The Next Generation: SAS/STAT® 9.22

Maura Stokes, Bob Rodriguez, Robert Cohen
SAS Institute, Cary NC

August 13, 2010

Abstract

Caught up with SAS/STAT® 9.2? Want to add even more statistical tools to your arsenal? Then get ready for the 9.22 release of SAS/STAT software.

New functionality for postprocessing equips many linear modeling procedures with comparable capabilities, including the latest techniques for testing complex research hypotheses. The new SURVEYPHREG procedure fits Cox regression models to survey data. The EFFECTPLOT statement uses ODS Graphics to create plots of effects from models produced by the GENMOD, LOGISTIC, and ORTHOREG procedures. Exact Poisson regression is one of several new exact methods for categorical data analysis. Other new features include updated spatial analysis capabilities, classification variable support in the ROBUSTREG procedure, and model averaging in the GLMSELECT procedure. This talk provides an overview of these exciting new enhancements to SAS/STAT.

Introduction

SAS/STAT 9.2 brought major enhancements to the product: Bayesian capabilities in the GENMOD, LIFEREG, and PHREG procedures; the MCMC procedure for fitting general Bayesian models; the experimental HPMIXED procedure for performing mixed model analysis with a large number of observations and/or class levels; the SEQDESIGN and SEQTEST procedures for group sequential analysis; and production ODS Statistical Graphics. In addition, more than 200 new features were added to the existing software. The MCMC procedure and the SEQDESIGN and SEQTEST procedures became production with the second maintenance release of SAS 9.2.

SAS/STAT 9.22 is a brand-new release which is coupled with the third maintenance release of SAS 9.2. SAS/STAT 9.22 also brings major enhancements to the SAS/STAT product. New internal architecture means a full complement of postfitting capabilities in a large number of linear modeling procedures. The new PLM procedure enables you take model information stored from these procedures and use it to perform additional inference and scoring without refitting the original model. This is especially useful in situations where fitting complex models requires considerable computation time.

The new SURVEYPHREG procedure provides survival analysis, in the form of Cox proportional hazards regression, for sample survey data. This, along with the expansion of replication methods to all analyses provided by the survey analysis procedures, means that SAS/STAT software now provides state-of-the-art analysis for a broad range of sample survey data. In addition, graphics are now available in the SURVEYFREQ procedure.

There are numerous other updates. You can perform exact Poisson regression with the GENMOD procedure as well as fit zero-inflated negative binomial models. You can perform seamless analysis and prediction of spatial processes with the spatial analysis procedures. Model averaging for high-dimension prediction is now available with the GLMSELECT procedure. The more powerful and customizable structural equation modeling first implemented with the experimental TCALIS procedure in SAS/STAT 9.2 has been rolled into the CALIS procedure. And more procedures now include the EFFECT statement, which defines a richer family of linear models.

Obviously, one paper can't describe all of the important enhancements in this release. The goal here is to provide a solid overview of the updates, illustrate some of the exciting new capabilities, and point out additional online references, including relevant papers in the proceedings of this conference.

Postfitting Inference in the Post-9.2 World

One of the strengths of SAS/STAT linear modeling procedures is the breadth of postfitting analyses available once you have fit your model and estimated its parameters. Inference is based on estimating and testing linear combinations

of parameters, $\mathbf{L}\beta$, for some matrix $\mathbf{L}$. Such statements as CONTRAST, ESTIMATE, LSMEANS, and LSMESTIMATE provide a way of requesting such linear inference, specifying the values of $\mathbf{L}$ and requesting specific tests. With SAS/STAT 9.22, you have access to even more postfitting analyses.

A Plethora of Postfitting Potential

Ever want an LSMEANS statement in PROC LOGISTIC? Do you wonder why PROC MIXED doesn't provide an LS-MESTIMATE statement like PROC GLIMMIX? Well, wonder no more! Over 30 postfitting statements have been added to the linear modeling procedures in SAS/STAT software. In addition, several existing statements have been updated. Table 1 provides an overview of the 9.22 SAS/STAT postfitting capabilities for linear modeling procedure. Checks indicate that new statements have been added to procedures, stars indicate existing functionality, and starred checks indicate that existing statements have been updated.

Table 1 Postfitting Statements Available in Linear Modeling Procedures

Procedure	CONTRAST	EFFECTPLOT	ESTIMATE	LSMEANS	LSMESTIMATE	SLICE	TEST
GENMOD	⋆	✓	⋆	⋆✓	✓	✓	
GLM	⋆		⋆	⋆			⋆
GLIMMIX	⋆		⋆	⋆	⋆✓	✓	
LOGISTIC	⋆	✓	✓	✓	✓	✓	⋆
MIXED	⋆		⋆	⋆	✓	✓	
ORTHOREG		✓	✓	✓	✓	✓	✓
PHREG	⋆		✓	✓	✓	✓	⋆
SURVEYLOGISTIC	⋆		✓	✓	✓	✓	⋆
SURVEYPHREG			✓	✓	✓	✓	✓
SURVEYREG	⋆		⋆	✓	✓	✓	✓

For example, the GENMOD procedure now offers the EFFECTPLOT, LSMESTIMATE, and SLICE statements, and its LSMEANS statement has been updated with additional features. The MIXED procedure picks up the LSMESTIMATE statement and the SLICE statement, and the PHREG procedure picks up the ESTIMATE, LSMEANS, LSMESTIMATE, and SLICE statements.

The SLICE statement is used to perform inferences on model effects that are made up of classification variables only. Effects listed in the SLICE statement must be high-order effects of at least two classification variables. The effect is then partitioned into subsets that correspond the variables that are used to form the effect. This type of analysis is also known as an analysis of simple effects (Winer 1971). Options used in the LSMEANS statement can be used in the SLICE statement. For example, the statements

```
class a b;
model y= a b a*b;
slice a*b / diff sliceby=a;
```

produce tests of the differences among the A levels for each level of B. If you left out the SLICEBY= option, these statements would produce the differences among the B levels for each level of A as well.

One procedure that has benefited greatly from the avalanche of new postfitting statements is the ORTHOREG procedure, which provides very precise regression estimation. Dressed up with a CLASS statement in a recent release, PROC ORTHOREG has now been outfitted with a full array of postfitting statements. It now provides a richer set of analyses for data sets that require numerically precise techniques.

To explore the new possibilities of PROC ORTHOREG, consider an example taken from Introducing PROC PLM and Postfitting Analysis for Very General Linear Models in SAS/STAT® 9.22 by Tobias and Cai (2010).

The following code displays the statements now available in the ORTHOREG procedure. The FLOWERS data set contains information about the effect of different amounts of fertilizer on the size of two flower species. You want to determine the influence of fertilizer on the size, and you decide to fit trends for the two species and to model those trends as B-splines. The new EFFECT statement in PROC ORTHOREG creates spline effects for each level of Fertilizer, and the new TEST statement tests the following effects in the model: overall species difference, overall fertilizer trend, and their interaction. Note that the ODS GRAPHICS ON statement is submitted with this block of statements and remains in effect for the rest of the examples in this paper.

```
ods graphics on;
proc orthoreg data=Flowers;
   class Species;
```

```
    effect SmoothF = spline(Fertilizer);
    model Size = Species|SmoothF;
    test Species|SmoothF;
    effectplot;
    title 'Flower Species Fertilizer Effects';
run;
```
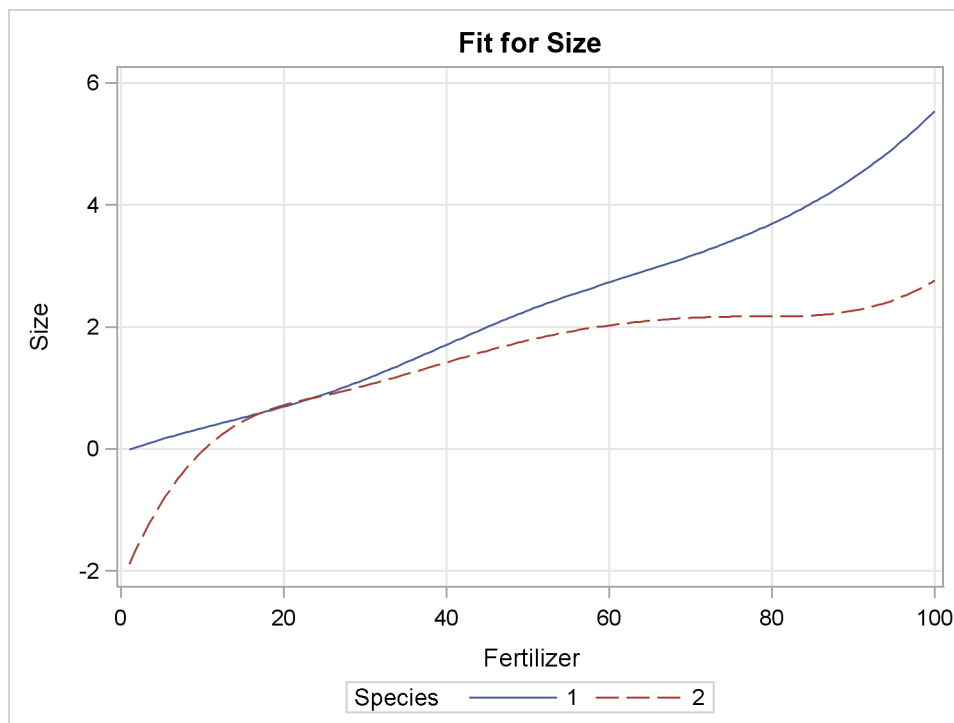
Figure 1 displays the TEST statements results: all effects are significant.

Figure 1 TEST Results

```
                    Type III Tests of Model Effects

                            Num      Den
            Effect           DF       DF     F Value    Pr > F

            Species           1       86       13.30    0.0005
            SmoothF           6       86      290.42    <.0001
            SmoothF*Species   6       86       30.47    <.0001
```

The plot produced by the EFFECTPLOT statement is shown in Figure 2.

Figure 2 Fit for Size

The graph displays the predicted smooth effect of fertilizer on size for each species; it helps you understand the nature of the interaction between fertilizer and species. In particular, larger amounts of fertilizer impact size for Species 1 much more than size for Species 2.

Internal architectural changes made these additional statements possible; in fact, with the exception of the CONTRAST statements, the postfitting statements are now "shared" by many procedures. Not only did these changes make it convenient to outfit the ORTHOREG procedure with the new statements illustrated above, but they also made it easier to produce a new procedure, the experimental SURVEYPHREG procedure, with a full complement of postfitting analyses.

But what if you want to run an additional postfitting analysis and you don't want to refit the model?

Running with the New PLM Procedure

The new PLM procedure performs postfitting inference with model fit information saved from a number of SAS/STAT modeling procedures. These procedures are equipped with the new STORE statement, which saves model information

as a SAS *item store*. An item store is a special SAS binary file that is used to store and restore information that has a hierarchical structure. Ten SAS/STAT procedures now provide the STORE statement: GENMOD, GLIMMIX, GLM, LOGISTIC, MIXED, ORTHOREG, PHREG, SURVEYLOGISTIC, SURVEYPHREG, and SURVEYREG.

The PLM procedure takes these item stores as input and performs tasks such as testing hypotheses, producing effect plots, and scoring a new data set. These tasks are specified through the usual complement of postfitting statements such as the TEST, LSMEANS, and new EFFECTPLOT and SCORE statements. Any procedure that offers the STORE statement can produce the item stores that are necessary for postfitting processing with the PLM procedure. This allows you to perform additional postfitting inference at a later time without having to refit your model, which is especially convenient for those models that are computationally expensive. In addition, with growing concerns for data confidentiality, storing and using intermediate results for remaining analyses might become a requirement in some organizations.

The PLM procedure also provides capabilities that are not necessarily available with the individual procedures. For example, you can use the EFFECTPLOT statement with item stores from the GLM procedure, which doesn't otherwise offer a EFFECTPLOT statement. And, the SCORE statement in PROC PLM enables you to score new data with the model saved from any procedure that has a STORE statement.

The most advanced postfitting inference techniques available in SAS/STAT are available in the PLM procedure; this includes step-down multiplicity adjustment for *p*-values, *F* tests with order restrictions, analysis of means (ANOM), and sampling-based linear inference based on Bayes posterior estimates.

If you include classification effects in your model, then that information is saved in the item store. If you use the EFFECT statement to construct effects in your model, and then use the STORE statement, those constructed effects are also available to you through the PLM procedure. The original model context is always preserved when that model information is saved with the STORE statement. For example, if you fit a model in the MIXED procedure that used the Satterthwaite or the Kenward-Roger degree-of-freedom method, that model context is available subsequently with analyses performed byPROC PLM.

Note that the SAS item stores created with the STORE statement are not currently transportable to different platforms.

STORE, STORE, STORE Your Beta

For a simple example, consider the GLM procedure. ALthough the postfitting inference capabilities of PROC GLM haven't changed with SAS/STAT 9.22, its new STORE statement means that you can apply the postfitting capabilities of PROC PLM to the model results you produced with the GLM procedure.

The following statements input the tried and true SAS Class data and fit an analysis of covariance. Model information– the parameter estimates and their covariances–are saved to a SAS item store named StudyResults in the Sasuser directory via the STORE statement.

```
proc glm data=sashelp.Class;
    class Sex;
    model Weight = Height Sex|Age;
    store sasuser.StudyResults;
run;
```

If you want to perform a test to compare the weights of boys and girls, adjusting for age and height, you may decide to test the difference of their least squares means. Since you created the StudyResults item store, you can apply PROC PLM to produce this analysis.

```
proc plm restore=sasuser.StudyResults;
    lsmeans Sex/ diff;
run;
```

Figure 3 displays the tabular results, which indicate no difference in weight after age and height adjustments.

Figure 3 Differences of Least Squares Means

			Differences of Sex Least Squares Means			
Sex	_Sex	Estimate	Standard Error	DF	t Value	Pr > |t|
F	M	−8.8303	5.9687	14	−1.48	0.1612

Figure 4 shows this graphically with the diffogram requested with the DIFF option in the LSMEANS statement.

Figure 4 Diffogram
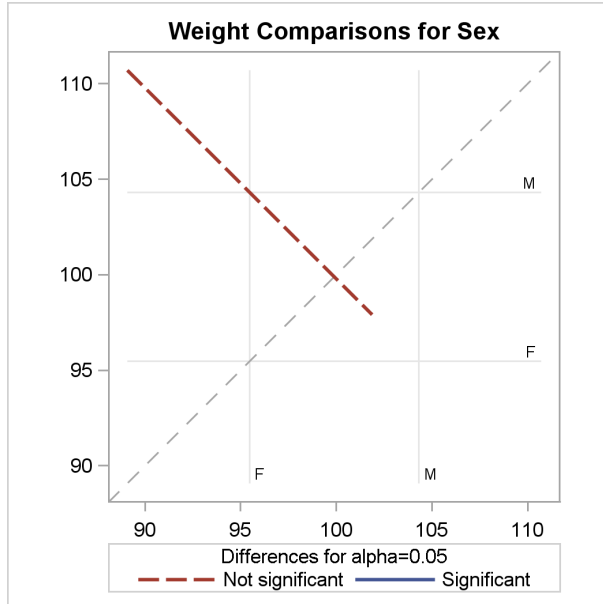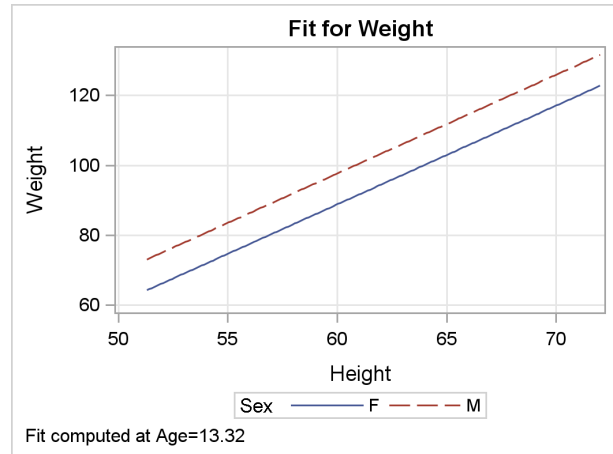

Weight Comparisons for Sex

Figure 5 Effect Plot

You later decide to produce a plot that displays the results of your analysis of covariance graphically. The GLM procedure doesn't offer the EFFECTPLOT statement, but you can employ the PLM procedure again for this purpose.

```
proc plm restore=sasuser.StudyResults;
   effectplot slicefit;
run;
```

The SLICEFIT option requests a curve of the predicted values versus a continuous variable grouped by the levels of the CLASS variable. Figure 5 displays this plot. The prediction lines are parallel with greater weights for males than females at the same height. The fit is computed at the overall mean age of 13.32.

For comprehensive information about the PLM procedure and new postfitting capabilities in SAS/STAT, refer to Introducing PROC PLM and Postfitting Analysis for Very General Linear Models in SAS/STAT® 9.22 by Tobias and Cai (2010).

The SURVEYPHREG Procedure

The experimental SURVEYPHREG procedure fits the semiparametric Cox model for proportional odds to sample survey data. Based on work by David Binder (1983), the procedure provides design-based variance estimates, confidence intervals, and hypothesis tests concerning the model parameters and model effects.

The SURVEYPHREG procedure produces a sample-based estimate $\hat{\beta}$ of finite-population proportional hazards regression parameters β_N by maximizing the partial pseudo-log-likelihood $l_\pi(\beta; \mathbf{Z}_i, t_i)$ based on observed covariates $\mathbf{Z}_i$ and observed survival time t_i. The procedure also produces an estimate of the sampling variance $V(\hat{\beta}|\mathcal{F}_N)$, which assumes the values of the finite population $\mathcal{F}_N$ are fixed.

For statistical inference, PROC SURVEYPHREG incorporates complex survey sample designs, including designs with stratification, clustering, and unequal weighting. PROC SURVEYREG provides both Taylor series and replication variance estimation procedures, and it also provides domain analysis through the DOMAIN statement. In addition, PROC SURVEYREG offers postfitting inference as performed with the ESTIMATE, LSMEANS, LSMESTIMATE, and SLICE statements. You can save model information with the STORE statement for further use with the new PLM procedure.

The paper titled Not Hazardous to Your Health: Proportional Hazards Modeling for Survey Data by Mukhopadhyay (2010) provides more information on this new procedure.

Consider the following example as an illustration of the use of PROC SURVEYPHREG.

A county librarian wants to study the length of time that borrowers actually keep their three-day books. She wants to take age and sex into consideration as well as the fact that some books never return. Assume that a list of 11,617 transactions for a three-month period is available. Stratified random sampling is used to select 100 transactions from

10 branch libraries. For each selection record, researchers conducted telephone interviews to obtain demographic information.

The following DATA step displays some of the records that were collected. The variable SamplingWeight contains the sampling weight for each observation.

```
data LibrarySurvey;
   input Branch         2.
         SamplingWeight 7.2
         CheckOut       date10.
         CheckIn        date10.
         Age Sex $;
   datalines;
 1 103.60 08NOV2008 13NOV2008 18 m
 1 103.60 01OCT2008 07OCT2008 30 m
 1 103.60 05NOV2008 06NOV2008 73 f
 1 103.60 25OCT2008 26OCT2008 53 m
 1 103.60 09NOV2008 10NOV2008 55 m
 2 127.50 10DEC2008 15DEC2008 39 m
 2 127.50 19DEC2008          . 33 f
 2 127.50 26NOV2008 27NOV2008 41 m
 2 127.50 03NOV2008 07NOV2008 33 f
 3 128.75 28DEC2008 29DEC2008 41 m
 ...
 ;
```

Additional data manipulations are required to prepare the data for analysis.

```
data LibrarySurvey;
   set LibrarySurvey;
   Returned = (CheckIn ^= .);
   if (Returned) then
      lenBorrow = CheckIn              - CheckOut;
   else
      lenBorrow = input('31Dec2008',date9.) - CheckOut;
run;
```

The following statements request Cox regression. The variable Branch is included in the STRATA statement, and the variable SamplingWeight is listed in the WEIGHT statement. These statements define the sample design to the SURVEYPHREG procedure. The MODEL statement defines lenBorrow as the dependent variable, and Age, Sex, and their interaction as the explanatory variables. The outcome can be censored, that is, a book might not be returned at the time of the survey, and so a censoring indicator variable is required. An observation is censored if Returned takes the value 0.

```
proc surveyphreg data = LibrarySurvey;
   class Sex;
   weight SamplingWeight;
   strata Branch;
   model lenBorrow*Returned(0) = Age Sex Age*Sex;
run;
```

Figure 6 displays model information.

Figure 6 Model Information

```
                  The SURVEYPHREG Procedure

                      Model Information

           Data Set                   WORK.LIBRARYSURVEY
           Dependent Variable         lenBorrow
           Censoring Variable         Returned
           Censoring Value(s)         0
           Weight Variable            SamplingWeight
           Stratum Variable           Branch
           Ties Handling              BRESLOW


       Number of Observations Read                   100
       Number of Observations Used                   100
       Sum of Weights Read                      11616.79
       Sum of Weights Used                      11616.79
```

6

Figure 7 displays survey design information; it also indicates that 10 observations are censored and Taylor Series based variance estimation is employed (by default).

Figure 7 Survey Design Information

```
                            Design Summary

                Number of Strata              10


           Summary of the Number of Event and Censored Values

                                                 Percent
                 Total        Event    Censored   Censored

                  100           90          10      10.00


         Summary of the Weighted Number of Event and Censored Values

                                                 Percent
                 Total        Event    Censored   Censored

               11616.79     10449.22    1167.57     10.05


                         Variance Estimation

                    Method    Taylor Series
```

Figure 8 contains the fit statistics.

Figure 8 Fit Statistics

```
                    Model Fit Statistics

                         Without        With
           Criterion    Covariates    Covariates

           -2 LOG L     184459.91     179942.66
           AIC          184459.91     179948.66
```

Figure 9 displays the parameter estimates, standard errors, and *p*-values.

Figure 9 Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
Age	90	0.065626	0.010197	6.44	<.0001	1.068
Sex f	90	-0.014866	0.542021	-0.03	0.9782	0.985
Sex m	90	0	.	.	.	1.000
Age*Sex f	90	-0.004757	0.011048	-0.43	0.6678	0.995
Age*Sex m	90	0	.	.	.	1.000

Analysis of Maximum Likelihood Estimates

The preliminary analysis indicates that only age matters in how long a borrower keeps out a book. You could continue the analysis by fitting a main effects model. For more analysis of these data, refer to the "The SURVEYPHREG Procedure" chapter in the *SAS/STAT 9.22 User's Guide.*

Analyze More Count Data with the GENMOD Procedure

The GENMOD procedure has been a powerhouse for statistical SAS users for years. Designed to fit generalized linear models, it also provides analyses for data that require related techniques, such as GEE analysis for repeated measurements data and zero-inflated Poisson regression for overdispersed data. SAS/STAT 9.22 now provides exact Poisson regression and zero-inflated negative binomial regression.

Exact Poisson Regression

For years, SAS/STAT software has provided a variety of methods for exact inference. Not only are exact counterparts of numerous table statistics available in the FREQ procedure (note that this release adds exact *p*-values for Kendall's tau-*b*, Stuart's Tau-*c*, Somers' D(C|R), and Somers' D(R|C) in this release), the LOGISTIC procedure offers exact logistic regression. In SAS/STAT 9.22, exact Poisson regression comes to the GENMOD procedure with the EXACT statement (which also provides exact logistic regression from PROC GENMOD).

Poisson regression is used to model count data. But often the counts represent events that occur over time or some other measure of size, and then modeling rates becomes of interest. If the exposure time or volume is denoted as N, you write the rate as Y/N and the expected value as μ/N.

The equation

$$\log \frac{\mu}{N} = \alpha + x\beta$$

can be rearranged as

$$\log \mu = \alpha + x\beta + \log(N)$$

The term $\log(N)$ is called an *offset* and must be accounted for in the estimation process.

If you have s independent groups referenced by $i = 1, 2, \ldots, s$, each with a vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{it})$ of t explanatory variables, you can write a likelihood function for the data as

$$\Phi(\mathbf{n}|\boldsymbol{\mu}) = \prod_{i=1}^{s} \mu_i^{n_i} \{\exp(-\mu_i)\}/n_i!$$

where $\mathbf{n} = (n_1, n_2, \ldots, n_s)'$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_s)'$.

The loglinear Poisson model is often written as

$$\log\{n_i\} = \log\{N_i\} + \mathbf{x}_i'\boldsymbol{\beta}$$

in the generalized linear models framework, where the quantity $\log\{N_i\}$ is the offset.

The GENMOD procedure provides Poisson regression when you specify DIST=POISSON and LINK=LOG in the MODEL statement. With the addition of the EXACT statement, PROC GENMOD provides exact Poisson regression, too. Exact Poisson regression is a useful strategy when you have small samples because it does not depend on asymptotic results. Instead, it relies on estimation of the model parameters by using the conditional distributions of the sufficient statistics of the parameters (hence the term exact conditional Poisson regression). For more information about the computational aspects of exact Poisson regression, which are similar to those for exact logistic regression, see the paper Performing Exact Logistic Regression With the SAS System–Revised 2009 by Bob Derr.

The EXACT statement works the same as it does in the LOGISTIC procedure; it requests exact tests of the parameters for the specified effects and optimally estimates those parameters and outputs the exact conditional distributions. You must specify either a logistic regression (LINK=LOGIT and DIST=BIN) or Poisson regression (LINK=LOG and DIST=POISSON) in the MODEL statement to specify an EXACT statement. Postfitting analyses and the BAYES and STORE statements are not available for an exact analysis. The EXACTOPTIONS statement specifies options that control the various computations that produce an exact analysis.

Consider the following data from Laird and Olivier (1981) which are analyzed in Agresti (2002). Patients were classified according to age, type of heart valve, and whether they survived heart valve replacement surgery. They were followed until they died or the study ended. In addition, information was collected about their time at risk or the months during which they were followed. The following DATA step creates these data:

```
data valve;
   input age $ valve $ deaths totalAR;
   TAR=log(totalAR);
datalines;
 <55 aortic 4 1259
 <55 mitral 1 2082
 >=55 aortic 7 1417
 >=55 mitral 9 1647
;
```

The counts are relatively small, so you might have doubts about meeting the sample size requirements for asymptotic analysis. You decide to pursue exact logistic regression.

The following statements request this analysis. In the MODEL statement, you specify DIST=POISSON (LINK=LOG is the default for this distribution) and the explanatory variables age and valve. You also specify the option OFFSET=TAR, which declares TAR as the offset variable. In the EXACT statement, you specify the variables for which you want exact estimation in addition to the ESTIMATE option for explicit parameter estimation. Otherwise, you obtain only the effect tests.

```
proc genmod;
   class age valve / param=ref;
   model deaths = age valve / dist=poisson offset=TAR;
   exact age valve / estimate;
run;
```

Note that the PARAM=REF option is specified in the CLASS statement. This parameterization is recommended for the exact analyses when you have classification variables, and in fact, using the default GLM coding for this model and these data results in infeasibility because of its linear dependencies.

Figure 10 displays the maximum likelihood results. Age appears to be an significant factor. The younger you are, the better off you are.

Figure 10 Parameter Estimates

```
         Analysis Of Maximum Likelihood Parameter Estimates

                                  Standard    Wald 95% Confidence        Wald
Parameter           DF   Estimate    Error          Limits        Chi-Square

Intercept            1    -5.4210   0.3456    -6.0985    -4.7436       245.98
age        <55       1    -1.2209   0.5138    -2.2280    -0.2139         5.65
valve      aortic    1     0.3299   0.4382    -0.5289     1.1887         0.57
Scale                0     1.0000   0.0000     1.0000     1.0000

             Analysis Of Maximum Likelihood
                   Parameter Estimates

             Parameter              Pr > ChiSq

             Intercept                 <.0001
             age         <55           0.0175
             valve       aortic        0.4515
             Scale

NOTE: The scale parameter was held fixed.
```

Figure 11 displays tests for the exact analysis, and Figure 12 displays the parameter estimates. These results are essentially the same as those for the maximum likelihood analysis in this case.

Figure 11 Exact Conditional Analysis

```
                      The GENMOD Procedure

                   Exact Conditional Analysis

                   Conditional Exact Tests

                                       --- p-Value ---
            Effect    Test        Statistic    Exact      Mid

            age       Score          6.3692   0.0145   0.0109
                      Probability    0.00720  0.0145   0.0109
            valve     Score          0.5718   0.5118   0.4467
                      Probability    0.1301   0.5118   0.4467
```

Figure 12 Exact Parameter Estimates

```
                           Exact Parameter Estimates

                                  Standard        95% Confidence
        Parameter          Estimate     Error          Limits           p-Value

        age      <55        -1.2209     0.5138     -2.4740     -0.1675     0.0194
        valve    aortic      0.3299     0.4382     -0.6260      1.2983     0.5889
```

Dealing Overdispersion Another Punch with ZINB Models

You often encounter count data in which you find more variance than what your distributional assumptions suggest. For example, the mean of Poisson data should be the same as the variance, and, when it's larger than the mean, you say that you have overdispersed data. One strategy for dealing with overdispersed data is the negative binomial model. The conditional variance of this distribution can exceed the conditional mean (these are equal for the Poisson distribution) and supports relatively higher frequencies of lower and higher counts.

However, when the overdispersion is thought to come from more zero counts in your data than the Poisson process would produce, zero-inflated methods might be appropriate. This approach assumes that some zero counts are produced from a different process than the other counts, and the analysis proceeds by estimating one set of parameters for a model for those zeros, usually a binomial logit model, and modeling the other counts with either Poisson regression (zero-inflated Poisson regression or ZIP) or the negative binomial regression (zero-inflated negative binomial regression or ZINB). ZIP models were introduced in the GENMOD procedure in SAS/STAT 9.2, and ZINB models are available in PROC GENMOD with the 9.22 release.

The ZEROMODEL statement requests zero-inflated analysis, and you use it to specify the explanatory variables for modeling some zeros with binomial logistic regression. If you specify Poisson regression in the MODEL statement, you get a ZIP analysis, and if you specify negative binomial regression in the MODEL statement, you get a ZINB analysis.

The following example concerns data taken from analyses discussed in Long (1997). Researchers examined how factors such as gender, marital status, number of young children, prestige of the graduate program, and number of articles published by a mentor affect the number of scholarly articles published by scientists.

Figure 13 displays the frequency of articles published.

Figure 13 Article Frequency Counts
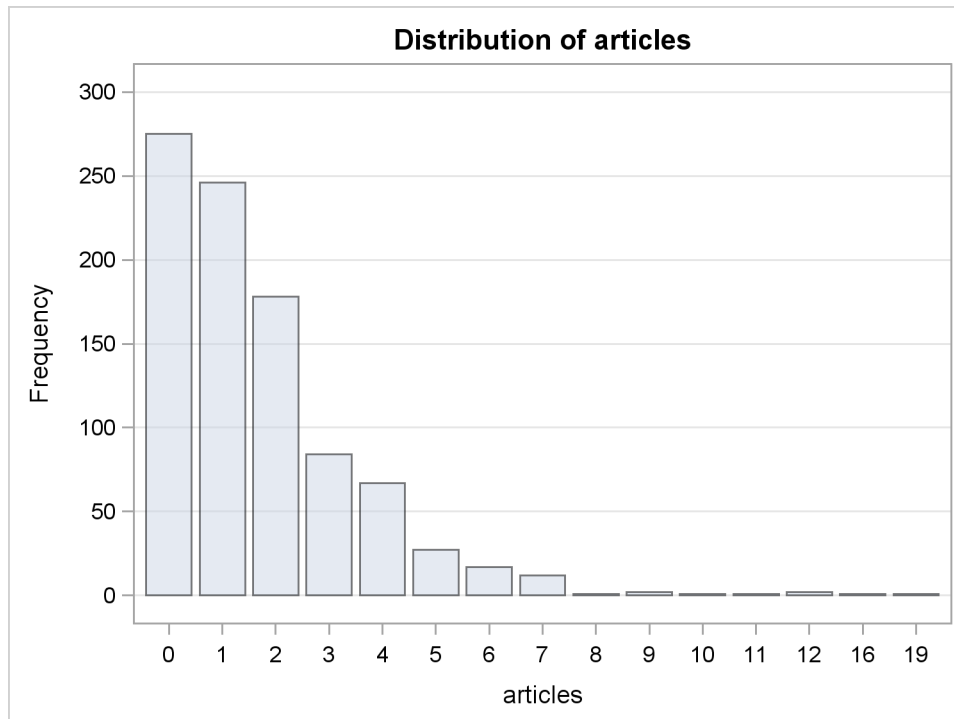
SInce interest lies in predicting article counts with these explanatory variables, the first step may be to fit a Poisson

regression model. If you do so, you will find that the Pearson chi-square/DF criterion has a value of 1.82 (not shown here), a sure indication of possible overdispersion as this ratio should be close to 1. The next step might be negative binomial regression (actually this fits the data fairly well), but you might want to consider zero-inflated methods since there are a considerable number of zero counts in these data.

It's possible that some of the zeros in the model are structural; that is, some scientists might be in positions where publishing is not an objective, and their zeros are thus different from those of other scientists who couldn't find writing time or adequate department support. Fitting zero-inflated models to these data makes sense.

The following statements produce a ZINB analysis with the GENMOD procedure. The ZEROMODEL statement specifies this analysis in conjunction with the DIST=ZINB option in the MODEL statement. The explanatory variables female, married, kid5, phd, and mentor are specified for both models.

```
proc genmod data=long97data;
   model articles=female married kid5 phd mentor / dist=zinb;
   zeromodel female married kid5 phd mentor ;
   ods output ParameterEstimates=zinb ZeroParameterEstimates=zeronb;
run;
```

Figure 14 displays the model fit information. The value of 1.03 for Pearson chi-square/DF indicates that overdispersion is no longer an issue.

Figure 14 Model Fit

```
              Criteria For Assessing Goodness Of Fit

      Criterion                    DF          Value        Value/DF

      Deviance                              3099.9818
      Scaled Deviance                       3099.9818
      Pearson Chi-Square          903        933.4044         1.0337
      Scaled Pearson X2           903        933.4044         1.0337
      Log Likelihood                       -1549.9909
      Full Log Likelihood                  -1549.9909
      AIC (smaller is better)               3125.9818
      AICC (smaller is better)              3126.3858
      BIC (smaller is better)               3188.6278
```

Figure 15 and Figure 16 display the two sets of parameter estimates for this model. Note that the estimates for the "zero" model mostly take the opposite sign of the standard parameter estimates.

Figure 15 Parameter Estimates

```
           Analysis Of Maximum Likelihood Parameter Estimates

                        Standard      Wald 95%           Wald
    Parameter   DF   Estimate    Error   Confidence Limits   Chi-Square  Pr > ChiSq

    Intercept    1     0.4167   0.1436    0.1353    0.6982       8.42      0.0037
    female       1    -0.1955   0.0756   -0.3437   -0.0473       6.69      0.0097
    married      1     0.0976   0.0845   -0.0679    0.2631       1.34      0.2479
    kid5         1    -0.1517   0.0542   -0.2580   -0.0455       7.84      0.0051
    phd          1    -0.0007   0.0363   -0.0718    0.0704       0.00      0.9846
    mentor       1     0.0248   0.0035    0.0179    0.0316      50.37     <.0001
    Dispersion   1     0.3767   0.0510    0.2888    0.4912

    NOTE: The negative binomial dispersion parameter was estimated by maximum
          likelihood.
```

Figure 16 Zero Inflation Parameter Estimates

```
        Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates

                        Standard      Wald 95%           Wald
    Parameter   DF   Estimate    Error   Confidence Limits   Chi-Square  Pr > ChiSq

    Intercept    1    -0.1917   1.3228   -2.7844    2.4010       0.02      0.8848
    female       1     0.6359   0.8489   -1.0279    2.2998       0.56      0.4538
    married      1    -1.4995   0.9387   -3.3392    0.3403       2.55      0.1102
    kid5         1     0.6284   0.4428   -0.2394    1.4963       2.01      0.1558
    phd          1    -0.0377   0.3080   -0.6414    0.5660       0.01      0.9025
    mentor       1    -0.8823   0.3162   -1.5021   -0.2625       7.78      0.0053
```

Finally, Figure 17 displays the model-predicted counts for four models.

Figure 17 Model Comparisons
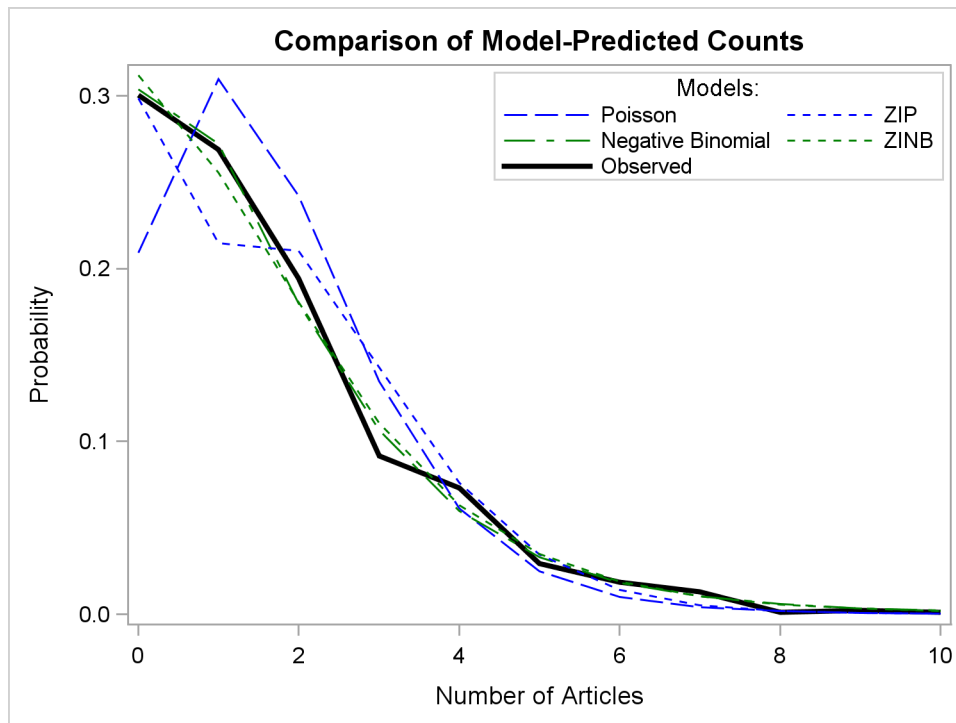
Long (1997) suggests that if an effect has a positive impact on counts and it also positively impacts the chances of being in the group for which positive counts are likely, then its parameter estimates will have opposite signs.

Poisson regression is clearly the worst-fitting model, greatly underpredicting zero counts. The ZIP model does better, but it underpredicts the lower counts. Both the negative binomial and the zero-inflated negative binomial models work well. Your ultimate choice might be based on your perception of whether a separate structural zeros process makes sense.

The Commons

You might have noticed that the SAS/STAT 9.2 documentation includes a new chapter called "Shared Concepts and Topics." One of the motivations for this new chapter was the introduction of the new EFFECT and EFFECTPLOT statements, which behave exactly the same in the procedures that support them. This chapter is expanded in the SAS/STAT 9.22 documentation to include the shared ESTIMATE, LSMEANS, LSMESTIMATE, SLICE, STORE, and TEST statements. While each procedure has a section devoted to these statements and provides procedure-specific information, these sections also link to this chapter for general information.

Speaking of the EFFECT Statement

The EFFECT statement remains experimental for SAS/STAT 9.22. This statement, which defines a richer family of linear models in several additional procedures, is still under evaluation and could change in future releases. It is now available in the HPMIXED, GLIMMIX, GLMSELECT, LOGISTIC, ORTHOREG, PHREG, PLS, QUANTREG, ROBUSTREG, SURVEYLOGISTIC, and SURVEYREG procedures. Effect types include splines for semiparametric modeling, effects for situations in which measurements can belong to more than one class, lag effects (new in SAS/STAT 9.22), and polynomials.

Interval Censoring

Usually, new SAS/STAT capabilities surface in new procedures or existing procedures, but sometimes they surface as SAS-written macros. This is the case for new facilities for interval-censoring, a technique used in survival analysis. Survival data analysis usually focuses on lifetimes using measured time to an event of interest, or the latest time

available if the event did not occur during the observation period. But sometimes the event is not directly observed and is only known to have occurred within some interval of time. This often occurs in longitudinal studies in which subjects are only assessed periodically.

Data measured in this way are called interval-censored survival data, and special methods are required to analyze data that include such observations. The LIFEREG and RELIABILITY procedures fit popular lifetime distributions such as the Weibull and lognormal to interval-censored data by maximum likelihood estimation of distribution parameters. However, nonparametric methods are also of interest.

SAS has recently released new macros for computing nonparametric maximum likelihood estimators of the survival function for these data, along with macros for appropriate logrank tests. See Analyzing Interval-Censored Survival Data with SAS by So, Johnston, and Kim (2010) for more information.

Updates to ROBUSTREG and QUANTILE Procedures

In recent releases, SAS/STAT software has introduced facilities for modern regression analysis. The ROBUSTREG procedure performs robust regression, which is mainly used to detect outliers and to produce stable results in their presence. In SAS/STAT 9.22, PROC ROBUSTREG has been updated with the ability to include effects specified in the CLASS statement for the LTS, S, and MM methods. In addition, PROC ROBUSTREG now computes a robust version of the Mahalanobis distance by using the generalized minimum covariance determinant (MCD) method. Leverage point analysis is updated to reflect the inclusion of classification variables.

And, as mentioned previously, the experimental EFFECT statement is now available in PROC ROBUSTREG.

The QUANTREG procedure models the effects of explanatory variables on the conditional quantiles of a response variable. Quantile regression is particularly useful when the rate of change in the conditional quantiles, as expressed by the regression coefficients, depends on the quantiles. With SAS/STAT 9.22, if you specify multiple quantiles in a MODEL statement in PROC QUANTREG, additional analyses, such as those specified in the TEST statement, are now produced for each quantile specified.

In addition, rank tests are now produced by PROC QUANTREG with the RANKSCORE option in the TEST statement. Normal scores, Wilcoxon scores, and sign scores are available as score functions.

A Step Up for Spatial Analysis in SAS

Spatial data surface in fields such as environmental sciences, epidemiology, health risk analysis, and energy exploration. To analyze such data you need to deal with observations for one or more attributes at a series of locations. You generally investigate and model correlations among these observations first in order to learn about the underlying process. After you produce a suitable mathematical characterization of the correlation behavior for each of the process attributes, you can predict values at unsampled spatial locations, produce spatial maps, and extract secondary information for further analysis.

SAS/STAT software includes three procedures for this type of analysis of spatial data, also known as geostatistics:

- The VARIOGRAM procedure explores and models underlying correlations in spatial data.

- The KRIGE2D procedure uses a given correlation model to interpolate values at unsampled locations.

- The SIM2D procedure simulates spatial processes with a specified correlation structure.

SAS/STAT 9.2 introduced ODS Statistical Graphics to these procedures, and SAS/STAT 9.22 introduces exciting additional updates to provide seamless analysis and prediction of spatial processes. These new features include:

- more correlation models

- automated semivariogram modeling

- item stores to pass semivariance information to KRIGE2D or SIM2D procedures

- customizable graphics for spatial data exploration and visualization of resulting predictions amd simulations

The paper Everything in Its Place: Efficient Geostatistical Analysis with SAS/STAT Spatial Procedures by Alexander Kolovos (2010) describes how to perform a typical spatial analysis using some of the new features of SAS/STAT 9.22. The following plots come from that paper and illustrate the types of graphs you produce as you explore spatial data, fit

13

a semivariogram, estimate an empirical semivariogram, compare empirical and fitted theoretical semivariograms, and evaluate model prediction.

Figure 18 displays the spatial distribution of daily average global solar irradiance (daGSI) in a square area. Figure 19 displays an empirical estimate of the daGSI semivariogram, which indicates the existence of surface trends.
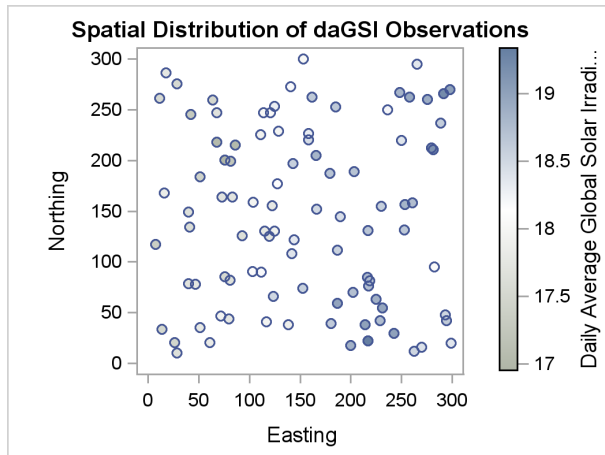
Figure 18   Spatial Distribution of Observations
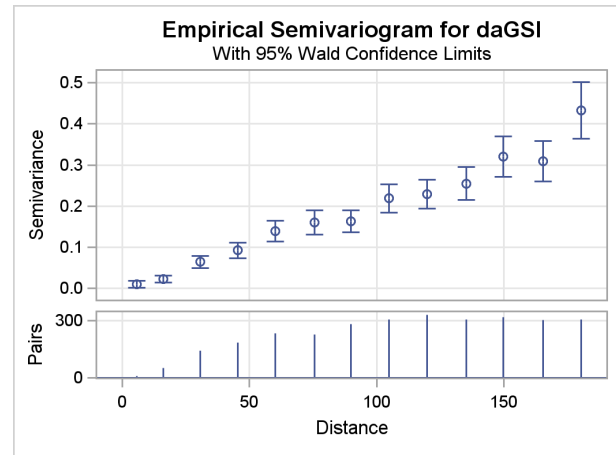

Figure 19   Classical Empirical Semivariogram

Figure 20 displays the empirical semivariogram based on residuals from the detrended data as well as a comparison of theoretical models. Figure 21 displays the predicted values and standard errors of the residual daGSI on a regular grid of unsampled locations in the target area.

Figure 20   Fitted Theoretical and Empirical Semivariogram

Figure 21   Prediction Contours and Standard Errors
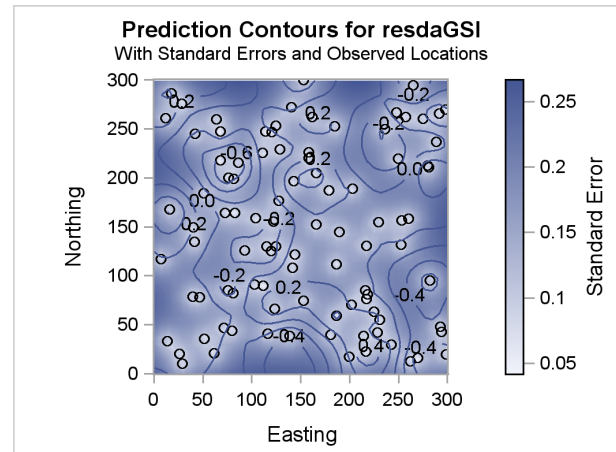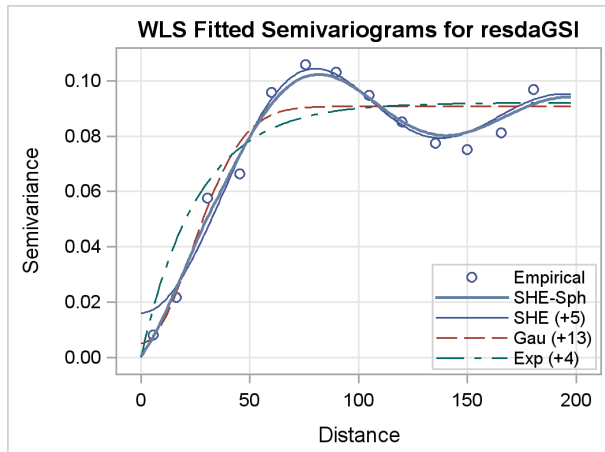
Model Averaging Comes to the GLMSELECT Procedure

While many data analysts employ automatic variable selection methods, more than a few statisticians caution against their use because of numerous pitfalls that plague such methods. Model averaging approaches provide a way to address these issues and make more stable inferences based on a set of models. Methods for doing this are presented in Burnham and Anderson (2002), and Bayesian approaches are discussed in Raftery, Madigan, and Hoeting (1997).

One approach for performing model averaging is to use resampled data as a proxy for multiple samples that are drawn from some conceptual probability distribution. A model is selected for each resampled set of data, and a predictive model is built by averaging the predictions of these selected models. You can perform this type of model averaging with the new MODELAVERAGE statement. Resampling-based methods, in which samples are obtained by drawing with replacement from your data, fall under the umbrella of the widely studied methodology known as the bootstrap (Efron and Tibshirani, 1993). For use of the bootstrap in the context of variable selection, see Breiman (1992).

14

The Hald data (Woods, Stein, and Starke, 1932) is a small data set that has been used by various authors in studying variable selection methods (see Burnham and Anderson, 2002 and the references therein). This data set consist of 13 observations of the heat given off per gram of cement after 180 days of hardening. The predictors $x1$–$x4$ are the percentages by weight of calcium aluminate, tricalcium silicate, tetracalcium alunimo ferrite and dicalcim silicate respectively.

```
data hald;
   input x1 x2 x3 x4 y;
datalines;
 7   26    6   60    78.5
 1   29   15   52    74.3
11   56    8   20   104.3
11   31    8   47    87.6
 7   52    6   33    95.9
11   55    9   22   109.2
 3   71   17    6   102.7
 1   31   22   44    72.5
 2   54   18   22    93.1
21   47    4   26   115.9
 1   40   23   34    83.8
11   66    9   12   113.3
10   68    8   12   109.4
;
```

The following statements show how you can use the MODELAVERAGE statement in PROC GLMSELECT to perform bootstrap-based model averaging for the Hald data. 10,000 bootstrap samples are requested, and the BEST=1 sub-option of the SUBSET option specifies that only the most frequently selected model be used in forming the average model.

```
proc glmselect data=hald plots=all;
   model y=x1-x4/selection=stepwise(select=aicc);
   modelAverage nsamples=10000 subset(best=1);
run;
```

Following the methodology used in Burnham and Anderson (2002), models are selected using stepwise selection based on the bias corrected Akaike information criterion (AICC) for each of 10,000 bootstrap samples of the data. Figure 22 shows the "Model Selection Frequency" table that displays the number of times each model is selected. These results suggest that the models with regressors {x1 x2} and {x1 x4} are appropriate for these data and that larger models are overparameterized. As noted by Burnham and Anderson (2002), these models are somewhat aliased because x2 and x4 are strongly negatively correlated ($r = -0.97$).

Figure 22 Model Selection Frequency

```
                 Comparison of Model-Predicted Counts

                     The GLMSELECT Procedure

                     Model Selection Frequency

                          Number
         Times      Selection      of    Frequency
       Selected    Percentage  Effects       Score    Effects in Model

         5254         52.54         3        5255    Intercept x1 x2
       ------------------------------------------------------------------
  *      2601         26.01         3        2602    Intercept x1 x4
  *       752          7.52         4        752.7   Intercept x1 x2 x3
  *       595          5.95         4        595.6   Intercept x1 x3 x4
  *       478          4.78         4        478.8   Intercept x1 x2 x4
  *       171          1.71         5        171.6   Intercept x1 x2 x3 x4
  *       131          1.31         4        131.6   Intercept x2 x3 x4
  *        11          0.11         3         11.52  Intercept x3 x4
  *         3          0.03         2          3.58  Intercept x3
  *         1          0.01         2          1.84  Intercept x2
  *         1          0.01         3          1.72  Intercept x1 x3
  *         1          0.01         2          1.70  Intercept x4
  *         1          0.01         3          1.62  Intercept x2 x3

                  * Not Used in Model Averaging
```

Figure 23 shows the average parameter estimates, and the distributions of the parameter estimates are shown graphically in Figure 24.

Figure 23 Average Parameter Estimates

```
                          Average Parameter Estimates

                      Number        Non-zero             Mean       Standard
          Parameter   Non-zero     Percentage         Estimate      Deviation

          Intercept     5254          100.00          51.976271      1.952022
          x1            5254          100.00           1.441448      0.141801
          x2            5254          100.00           0.676292      0.042635

                          Average Parameter Estimates

                              ----------Estimate Quantiles----------
                 Parameter           25%         Median           75%

                 Intercept      50.703669      51.925759      53.110692
                 x1              1.359533       1.466379       1.538196
                 x2              0.650838       0.673397       0.699057
```

Figure 24 Effect Selection Percentages
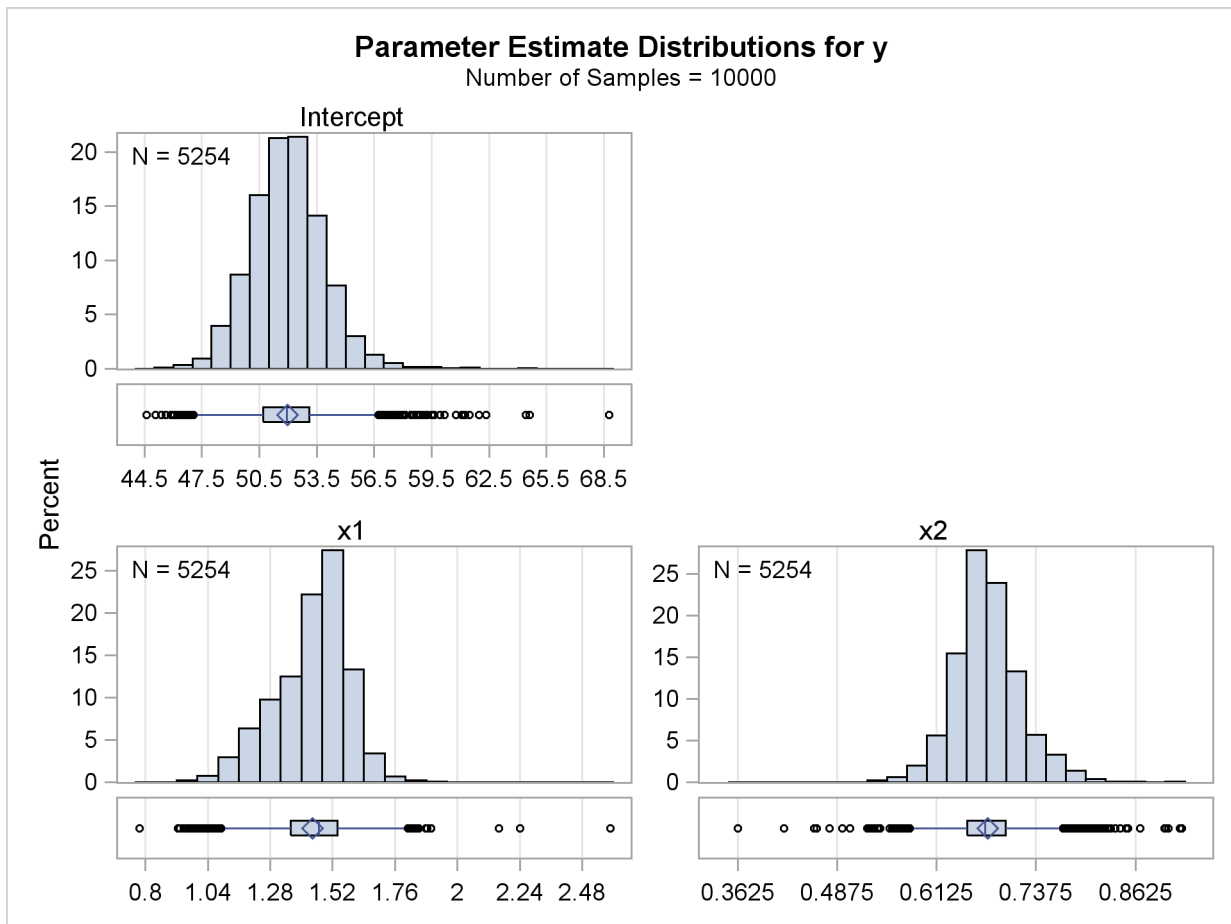
Last But Certainly Not Least

SAS/STAT 9.22 software contains a host of other enhancements as well. The following sections describe the other highlights of this release.

Bayesian Software Updates

Bayesian capabilities continue to grow in SAS/STAT software. First available via web-downloadable procedures, the capabilities provided by the BAYES statement in the GENMOD, LIFEREG, and PHREG procedures have been updated with new sampling methods. Conjugate sampling for the linear regression in the GENMOD procedure is now the

default, which provides computation time reductions. In addition, you can specify either the Gamerman algorithm or the independent Metropolis algorithm with the SAMPLING= option in the BAYES statement in the GENMOD procedure for other generalized linear models. You can choose the random-walk Metropolis algorithm as an alternative sampling method in the PHREG procedure. PROC PHREG also added the Zellner g-prior for the regression coefficients.

The MCMC procedure introduces the PREDDIST statement in this release. This statement enables you to create random samples from the posterior predictive distribution of the response variables. The posterior predictive distribution is the distribution of unobserved observations (predictions) conditional on the observed data. You can find a good introduction to the MCMC procedure in the SAS Global Forum paper Bayesian Modeling Using the MCMC Procedure by Fang Chen (2009).

HPMIXED Grows Up

The HPMIXED procedure provides mixed models analysis for when you have large amounts of data and/or large numbers of class levels that may be beyond the grasp of the MIXED procedure. PROC HPMIXED employs sparse matrix methods and other numerical techniques to fit the required mixed models. Experimental in SAS/STAT 9.2, the HPMIXED procedure becomes production with SAS/STAT 9.22. While it does not, and will not, provide the full-fledged capabilities of the MIXED procedure, PROC HPMIXED fits a very useful subset of mixed models for the purposes of large data/number of effects analysis. The paper All the Cows in Canada: Massive Mixed Modeling with the HPMIXED Procedure in SAS® 9.2 provides more information.

With SAS/STAT 9.22, the HPMIXED procedure includes the experimental EFFECT statement. The BLUP= option in the PROC HPMIXED statement creates a data set that contains the BLUE and BLUP options. In addition, the SLICE and DIFF options are now supported in the LSMEANS statement.

Welcome to a New CALIS Procedure

The CALIS procedure provides the analysis of structural equation modeling in SAS. This procedure has been completely redesigned to provide much more powerful and customizable modeling. Most of these updates were introduced in SAS/STAT 9.2 with the TCALIS procedure, and those updates have now been rolled into the production CALIS procedure, along with some additional features.

These capabilities include:

- new modeling languages such as LISMOD, MSTRUCT, and PATH

- multiple group analysis

- improved mean structure analysis

- general parametric function testing

- improved effect analysis

In addition, PROC CALIS introduces several experimental features, including the full information likelihood method (FIML), mean structure analysis with the COSAN model, unnamed free parameter specification, and an extended path modeling language. For an overview of the enhancements, see Structural Equation Modeling and Path Analysis Using PROC TCALIS in SAS® 9.2.

More ODS Statistical Graphics, and Oh, Did I Show You the EFFECTPLOT Statement?

ODS Statistical Graphics was made production in SAS/STAT 9.2, and it continues to come on strong with this latest release. The majority of SAS/STAT procedures now offer ODS graphics and provide over 400 graphs in all. Graphics have arrived in the SURVEYFREQ and TPSPLINE procedures, and new graphs are now available in the FREQ, SIM2D, KRIGE2D, and VARIOGRAM procedures. For example, PROC FREQ now produces relative risk and risk difference plots, and PROC VARIOGRAM now offers the Moran plot, as displayed in Figure 25 and Figure 26, respectively.

Figure 27 displays contour surface plots created by the TPSPLINE procedure to assess the fit of models with different smoothing values. These plots come from "Example 90.3 Multiple Minima of the GVC Function" in the chapter "The TPSPLINE Procedure" in the *SAS/STAT User's Guide*.

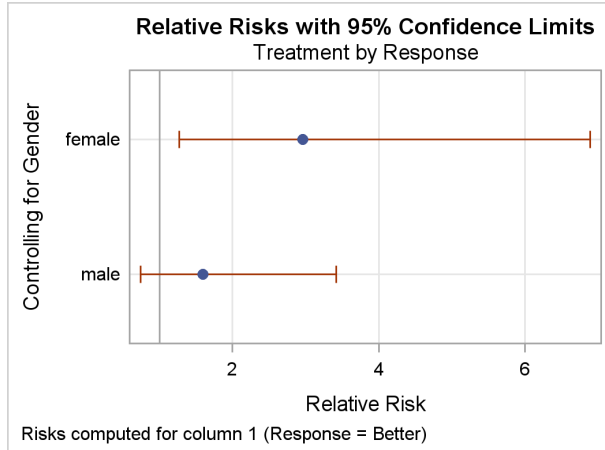**Figure 25** Relative Risk Plot
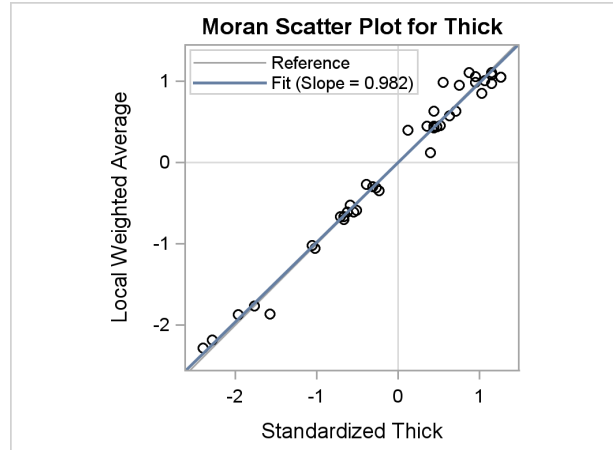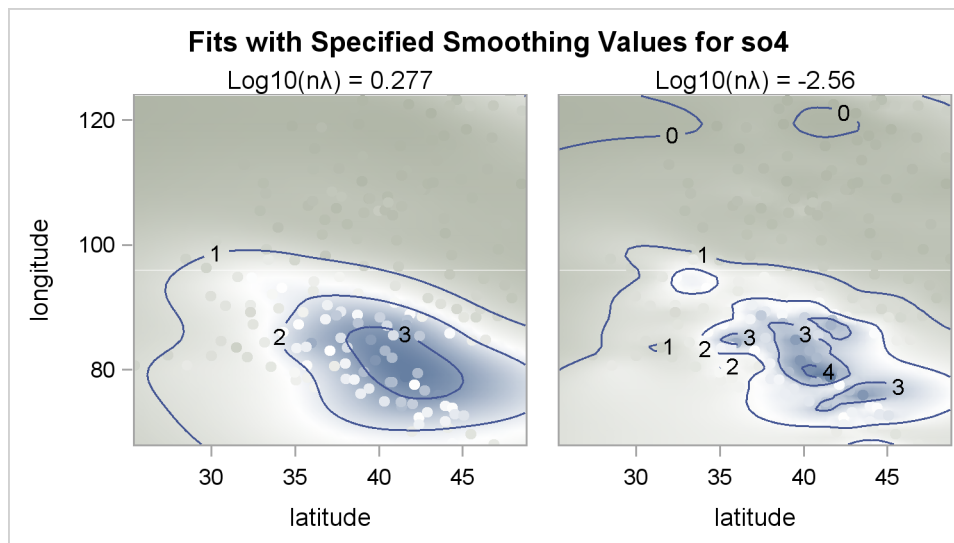

Figure 26 Moran Plot

Figure 27 Contour Surfaces for Models Fit with PROC TPSPLINE

The EFFECTPLOT statement was introduced in SAS/STAT 9.2 in the LOGISTIC procedure, and it is now available in the SURVEYLOGISTIC and the GENMOD procedure, too. This statement, along with its many options, is extremely helpful in visualizing the influence of the levels of your model effects, by themselves, sliced by other effects, or for a specific value of a continuous effect in the model.

For an example of the usefulness of this plot, consider the following analysis of the Neuralgia data that are displayed in "Example 51.2 Logistic Modeling with Categorical Predictors" in "The LOGISTIC Procedure" chapter in the *SAS/STAT 9.22 User's Guide*. Two test treatments for managing the pain of neuralgia in the elderly are compared, along with a placebo. Age, gender, and duration of complaint are also recorded. The final model chosen is one with main effects for treatment, sex, and age; treatment and sex are classification effects.

```
proc logistic data=Neuralgia;
   class Treatment Sex /param=ref;
   model Pain= Treatment Sex Age;
   effectplot / at(Sex=all);
   effectplot slicefit(sliceby=Sex plotby=Treatment) / noobs;
run;
```

The first EFFECTPLOT statement requests the predicted probabilities for no pain (the predicted outcome level) by age by sex (requested with the AT(SEX=ALL) option, with each treatment represented separately on the plot. Figure 28 displays this graph, and it clearly shows that the placebo does worse than the other treatments and younger patients do better than older patients. The second EFFECTPLOT statement produces the same plot except that sex and treatment are reversed. Figure 29 shows that females do better than males for all treatments.

Numerous other plots are available in the LOGISTIC procedure, including influence plots, ROC curves, and odds ratio plots.

Figure 28 Predicted Probabilities by Sex
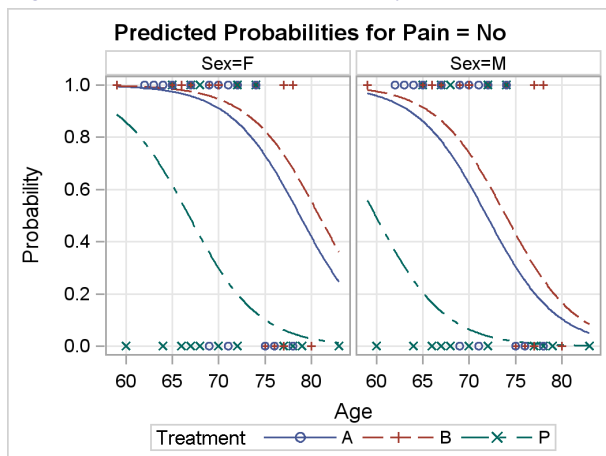
Figure 29 Predicted Probabilities by Treatment
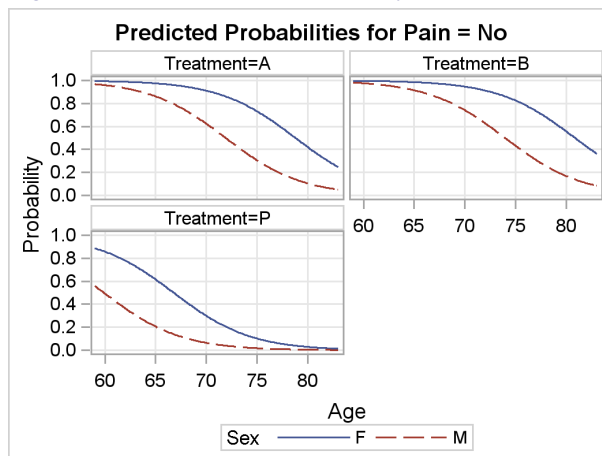
See Getting Started with ODS Statistical Graphics in SAS 9.2–Revised 2009 by Bob Rodriguez (2009) for a good discussion of ODS Statistical Graphics and how it puts powerful statistical graphs at your fingertips.

For Further Information

A good place to start for further information is the "What's New in SAS/STAT 9.22" chapter in the online documentation for SAS/STAT 9.22. The link `support.sas.com/whatsnewstat922/` will take you there as soon as the software is released in the second quarter of 2010. In addition, the Statistics and Operations Focus Area is chock full of useful information about the statistical products, and can be found at `support.sas.com/statistics/`. The quarterly e-newsletter for that site is available on its home page. And of course, full information is available in the online documentation located here: `support.sas.com/documentation/onlinedoc/stat/`.

The papers referenced in the text link to the appropriate paper if you read it online.

References

Agresti, A. (2002), *Categorical Data Analysis, Second Edition*, New York: John Wiley & Sons.

Binder, D. A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *International Statistical Review,* 51, 279–292.

Burnham, K. P., and Anderson, D. R. (2002), *Model Selection and Multimodel Inference,* Second Edition, New York: Springer-Verlag.

Chen, F. (2009), "Bayesian Modeling Using the MCMC Procedure," in *Proceedings of the SAS Global Forum 2008 Conference,* Cary NC: SAS Institute Inc.

Derr, R. (2009). "Performing Exact Logistic Regression with the SAS System–Revised 2009," SAS Institute, Inc. `support.sas.com/rnd/app/papers/exactlogistic2009.pdf`

Efron B., and Tibshirani, R. (1993), *An Introduction to the Bootstrap,* Chapman & Hall/CRC.

Gentleman, W. M. (1972), *Basic Procedures for Large, Sparse, or Weighted Least Squares Problems*, Technical Report CSRR-2068, University of Waterloo, Ontario.

Gentleman, W. M. (1973), "Least Squares Computations by Givens Transformations without Square Roots," J. Inst. Math. Appl., 12, 329–336.

Kolovos, A. (2010), "Everything in Its Place: Efficient Geostatistical Analysis with SAS/STAT Spatial Procedures," in Proceedings of the SAS Global Forum 2010 Conference, Cary NC: SAS Institute Inc.

Laird, N. M. and D. Olivier (1981),"Covariance analysis of censored survival data using long-linear analysis techniques," *Journal of the American Statistical Association*, 76, 231–240.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association,* 92, 179–191.

Rodriguez, R. (2009) "Getting Started with ODS Statistical Graphics in SAS 9.2–Revised 2009," SAS Institute, Inc `support.sas.com/rnd/app/papers/intodsgraph.pdf`

Schabenberger, O. (2007), "Growing Up Fast: SAS 9.2 Enhancements to the GLIMMIX Procedure," in *Proceedings of the SAS Global Forum 2007 Conference,* Cary NC: SAS Institute Inc.

Silvapulle, M. J. and Sen, P. K. (2004), *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*, New York: John Wiley & Sons.

So, Y., Johnston, G., and S.H. Kim (2010), "Analyzing Interval-Censored Survival Data with SAS Software," in *Proceedings of the SAS Global Forum 2010 Conference,* Cary NC: SAS Institute Inc.

Tobias, R. and W. Cai (2010), "Introducing PROC PLM and Postfitting Analysis for Very General Linear Models in SAS/STAT 9.22" in *Proceedings of the SAS Global Forum 2010 Conference,* Cary NC: SAS Institute Inc.

Wang, T. and R. Tobias (2009), "All the Cows in Canada: Massive Mixed Modeling with the HPMIXED Procedure in SAS" in *Proceedings of the SAS Global Forum 2009 Conference,* Cary NC: SAS Institute Inc.

Winer, B. J. (1971), *Statistical Principles in Experimental Design, Second Edition,* New York: McGraw-Hill.

Woods, H., Steinour, H.H., and Starke, H.R., "Effect of Composition of Portland Cement on Heat Evolved During Hardening," *Industrial and Engineering Chemistry*, 24, 1207–1214.

Yung, Y. (2008), "Structural Equation Modeling and Path Analysis Using PROC TCALIS in SAS 9.2," in *Proceedings of the SAS Global Forum 2008 Conference,* Cary NC: SAS Institute Inc.

ACKNOWLEDGMENTS

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Maura Stokes
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
maura.stokes@sas.com