Paper SP03_2011

On Deck: SAS/STAT® 9.3

Maura Stokes, Fang Chen, and Ying So
SAS Institute, Cary NC
Presented by Funda Gunes

Abstract

SAS/STAT® 9.3, coming soon to a site near you, delivers numerous enhancements to the statistical software. The PHREG procedure supports frailty models for incorporating random effects in Cox regression, and the MCMC procedure provides a RANDOM statement to facilitate fitting Bayesian models with random effects. The NLIN procedure has been updated, and the MI procedure offers additional flexibility by providing a fully conditional specification method. The new SURVEYPHREG and HPMIXED procedures are also outfitted with additional capabilities.

This talk reviews the highlights of SAS/STAT 9.22 and then describes important SAS/STAT 9.3 enhancements with practical illustrations, mainly from the SAS/STAT 9.3 documentation.

Introduction

SAS/STAT has undergone considerable updates in recent years: 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.

In 2009, SAS/STAT 9.22 arrived. It made a full complement of postfitting capabilities available in a large number of linear modeling procedures as well as introducing the PLM procedure, which enables you to take stored model information and use it to perform additional inference and scoring without refitting the original model. The SURVEYPHREG procedure provides survival analysis, in the form of Cox proportional hazards regression, for sample survey data. The more powerful and customizable structural equation modeling first implemented with the experimental TCALIS procedure in SAS/STAT 9.2 was rolled into the CALIS procedure. Other enhancements include exact Poisson regression, zero-inflated negative binomial models, model-averaging for high-dimension prediction, and improved prediction of spatial processes with the spatial analysis procedures.

SAS/STAT 9.3, available later this year, both introduces new methodology as well as continues to refine existing functionality. The SURVEYPHREG procedure becomes production and now handles time-dependent covariates. The EFFECT statement also becomes production and is available in eleven linear modeling procedures. The MCMC procedure provides a RANDOM statement, which simplifies the specification of hierarchical random-effects models and significantly reduces simulation time which improving convergence. This statement defines random effects that can enter the model in a linear or nonlinear fashion and supports univariate and multivariate prior distributions.

The experimental FMM procedure fits statistical models to data where the distribution of the response is a finite mixture of univariate distributions. These models are useful for applications such as estimating multimodal or heavy-tailed densities, fitting zero-inflated or hurdle models to count data with excess zeros, modeling overdispersed data, and fitting regression models with complex error distributions. The MI procedure now includes the experimental FCS statement, which specifies a multivariate imputation by fully conditional specification (FCS) methods. Finally, The NLIN procedure has been updated with experimental features for diagnosing the nonlinear model fit.

This paper reviews some of the new capabilities in the SAS/STAT 9.3 release. It draws heavily from the documentation in progress for this release.

A Review of SAS/STAT 9.22

Many users are still finding out about SAS/STAT 9.22, a new release that 'snuck out' with the third maintenance of SAS® 9.3. This release contained many important enhancements to SAS/STAT software.

New architectural changes mean that many procedures added additional postfitting capabilities to their arsenal. Over 30 postfitting statements have been added to the linear modeling procedures. In addition, several existing statements have been updated. Table 1 provides an overview of these capabilities. 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	⋆		⋆	✓	✓	✓	✓

In addition, the 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 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. For statistical inference, PROC SURVEYPHREG incorporates complex survey sample designs, including designs with stratification, clustering, and unequal weighting.

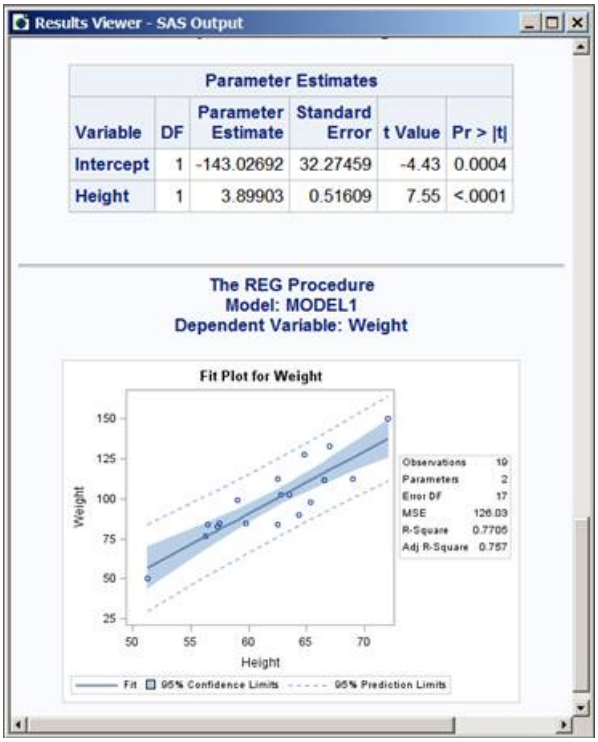
The GENMOD procedure is equipped with an EXACT statement and performs exact Poisson regression as well as exact logistic regression. In addition, it performs zero-inflated negative binomial regression as well as zero-inflated Poisson regression. The EFFECT statement is available in the HPMIXED, GLIMMIX, 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, and polynomials.

The SAS/STAT 9.22 release contains numerous other additions as well. See the paper "The Next Generation: SASS/STAT® 9.22" by Stokes, Rodriguez, and Cohen (2010)

ODS Statistical Graphics Footprint Grows Larger

The 9.3 release marks another milestone in the evolution of statistical graphics in SAS. The ODS graphical capabilities move to SAS® Base so they will be available to all SAS/STAT customers. This means that a SAS/GRAPH® license will no longer be required to generate the plots produced by the statistical procedures. The SG family of procedures is also part of Base SAS, and so is the ODS Graphics editor and the Graphical Template Language. You might be interested in the ODS Graphics Designer, which enables you to create new graphics via a point-and-click interface.(Note that SAS/GRAPH remains a SAS product with extensive graphics functionality.)

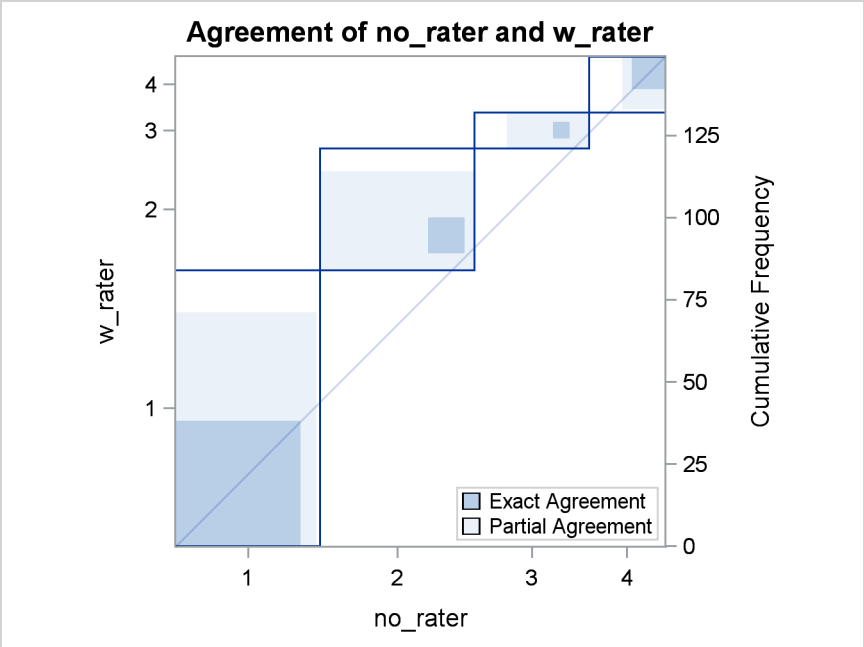
Figure 1 New DMS Look



In addition, the default Display Manager settings have been changed so that resulting tables and graphs are interspersed. ODS Graphics is now automatically turned on in Display Manager, and the default output destination is HTML. The new HTMLBlue style affords a modern appearance. As a result of these changes, a few graphs behave somewhat differently, mostly having to do with the number of points being plotted. See the "What's Changed" section in the "What's New" chapter of the SAS/STAT 9.3 documentation for more detail.

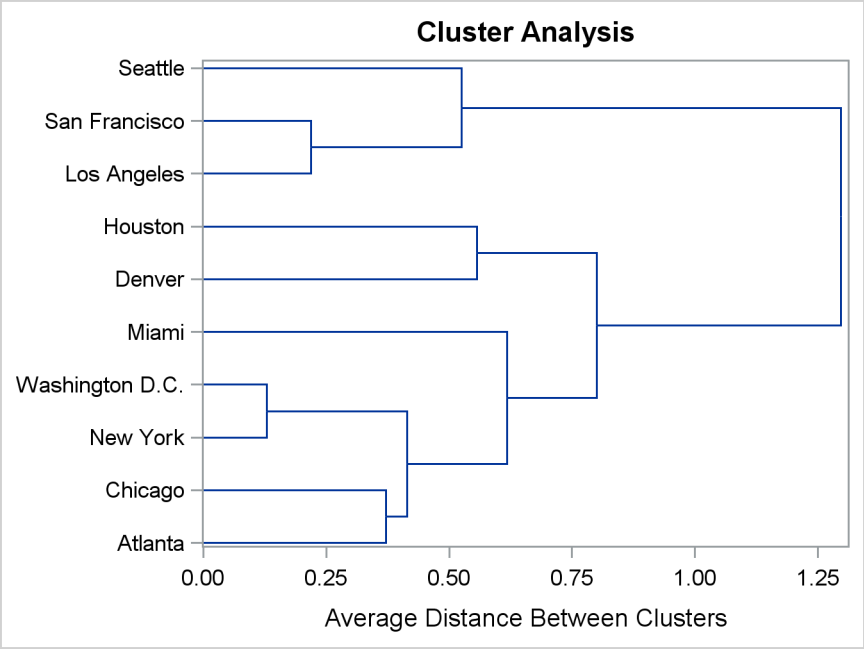
In addition, several more graphics have been added to SAS/STAT. Figure 2 displays the agreement plot corresponding to observer agreement analysis in the FREQ procedure:

Figure 2 Agreement Plot for MS Patient Study



And the CLUSTER and VARCLUS procedures now produce dendrograms when ODS Graphics is enabled as illustrated in Figure 3.

Figure 3 Dendrogram for Mileage Data



Finite Mixture Models

Finite mixture models allow you to fit statistical models to data when the distribution of the response is a finite mixture of univariate distributions. These models are useful for applications such as estimating multimodal or heavy-tailed densities, fitting zero-inflated or hurdle models to count data with excess zeros, modeling overdispersed data, and fitting regression models with complex error distributions. Many well-known statistical models for dealing with overdispersed data are members of the finite mixture model family: for example, zero-inflated Poisson models and zero-inflated negative binomial models.

The experimental FMM procedure in SAS/STAT 9.3 fits finite mixtures of regression models of generalized linear models in which the regression structure and the covariates are the same or different across the components. PROC FMM performs maximum likelihood estimation for all models, and it provides Markov Chain Monte Carlo estimation for many models, including zero-inflated Poisson models. The procedure includes many built-in link and distribution functions, including the beta, shifted, Weibull, beta-binomial, and generalized Poisson distributions, as well as standard members of the exponential family of distributions. In addition, several specialized built-in mixture models are provided, such as the binomial cluster model (Morel and Nagaraj, 1993).

The FMM procedure uses the typical model-building syntax of SAS statistical procedures, including the CLASS and MODEL statements, although multiple MODEL statements are required to build the mixture models in which model effects, distributions, or link functions vary across mixture components. It also provides a number of features particular to fitting finite mixture models. For example, it evaluates sequences of mixture models when you specify ranges for the number of components, it has the ability to model regression and classification effects in the mixing probabilities, and it can incorporate full or partially known component membership into the analysis. In addition, Bayesian methods are available for several models.

The following data illustrates the use of a zero-inflated model. In a survey of park attendees, randomly selected individuals were asked about the number of fish they caught in the last six months. Along with that count, the gender and age of each sampled individual were recorded. The following DATA step displays the data for the analysis:

```
data catch;
   input gender $ age count @@;
   datalines;
  F   54  18   M   37   0   F   48  12   M   27   0
  M   55   0   M   32   0   F   49  12   F   45  11
  M   39   0   F   34   1   F   50   0   M   52   4
  M   33   0   M   32   0   F   23   1   F   17   0
```

```
  F   44   5   M   44   0   F   26   0   F   30   0
  F   38   0   F   38   0   F   52  18   M   23   1
  F   23   0   M   32   0   F   33   3   M   26   0
  F   46   8   M   45   5   M   51  10   F   48   5
  F   31   2   F   25   1   M   22   0   M   41   0
  M   19   0   M   23   0   M   31   1   M   17   0
  F   21   0   F   44   7   M   28   0   M   47   3
  M   23   0   F   29   3   F   24   0   M   34   1
  F   19   0   F   35   2   M   39   0   M   43   6
;
```

If you simply look at the data, it appears that many park attendees did not catch any fish. The zero counts are made up of two populations: attendees who do not fish and attendees who fish badly. A zero-inflation mechanism thus appears reasonable since a zero count can be produced by two separate distributions.

The following PROC FMM statements fit a standard Poisson regression model to these data. A common intercept is assumed for men and women, and the regression slope varies with gender.

```
proc fmm data=catch;
   class gender;
   model count = gender*age / dist=Poisson;
run;
```

Figure 4 displays information about the model and data set. The "Model Information" table conveys that the model is a single-component Poisson model (a Poisson GLM) and that parameters are estimated by maximum likelihood.

Figure 4 Model Information and Class Levels in Poisson Regression

Confidence Intervals for the Difference of Proportions

The FMM Procedure

Model Information

Data Set	WORK.CATCH
Response Variable	count
Type of Model	Generalized Linear (GLM)
Distribution	Poisson
Components	1
Link Function	Log
Estimation Method	Maximum Likelihood

Class Level Information

Class	Levels	Values
gender	2	F M

Number of Observations Read	52
Number of Observations Used	52

Figure 5 displays the "Fit Statistics" and "Parameter Estimates" tables from the maximum likelihood estimation. If the model is not overdispersed, the Pearson statistic should roughly equal the number of observations in the data set minus the number of parameters. With $n = 52$, there is evidence of overdispersion in these data with Q_P=85.96.

Figure 5 Fit Results in Poisson Regression

Fit Statistics

-2 Log Likelihood	182.7
AIC (smaller is better)	188.7
AICC (smaller is better)	189.2
BIC (smaller is better)	194.6
Pearson Statistic	85.9573

Parameter Estimates for 'Poisson' Model

Effect	gender	Estimate	Standard Error	z Value	Pr > |z|
Intercept		-3.9811	0.5439	-7.32	<.0001
age*gender	F	0.1278	0.01149	11.12	<.0001
age*gender	M	0.1044	0.01224	8.53	<.0001

Suppose that the cause of overdispersion is zero-inflation of the count data. The following statements fit a zero-inflated Poisson model.

```
proc fmm data=catch;
   class gender;
   model count = gender*age / dist=Poisson ;
   model       +            / dist=Constant;
run;
```

There are two MODEL statements, one for each component of the mixture. The distributions are different for the components, so you cannot specify the mixture model with a single MODEL statement. The first MODEL statement identifies the response variable for the model for COUNT, and it defines a Poisson model with intercept and gender-specific slopes.

The second MODEL statement uses the continuation operator ("+") and adds a model with a degenerate distribution by using the DIST=CONSTANT option. Because the mass of the constant is placed by default at zero, the second MODEL statement adds a zero-inflation component to the model. When you use the "+" sign, you pick up the response variable and effects from the preceding MODEL statement; any additional effects you list after the "+" sign will be added to them.

Figure 6 displays the "Model Information" and "Optimization Information" tables for this run of the FMM procedure. The model is now identified as a zero-inflated Poisson (ZIP) model with two components, and the parameters continue to be estimated by maximum likelihood. The "Optimization Information" table shows that there are four parameters in the optimization (compared to three parameters in the Poisson GLM model). The four parameters correspond to three parameters in the mean function (intercept and two gender-specific slopes) and the mixing probability.

Figure 6 Model and Optimization Information in the ZIP Model

Confidence Intervals for the Difference of Proportions

The FMM Procedure

Model Information

Data Set	WORK.CATCH
Response Variable	count
Type of Model	Zero-inflated Poisson
Components	2
Estimation Method	Maximum Likelihood

Optimization Information

Optimization Technique	Dual Quasi-Newton
Parameters in Optimization	4
Mean Function Parameters	3
Scale Parameters	0
Mixing Prob Parameters	1
Number of Threads	2

Figure 7 displays the results from fitting the ZIP model by maximum likelihood. The -2 log likelihood and the information criteria suggest a much-improved fit over the single-component Poisson model (compare Figure 7 to Figure 5). The Pearson statistic is reduced by a factor of 2 compared to the Poisson model and suggests a better fit than the standard Poisson model.

Figure 7 Maximum Likelihood Results for the ZIP model

Fit Statistics

-2 Log Likelihood	145.6
AIC (smaller is better)	153.6
AICC (smaller is better)	154.5
BIC (smaller is better)	161.4
Pearson Statistic	43.4467
Effective Parameters	4
Effective Components	2

Parameter Estimates for 'Poisson' Model

Component	Effect	gender	Estimate	Standard Error	z Value	Pr > |z|
1	Intercept		-3.5215	0.6448	-5.46	<.0001
1	age*gender	F	0.1216	0.01344	9.04	<.0001
1	age*gender	M	0.1056	0.01394	7.58	<.0001

Figure 7 *continued*

Parameter Estimates for Mixing Probabilities

	Linked Scale				
Effect	Estimate	Standard Error	z Value	Pr > |z|	Probability
Intercept	0.8342	0.4768	1.75	0.0802	0.6972

The number of effective parameters and components shown in Figure 5 equals the values from Figure 6. This is not always the case because components can collapse (for example, when the mixing probability approaches zero or when two components have identical parameter estimates). In this example, both components and all four parameters are identifiable. The Poisson regression and the zero process mix, with a probability of approximately 0.6972 attributed to the Poisson component.

Bayesian Techniques in the FMM Procedure

The FMM procedure enables you to fit some mixture models by Bayesian techniques. The following statements add the BAYES statement to the previous PROC FMM statements:

```
proc fmm data=catch seed=12345;
   class gender;
   model count = gender*age / dist=Poisson;
   model       +            / dist=constant;
   performance cpucount=2;
   bayes;
run;
```

The "Model Information" table indicates that the model parameters are estimated by Markov chain Monte Carlo techniques, and it displays the random number seed (Figure 8). This is useful if you did not specify a seed to identify the seed value that reproduces the current analysis. The "Bayes Information" table provides basic information about the Monte Carlo sampling scheme. The sampling method uses a data augmentation scheme to impute component membership and then the Gamerman (1997) algorithm to sample the component-specific parameters. The 2,000 burn-in samples are followed by 10,000 Monte Carlo samples without thinning.

Figure 8 Model, Bayes, and Prior Information in the ZIP Model

Confidence Intervals for the Difference of Proportions

The FMM Procedure

Model Information

Data Set	WORK.CATCH
Response Variable	count
Type of Model	Zero-inflated Poisson
Components	2
Estimation Method	Markov Chain Monte Carlo
Random Number Seed	12345

Bayes Information

Sampling Algorithm	Gamerman
Data Augmentation	Latent Variable
Initial Values of Chain	ML Estimates
Burn-In Size	2000
MC Sample Size	10000
MC Thinning	1
Parameters in Sampling	4
Mean Function Parameters	3
Scale Parameters	0
Mixing Prob Parameters	1
Number of Threads	2

Prior Distributions

Component	Effect	gender	Distribution	Mean	Variance	Initial Value
1	Intercept		Normal(0, 1000)	0	1000.00	-3.5215
1	age*gender	F	Normal(0, 1000)	0	1000.00	0.1216
1	age*gender	M	Normal(0, 1000)	0	1000.00	0.1056
1	Probability		Dirichlet(1, 1)	0.5000	0.08333	0.6972

The "Prior Distributions" table identifies the prior distributions, their parameters for the sampled quantities, and their initial values. The prior distribution of parameters associated with model effects is a normal distribution with mean 0 and variance 1,000. The prior distribution for the mixing probability is a Dirichlet(1,1), which is identical to a uniform distribution (Figure 8). Since the second mixture component is a degeneracy at zero with no associated parameters, it does not appear in the "Prior Distributions" table in Figure 8.

Figure 9 displays descriptive statistics about the 10,000 posterior samples. Recall from Figure 7 that the maximum likelihood estimates were -3.5215, 0.1216, 0.1056, and 0.6972, respectively. With this choice of prior, the means of the posterior samples are generally close to the MLEs. The "Posterior Intervals" table displays 95% intervals of equal-tail probability and 95% intervals of highest posterior density (HPD) intervals.

Figure 9 Posterior Summaries and Intervals in the ZIP Model

Posterior Summaries

Component	Effect	gender	N	Mean	Standard Deviation
1	Intercept		10000	−3.5524	0.6509
1	age*gender	F	10000	0.1220	0.0136
1	age*gender	M	10000	0.1058	0.0140
1	Probability		10000	0.6938	0.0945

Posterior Summaries

			Percentiles		
Component	Effect	gender	25%	50%	75%
1	Intercept		−3.9922	−3.5359	−3.0875
1	age*gender	F	0.1124	0.1218	0.1314
1	age*gender	M	0.0961	0.1055	0.1153
1	Probability		0.6293	0.6978	0.7605

Posterior Intervals

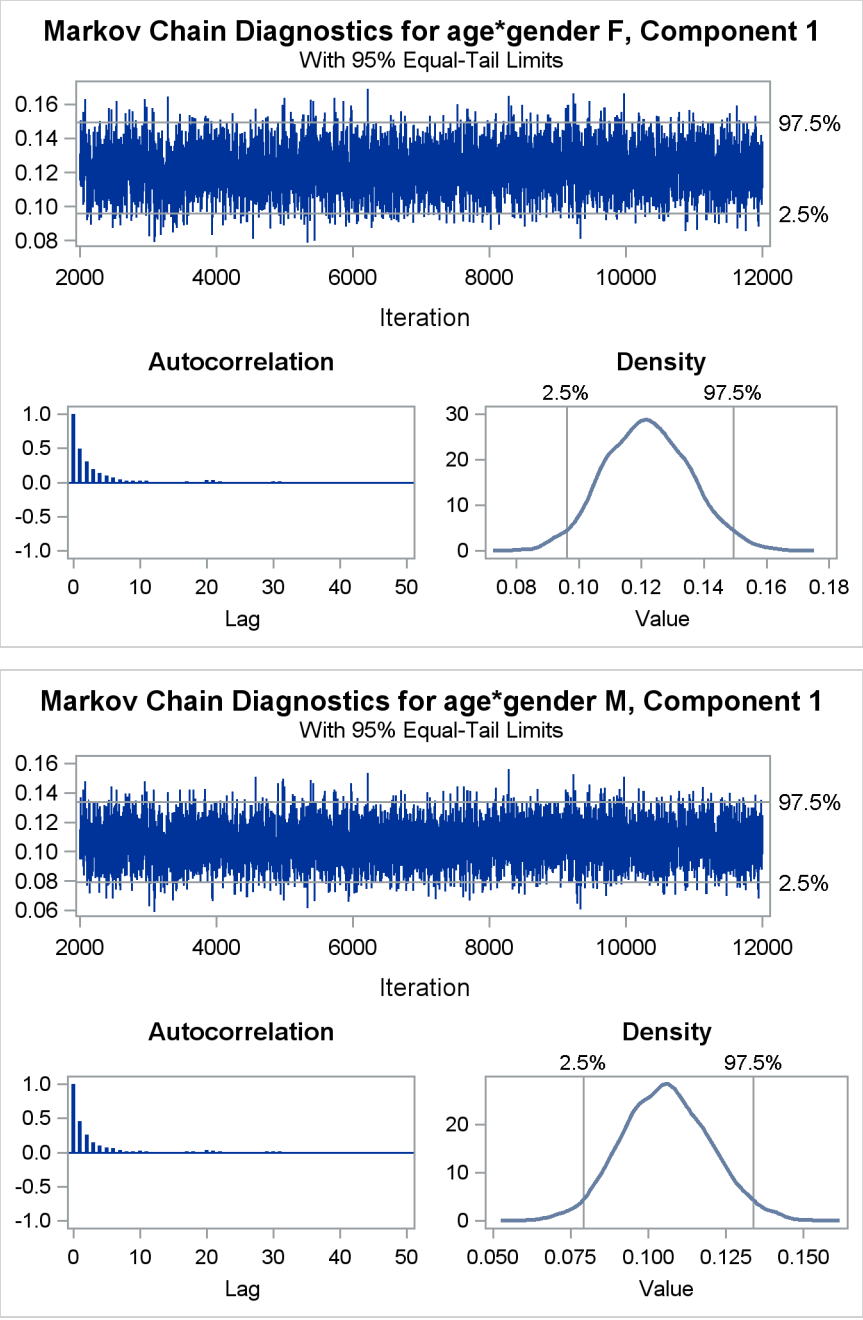
				Equal-Tail Interval		HPD Interval	
Component	Effect	gender	Alpha				
1	Intercept		0.050	−4.8693	−2.3222	−4.8927	−2.3464
1	age*gender	F	0.050	0.0960	0.1494	0.0961	0.1494
1	age*gender	M	0.050	0.0792	0.1339	0.0796	0.1341
1	Probability		0.050	0.5041	0.8688	0.5025	0.8666

You can generate trace plots for the posterior parameter estimates by enabling ODS Graphics:

```
ods graphics on;
ods select TADPanel;
proc fmm data=catch seed=12345;
   class gender;
   model count = gender*age / dist=Poisson;
   model       +            / dist=constant;
   performance cpucount=2;
   bayes;
run;
ods graphics off;
```

A separate trace panel is produced for each sampled parameter, and the panels for the gender-specific slopes are shown in Figure 10. There is good mixing in the chains: the modest autocorrelation diminishes after about 10 successive samples. By default, the FMM procedure transfers the credible intervals for each parameter from the "Posterior Intervals" table to the trace plot and the density plot in the trace panel.

Figure 10 Trace Panels for Gender-Specific Slopes



New Confidence Intervals for the Difference of Proportions

It is often useful to describe the nature of the association in a 2×2 table by examining the difference of proportions with the 'event' response in each of the groups. For example, for the table in Table 2 you would be interested in looking at the difference in the favorable response for the two treatment groups, Test and Control.

Table 2 Respiratory Treatment Outcomes

Treatment	Favorable	Unfavorable	Total
Test	10	2	12
Control	2	4	6
Total	12	6	18

The proportion p_i of favorable events in the ith row is defined as n_{i1}/n_{i+}. When the groups represent simple random samples, and the difference $d = p_1 - p_2$ for the proportions p_1 and p_2,

$$E\{p_1 - p_2\} = \pi_1 - \pi_2$$

The variance is

$$V\{p_1 - p_2\} = \frac{\pi_1(1-\pi_1)}{n_{1+}} + \frac{\pi_2(1-\pi_2)}{n_{2+}}$$

for which a consistent estimate is

$$v_d = \frac{p_1(1-p_1)}{n_{1+}} + \frac{p_2(1-p_2)}{n_{2+}}$$

A $100(1-\alpha)$% confidence interval for $(\pi_1 - \pi_2)$ with continuity correction is written as

$$d \pm \left\{ z_{\alpha/2}\sqrt{v_d} + \frac{1}{2}\left\{ \frac{1}{n_{1+}} + \frac{1}{n_{2+}} \right\} \right\}$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution; this confidence interval is based on Fleiss, Levin, and Paik (2003). These confidence limits include a continuity adjustment to the Wald asymptotic confidence limits that adjust for the difference between the normal approximation and the discrete binomial distribution.

The default Wald asymptotic confidence limits (without the continuity correction) in the FREQ procedure are appropriate for larger sample sizes, say cell counts of at least 12. The continuity-corrected Wald confidence limits are more appropriate for moderate sample sizes, say cell counts of at least 8.

Since the cell counts are relatively small in Table 2, you might consider exact methods for not only assessing association (Fisher's exact test) but also for the construction of the confidence intervals for the difference of proportions. The EXACT statement provides an exact confidence interval based on the raw risk difference. You specify the RISKDIFF option in the EXACT statement to generate it. However, this interval is known to be fairly conservative (wide). New in SAS/STAT 9.3 is a method for computing the exact confidence interval based on the Farrington-Manning score statistic which has better coverage properties.

Also new in SAS/STAT 9.3 are several other methods for constructing the confidence interval for the difference of proportions: Newcombe, Farrington-Manning, and Hauck-Anderson. These methods can be more suitable for small counts than the exact intervals. See Stokes and Koch (2011) for more detail, as well as the PROC FREQ documentation. The following example produces these new confidence intervals with the FREQ procedure.

The following SAS statements create the SAS data set SMALL:

```
data small;
   input treat $ outcome $ count @@ ;
   datalines;
Test Favorable 10 Test Unfavorable 2
Placebo Favorable 2 Placebo Unfavorable 4
;
```

The PROC FREQ statements request the new confidence intervals as suboptions of the RISKDIFF option in the TABLES statement. The EXACT statement, with RISKDIFF=(FMSCORE), specifies the exact confidence interval based on the Farrington-Manning score statistic.

```
proc freq order=data data=small;
   weight count;
   tables treat*outcome / riskdiff(cl=(wald newcombe fm ha exact) correct);
   exact riskdiff(fmscore);
   ods output PdiffCLs=pd;
   ods output RiskDiffCol1=dif(keep=risk row where=(row='Difference'));
run;
```

The ODS OUTPUT statements are needed in order to plot the intervals in a later step.

Figure 11 displays the generated confidence intervals.

Figure 11 Confidence Intervals for Difference of Proportions

Random-Effects Model

The FREQ Procedure

Statistics for Table of treat by outcome

Confidence Limits for the Proportion (Risk) Difference
Column 1 (outcome = Favorabl)
Proportion Difference = 0.5000

Type	95% Confidence Limits	
Exact (FM Score)	-0.0000	0.8433
Farrington-Manning	0.0380	0.9620
Hauck-Anderson	-0.0516	1.0000
Newcombe Score (Corrected)	-0.0352	0.8059
Wald (Corrected)	-0.0571	1.0000

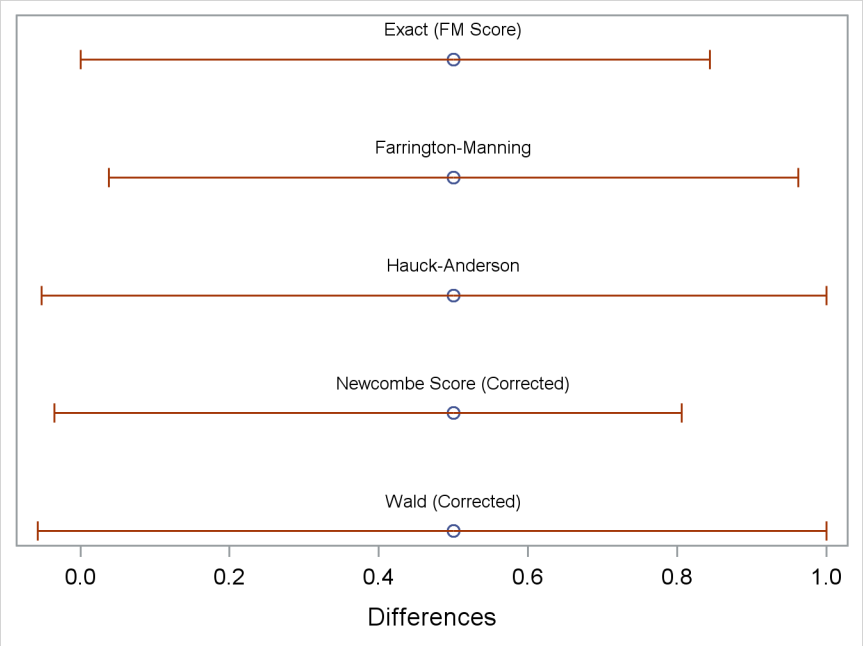
Note that the corrected Wald-based confidence interval is the widest interval. The scored-based exact (unconditional) confidence interval has a relatively wide interval, (−0.0296, 0.8813) as does the interval based on Farrington-Manning scores, (0.0380, 0.9620), which also excludes the value zero. The score-based exact confidence interval has a narrower interval of (0.000, 0.8433), but it comes with a zero boundary. Not shown here, the default exact unconditional interval is (−0.0296, 0.8813), which is fairly wide. The corrected Newcombe interval is the narrowest at (−0.0352, 0.8059), and it might be the most suitable for these data as it does not exclude zero and thus is in harmony with the Fisher exact test result.

By using the SGPLOT procedure, you can easily compare these confidence intervals graphically.

```
data risk;
   set pd;
   if _n_ = 1 then set dif(drop=row);
   y = -_n_;
   l = y + 0.25;
run;
ods graphics on;
proc sgplot data=risk noautolegend;
   title 'Confidence Intervals for the Difference of Proportions';
   scatter x = risk y = y / xerrorlower=lowercl xerrorupper=uppercl;
   scatter x = risk y = l / markerchar=type;
   yaxis display=none;
   xaxis offsetmin=0.025 offsetmax=0.025;
run;
ods  graphics off;
```

Figure 12 displays the resulting graph.

Figure 12 Confidence Intervals for Difference of Proportions width



Frailty Models in Survival Analysis

When experimental units are clustered, the failure times of those units within a cluster tend to be correlated. One approach to managing this correlation when you want to apply Cox regression is to estimate the regression parameters by the maximum partial likelihood estimates, assuming an independent working assumption and accounting for the correlation with a robust sandwich covariance matrix estimate (Lee, Wei, and Amato 1992).

Another approach is to account for within-cluster correlation by using a shared frailty mode in which the cluster effects are incorporated in the model as normally distributed random variables.

The hazard rate for the jth individual in the ith cluster is

$$\lambda_{ij}(t) = \lambda_0(t)e^{\boldsymbol{\beta}'\mathbf{Z}_{ij}(t)+\gamma_i}$$

where $\lambda_0(t)$ is an arbitrary baseline hazard rate, $\mathbf{Z}_{ij}$ is the vector of (fixed-effect) covariates, $\boldsymbol{\beta}$ is the vector of regression coefficients, and γ_i is the random effect for cluster i. The random components $\gamma_1, \ldots, \gamma_s$ are assumed to be independent and identically distributed as a normal random variable with mean 0 and an unknown variance θ.

In terms of the frailties $u_1, \ldots, u_s$, given by $\gamma_i = \log(u_i)$, the frailty model can be written as

$$\lambda_{ij}(t) = \lambda_0(t)u_i e^{\boldsymbol{\beta}'\mathbf{Z}_{ij}(t)}$$

Each frailty has a lognormal distribution with median 1. This gives the interpretation that individuals in cluster i with $u_i > 1 (u_i < 1)$ tend to fail at a faster (slower) rate than that under an independence model. The PHREG procedure uses a penalized partial likelihood approach to fit frailty models.

The following DATA step creates the data set BLIND that represents 197 diabetic patients who have a high risk of experiencing blindness in both eyes. One eye of each patient is treated with laser photocoagulation, and the other eye is treated with standard remedies. The hypothesis of interest is whether the laser treatment delays the occurrence of blindness. Since juvenile and adult diabetes have very different courses, it is also desirable to examine how the age of onset of diabetes might affect the time of blindness. Since there are no biological differences between the left eye and the right eye, it is natural to assume a common baseline hazard function for the failure times of the left and right eyes.

Each patient is a cluster that contributes two observations to the input data set, one for each eye. The following variables are in the input data set BLIND:

- ID, patient's identification
- Time, time to blindness

- Status, blindness indicator (0:censored and 1:blind)
- Treat, treatment received (Laser or Others)
- Type, type of diabetes (Juvenile: onset at age $\leq$ 20 or Adult: onset at age $>$ 20)

The following SAS statements create the data set BLIND:

```
proc format;
   value type 0='Juvenile' 1='Adult';
   value Rx  1='Laser' 0='Others';
run;
data Blind;
input ID Time Status dty trt @@;
Type= put(dty, type.);
Treat= put(trt, Rx.);
datalines;
  5 46.23 0 1 1    5 46.23 0 1 0   14 42.50 0 0 1   14 31.30 1 0 0
 16 42.27 0 0 1   16 42.27 0 0 0   25 20.60 0 0 1   25 20.60 0 0 0
 29 38.77 0 0 1   29  0.30 1 0 0   46 65.23 0 0 1   46 54.27 1 0 0
 49 63.50 0 0 1   49 10.80 1 0 0   56 23.17 0 0 1   56 23.17 0 0 0
 61  1.47 0 0 1   61  1.47 0 0 0   71 58.07 0 1 1   71 13.83 1 1 0
100 46.43 1 1 1  100 48.53 0 1 0  112 44.40 0 1 1  112  7.90 1 1 0
...
```

The next PROC PHREG statements fit a shared frailty model to the BLIND data set. The RANDOM statement specifies the analysis, and it identifies ID as the clustering variable. The variable ID must also be listed in the CLASS statement.

```
proc phreg data=Blind;
   class ID Treat Type;
   model Time*Status(0)=Treat|Type;
   random ID;
   hazardratio 'Frailty Model Analysis' Treat;
run;
```

Figure 13 shows a tabulation of the data. 54 eyes treated with laser photocoagulation developed blindess, as did 101 eyes treated with other remedies.

Figure 13 Crosstabulations

Confidence Intervals for the Difference of Proportions

The FREQ Procedure

Table of Treat by Status

Treat	Status		
Frequency	0	1	Total
Laser	143	54	197
Others	96	101	197
Total	239	155	394

Figure 14 displays the estimate and asymptotic estimated standard error of the common variance parameter of the normal random effects.

Figure 14 Variance Estimate of the Normal Random Effects

Confidence Intervals for the Difference of Proportions

The PHREG Procedure

Covariance Parameter Estimates

Cov Parm	REML Estimate	Standard Error
ID	0.8308	0.2145

Table 15 displays the Wald tests for both the fixed effects and the random effects. The random effects are statistically significant (p=0.0042). Laser photocoagulation appears to be effective (p=0.0252) in delaying the occurrence of blindness, although there is also a significant treatment by diabetes type interaction effect (p=0.0071).

Figure 15 Type 3 Tests

Type 3 Tests

Effect	Wald Chi-Square	DF	Pr > ChiSq	Adjusted DF	Adjusted Pr > ChiSq
Treat	4.8964	1	0.0269	0.9587	0.0252
Type	2.6386	1	0.1043	0.6795	0.0629
Treat*Type	7.1336	1	0.0076	0.9644	0.0071
ID	110.3916	.	.	74.2776	0.0042

Figure 16 contains the maximum likelihood parameter estimates.

Figure 16 Parameter Estimates

Analysis of Maximum Likelihood Estimates

Parameter			DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Treat	Laser		1	−0.49849	0.22528	4.8964	0.0269
Type	Adult		1	0.39781	0.24490	2.6386	0.1043
Treat*Type	Laser	Adult	1	−0.96530	0.36142	7.1336	0.0076

Analysis of Maximum Likelihood Estimates

Parameter			Hazard Ratio	Label
Treat	Laser		.	Treat Laser
Type	Adult		.	Type Adult
Treat*Type	Laser	Adult	.	Treat Laser * Type Adult

Figure 17 displays the hazard ratio estimates. They show that laser photocoagulation is effective in delaying blindness for both types of diabetes, and it is more effective for the adult-onset diabetes than for juvenile-onset diabetes.

Figure 17 Hazard Ratio Estimates for Frailty Model

Frailty Model Analysis: Hazard Ratios for Treat

Description	Point Estimate	95% Wald Confidence Limits	
Treat Laser vs Others At Type=Adult	0.231	0.133	0.403
Treat Laser vs Others At Type=Juvenile	0.607	0.391	0.945

These results are similar to those found in a marginal model analysis, which are described in the PHREG documentation.

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 PROC GENMOD. 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 RANDOM statement in this release. It simplifies the construction of hierarchical random-effects models and significantly reduces simulation time while improving convergence, especially in models

with a large number of subjects or clusters. This statement defines random effects that can enter the model in a linear or nonlinear fashion and supports univariate and multivariate prior distributions.

Consider a scenario in which data are collected in groups and you want to model group-specific effects. You can use a random-effects model (sometimes also known as a variance-components model):

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_i + e_{ij}, \quad e_{ij} \sim \text{normal}(0, \sigma^2)$$

where $i = 1, 2, \cdots, I$ is the group index and $j = 1, 2, \cdots, n_i$ indexes the observations in the ith group. In the regression model, the fixed effects β_0 and β_1 are the intercept and the coefficient for variable x_{ij}, respectively. The random effect γ_i is the mean for the ith group, and e_{ij} are the error term.

Consider the following SAS data set:

```
title 'Random-Effects Model';
data heights;
   input Family Gender$ Height @@;
   datalines;
1 F 67   1 F 66   1 F 64   1 M 71   1 M 72   2 F 63
2 F 63   2 F 67   2 M 69   2 M 68   2 M 70   3 F 63
3 M 64   4 F 67   4 F 66   4 M 67   4 M 67   4 M 69
;
```

Height, gender, and family are recorded for 18 individuals. Since GENDER is a character variable and PROC MCMC does not support a CLASS statement, you need to create the corresponding design matrix columns. The next DATA step creates a dummary variable IGENDER. In this example, the design matrix for a factor variable of level 2 (M and F) can be constructed using the following statement:

```
data iheights;
   set heights;
   IGender=('F');
run;
```

To model the data, you can assume that HEIGHT is normally distributed:

$$y_{ij} \sim \text{normal}(\mu_{ij}, \sigma^2), \quad \mu_{ij} = \beta_0 + \beta_1 \text{gf}_{ij} + \gamma_i$$

The priors on the parameters β_0, β_1, γ_i are also assumed to be normal:

$$\begin{aligned} \beta_0 &\sim \text{normal}(0, \text{var} = 1e5) \\ \beta_1 &\sim \text{normal}(0, \text{var} = 1e5) \\ \gamma_i &\sim \text{normal}(0, \text{var} = \sigma_\gamma^2) \end{aligned}$$

Priors on the variance terms, σ^2 and σ_γ^2, are inverse-gamma:

$$\begin{aligned} \sigma^2 &\sim \text{igamma}(\text{shape} = 0.01, \text{scale} = 0.01) \\ \sigma_\gamma^2 &\sim \text{igamma}(\text{shape} = 0.01, \text{scale} = 0.01) \end{aligned}$$

The inverse-gamma distribution is a conjugate prior for the variance in the normal likelihood and the variance in the prior distribution of the random effect.

The following statements fit a linear random-effects model to the data and produce the output shown in Figure 18:

```
ods graphics on;
proc mcmc data=iheights outpost=postout nmc=50000 thin=5 seed=7893;
   ods select Parameters REparameters PostSummaries PostIntervals
      tadpanel;
   parms b0 0 b1 0 s2 1 s2g 1;
   prior b: ~ normal(0, var = 10000);
   prior s: ~ igamma(0.01, scale = 0.01);
   random gamma ~ normal(0, var = s2g) subject=family monitor=(gamma);
   mu = b0 + b1 * gf + gamma;
   model height ~ normal(mu, var = s2);
run;
ods graphics off;
```

The PROC MCMC statement specifies the input and output data sets, the simulation size, the thinning rate, and a random number seed. The ODS SELECT statement displays the model parameter and random-effects parameter information tables, summary statistics table, the interval statistics table, and the diagnostics plots.

The PARMS statement lumps all four model parameters in a single block. They are B0 (overall intercept), B1 (main effect for IGENDER), S2 (variance of the likelihood function), and §2G (variance of the random effect). If a random walk Metropolis sampler is the only applicable sampler for all parameters, then these four parameters are updated in a single block. However, since the conjugate updater is used to draw posterior samples of S2 and S2G, PROC MCMC updates these parameters separately (see the BLOCK column in the "Parameters" table in Figure 18).

The PRIOR statements specify priors for all the parameters. The notation b: is a shorthand for all symbols that start with the letter 'b'. In this example, b: includes B0 and B1. Similarly, S: stands for both S2 and S2G.

The RANDOM statement specifies a single random effect to be GAMMA, and it specifies that it has a normal prior centered at 0 with variance S2G. The SUBJECT= argument in the RANDOM statement defines a group index FAMILY in the model, where all observations from the same family should have the same group indicator value. The MONITOR= option outputs analysis for all of the random-effects parameters.

Finally, the MU assignment statement calculates the expected value of HEIGHT in the random-effects model. The MODEL statement specifies the likelihood function for HEIGHT.

The "Parameters" and "Random-Effects Parameters" tables, shown in Figure 18, contain information about the model parameters and the four random-effects parameters.

Figure 18 Model and Random-Effects Parameter Information

Random-Effects Model

The MCMC Procedure

Parameters

Block	Parameter	Sampling Method	Initial Value	Prior Distribution
1	s2	Conjugate	1.0000	igamma(0.01, scale = 0.01)
2	s2g	Conjugate	1.0000	igamma(0.01, scale = 0.01)
3	b0	N-Metropolis	0	normal(0, var = 10000)
	b1		0	normal(0, var = 10000)

Random Effects Parameters

Parameter	Subject	Levels	Prior Distribution
gamma	family	4	normal(0, var = s2g)

The posterior summary and interval statistics for B0 and B1 are shown in Figure 19 and Figure 20:

Figure 19 Posterior Summary Statistics

Random-Effects Model

The MCMC Procedure

Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
b0	10000	68.4591	1.2085	67.7939	68.4369	69.0863
b1	10000	-3.5381	0.9622	-4.1439	-3.5351	-2.9466
s2	10000	4.1495	1.9089	2.8251	3.7353	4.9854
s2g	10000	4.8368	17.4110	0.2218	1.2534	4.0669
gamma_1	10000	0.9374	1.2817	0.0832	0.6802	1.6250
gamma_2	10000	0.0167	1.1399	-0.4145	0.0325	0.5038
gamma_3	10000	-1.3313	1.6080	-2.1514	-0.9247	-0.1434
gamma_4	10000	0.0979	1.1495	-0.3470	0.0537	0.5802

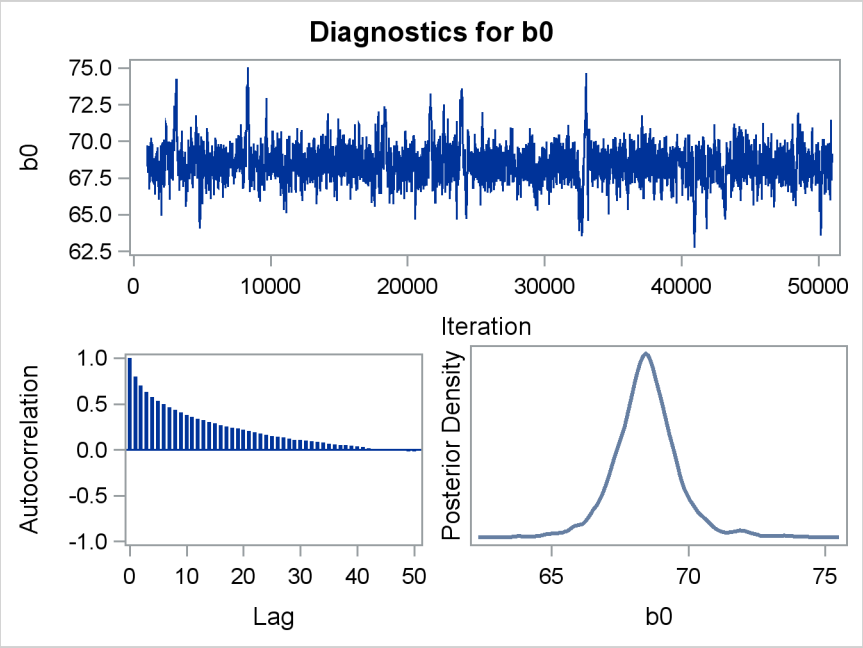
Figure 20 Posterior Summary Statistics

Posterior Intervals

Parameter	Alpha	Equal-Tail Interval		HPD Interval	
b0	0.050	66.0454	71.1125	65.8787	70.7985
b1	0.050	−5.4303	−1.5336	−5.4941	−1.6259
s2	0.050	1.7532	9.0102	1.4424	8.0066
s2g	0.050	0.0117	29.7402	0.00121	18.3336
gamma_1	0.050	−1.1857	3.8472	−1.1522	3.8666
gamma_2	0.050	−2.6485	2.4385	−2.3792	2.6240
gamma_3	0.050	−5.4020	0.6187	−4.8669	0.8624
gamma_4	0.050	−2.5324	2.6004	−2.2341	2.7602

Trace plots, autocorrelation plots, and posterior density plots for all the parameters are shown in Figure 20. The mixing looks very reasonable, suggesting convergence.

Figure 21 Plots for b_1 and Log of the Posterior Density



From the interval statistics table, you see that both the equal-tail and HPD intervals for B0 are positive, strongly indicating the positive effect of the parameter. On the other hand, both intervals for B1 cover the value zero, indicating that IGENDER does not have a strong impact on predicting HEIGHT in this model.

In addition to the default Metropolis-based algorithms, PROC MCMC now takes advantages of certain forms of conjugacy in the model in order to sample directly from the target conditional distributions. In many situations, the conjugate sampler increases sampling efficiency and provides a substantial reduction in computing time.

The MCMC procedure now supports multivariate distributions including the Dirichlet, inverse Wishart, multivariate normal, and multinomial distributions. You can find more information in the SAS Global Forum papers "Bayesian Modeling Using the MCMC Procedure" and "The RANDOM Statement and More: PROC MCMC Moves On" by Fang Chen (2009, 2011).

More Highlights of SAS/STAT 9.3

A number of other important enhancements are available in this release as well.

The EFFECT statement

The EFFECT statement is now production. This statement is available in the HPMIXED, GLIMMIX, GLMSELECT, LOGISTIC, ORTHOREG, PHREG, PLS, QUANTREG, ROBUSTREG, SURVEYLOGISTIC, and SURVEYREG proce-

dures. The NATURALCUBIC option specifies a natural cubic spline basis for the spline expansion.

CALIS Procedure

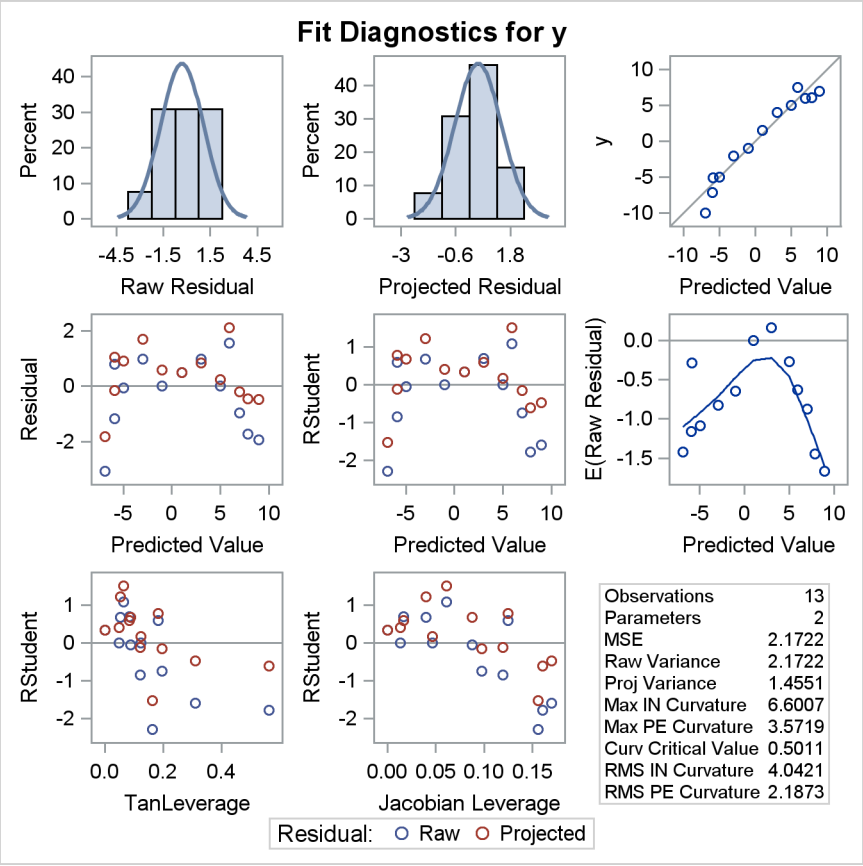
The following features are now production: METHOD=FIML, mean structure analysis with the COSAN model, extended PATH modeling language that supports the specification of variances or covariances as paths, unnamed free parameter specification in all model, and improved improved RAM model specification.

In addition, PROC CALIS now provides detailed analysis of the missing patterns with the FIML estimation method. With the COVPATTERN= and MEANPATTERN= options, you can specify various standard mean and covariance patterns by using keywords. PROC CALIS then generates the required covariance and mean structures automatically.

NLIN Procedure

The NLIN procedure provides several experimental features for diagnosing the nonlinear model fit, including the PLOTS, NLINMEASURES, and BIAS options in the PROC NLIN statement. The NLINMEASURES displays global measures of nonlinearity, and the BIAS option computes Box's bias statistics for the parameter estimates. The PLOTS option enables you to plot the fitted model, fit diagnostics, tangential and Jacobian leverage, and local influence. You can add the leverage, local influence, and residual diagnostics to the output data set.

Figure 22 Diagnostics Panel from NLIN Procedure



HPMIXED Procedure

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. The HPMIXED procedure became production with SAS/STAT 9.22.

PROC HPMIXED provides the REPEATED statement, which defines the repeated effect and the residual covariance structure in the mixed model. The AR(1), CS, CSH, UC, UCH, and UN covariance structures are now available with the TYPE= option in the RANDOM statement.

MI Procedure

The experimental FCS statement specifies a multivariate imputation by fully conditional specification (FCS) methods. For data with an arbitrary missing data pattern, these methods enable you to impute missing values for all variables, assuming that a joint distribution for these variables exists. The FCS method requires fewer iterations than the MCMC method.

QUANTREG Procedure

The new QINTERACT option in the TEST statement enables you to test whether any difference exists among the coefficients across quantiles if several quantiles are specified in the MODEL statement. The RANKSCORE option in the TEST statement now supports the tau score function, which is appropriate for non-iid error models.

For Further Information

A good place to start for further information is the "What's New in SAS/STAT 9.3" chapter in the online documentation when it becomes available.In addition, the Statistics and Operations Focus Area includes substantial 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, complete information is available in the online documentation located here: `support.sas.com/documentation/onlinedoc/stat/`.

References

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

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

Chen, F. (2011), "The RANDOM Statement and More: Moving on with PROC MCMC," in *Proceedings of the SAS Global Forum 2011 Conference,* Cary NC: SAS Institute Inc.

Gamerman, D. (1997), "Sampling from the Posterior Distribution in Generalized Linear Mixed Models," *Statistics and Computing*, 7, 57–68.

Lee, E. W., Wei, L. J., and Amato, D. (1992), "Cox-Type Regression Analysis for Large Numbers of Small Groups of Correlated Failure Time Observations," In Klein J.P., Goel, P.K., eds. *Survival Analysis: State of the Art* 237-247, Netherlands: Kluwer Academic.

Morel, J. G., and Nagaraj, N. K. (1993), "A Finite Mixture Distribution for Modelling Multinomial Extra Variation,"*Biometrika*, 80, 363-371.

Morel, J. G., and Neerchal, N. K. (1997), "Clustered Binary Logistic Regression in Teratology Data Using a Finite Mixture Distribution,"*Statistics in Medicine*, 16, 2843-2853.

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`

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.

Stokes, M. and Koch G. (2011), "Up To Speed With Categorical Data Analysis," in *Proceedings of the SAS Global Forum 2011 Conference,* Cary NC: SAS Institute Inc.

Stokes, M., Rodriguez, R. and Cohen, R. (2010), "SAS/STAT 9.22: The Next Generation," in *Proceedings of the SAS Global Forum 2011 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.

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.

Yung, Y. and Zhang, W. (2011), “Making Use of Incomplete Observations in the Analysis of Structural Equation Models: The CALIS Procedures Full Information Maximum Methods in SAS/STAT® 9.3,” in *Proceedings of the SAS Global Forum 2011 Conference,* Cary NC: SAS Institute Inc.

ACKNOWLEDGMENTS

The authors are grateful to Warren Kuhfeld for his graphics contributions. They are also grateful to Tim Arnold for his documentation programming, as well as Jennifer Waller and Rachael Biel for their support of this project.

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