHREG
- QUANTREG procedure
- GLMSELECT procedure
- RESIDUAL and INFLUENCE options in PROC MIXED
- cumulative residuals in PROC GENMOD and PROC PHREG
- SIMNORM procedure

UPCOMING EXPERIMENTAL PROCEDURES

SAS 9.2 includes several exciting new experimental procedures still under development at the time of this writing. Stay tuned to www.sas.com/statistics/ for up to date information about SAS/STAT software in the SAS 9.2 release.

CONCLUSION

Statisticians of all shapes and sizes should find something to suit their needs in the latest edition of SAS statistical software. Statistical modelers can now perform generalized linear mixed models and quantile regression, as well as conduct model selection with the GLMSELECT procedure. Bayesians can use SAS for survival analyses. High-end data analysts can look for trends in their data with the new Stat Studio. Survey statisticians can now choose from several variance estimation techniques. And everyone will benefit from the statistical graphics now available for nearly every analysis. And again, please look for an update of this paper on www.sas.com/statistics when the SAS 9.2 release comes out.

REFERENCES

Berry, D. A. (1996). *Statistics: A Bayesian Perspective*, London: Duxbury Press.

Box, G. E. P. and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis*, New York: John Wiley & Sons.

Castelloe, J. and Tobias, R. (2006). "Like Wine, the TTEST Procedure Improves with Age," *Proceedings of the Thirty-first Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc.

Chen L., (2005). "An Introduction to Quantile Regression and the QUANTREG Procedure," *Proceedings of the Thirtieth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc.

Cohen, R. (2006). "Introducing the GLMSELECT Procedure for Model Selection," *Proceedings of the Thirty-First Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). "Least Angle Regression (with discussion)," *Annals of Statistics*, 32, 407–499.

Fay M. P., Graubard, B. I. (2001). "Small-sample adjustments for Wald-type tests using sandwich estimators," *Biometrics*, 57, 1198–1206.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis, 3rd Edition*, London: Chapman & Hall.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning,* New York: Springer-Verlag, Inc.

Ibrahim, J. G., Chen, Ming-Hui, and Lipsitz, S.R. (1999). "Monte Carlo EM for Missing Covariates in Parametric Regression Models," *Biometrics*, 55, 591–596.

Ibrahim, J. G., Chen, M., and Sinha, D. (2001). *Bayesian Survival Analysis*, New York: Springer.

Lee, P. M. (2004). *Bayesian Statistics: An Introduction, 3rd Edition*, London: Arnold.

Preisser JS, Qaqish BF (1999), "Robust Regression for Clustered Data with Application to Binary Responses", *Biometrics*, 55, 574–579.

Rodriguez, R. and Balan, T. (2006). "Creating Statistical Graphics in SAS 9.2: What Every Statistical User Should Know," *Proceedings of the Thirty-first Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc.

Schabenberger, O. (2005). "Introducing the GLIMMIX procedure for Generalized Linear Models," *Proceedings of the Thirtieth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc.

Senn, S. (1993). *Cross-over Trials in Clinical Research*, New York: John Wiley & Sons, Inc.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, New York: John Wiley & Sons.

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.

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society Series B*, 58, 267–288.

Wolfinger, R. and O'Connell, M. (1993). "Generalized Linear Mixed Models: A Pseudo-Likelihood Approach," *Journal of Statistical Computation and Simulation*, 4, 233–243.

ACKNOWLEDGEMENTS

We are grateful to Tim Arnold, Tonya Balan, John Castelloe, Fang Chen, Lin Chen, Virginia Clark, Robert Cohen, Bob Derr, Gordon Johnson, Oliver Schabenberger, Ying So, Wayne Watson, and Rick Wicklin for valuable assistance in the preparation of this paper.

CONTACT INFORMATION

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

Robert N. Rodriguez SAS Institute Inc. SAS Campus Drive Cary, NC 27513