Paper SAS181-2014

SAS/STAT® 13.1 Round-Up

Robert N. Rodriguez and Maura Stokes, SAS Institute Inc.

Abstract

SAS/STAT® 13.1 brings valuable new techniques to all sectors of the audience for SAS statistical software. Updates for survival analysis include nonparametric methods for interval censoring and models for competing risks. Multiple imputation methods are extended with the addition of sensitivity analysis. Bayesian discrete choice models offer a modern approach for consumer research. Path diagrams are a welcome addition to structural equation modeling, and item response models are available for educational assessment.

This paper provides overviews and introductory examples for each of the new focus areas in SAS/STAT 13.1. The paper also provides a sneak preview of the follow-up release, SAS/STAT 13.2, which brings additional strategies for missing data analysis and other important updates to statistical customers.

Introduction

The goal of SAS/STAT developers has always been to provide SAS customers with powerful, versatile statistical methods. Driving the requirements for new methods are the increasing value, complexity, and size of data, coupled with advances in methodology and computing.

To meet these needs, SAS has accelerated the pace of SAS/STAT releases. These releases have their own numbering scheme because they occur more frequently than new versions of Base SAS®. SAS/STAT 13.1, the current production release, was delivered in December 2013. SAS/STAT 13.2 is planned for release during the summer of 2014.

The new numbering began with SAS/STAT 12.1, which was released in August 2012 for SAS 9.3. SAS/STAT 12.3 was released in July 2013 with the introduction of SAS 9.4. Highlights of these two releases are described by Stokes et al. (2012) and Stokes (2013).

This paper is an overview of SAS/STAT 13.1, with sections for the following major enhancements:

- nonparametric survival analysis of interval-censored data
- analysis of competing risks
- analysis of sensitivity to the MAR (missing at random) assumption in multiple imputation
- Bayesian analysis of discrete choice models
- path diagrams for visualizing structural equation models
- item response models
- bootstrap estimation for nonlinear regression models
- high-performance statistical capability, including effect selection for generalized linear models (added in SAS/STAT 12.3)

SAS/STAT 13.1 provides many enhancements that are not discussed here, including the following:

- the Tweedie distribution in generalized linear models
- the elastic net method for selecting effects in general linear models
- score confidence limits for the odds ratio
- Mantel-Haenszel and summary score estimates of the common risk (proportion) difference
- mid p-values for exact tests
- performance improvements in Markov chain Monte Carlo simulation
- power computations for PROC GLM–type MANOVA and repeated measurements analysis (Castelloe 2014)
- domain quantile estimates computed from a survey sample

Sections of this paper also preview two of the major enhancements in SAS/STAT 13.2: analysis of longitudinal data that have missing values and analysis of weighted multilevel models.

Survival Analysis of Interval-Censored Data with the ICLIFETEST Procedure

Time-to-event data are common in clinical trials and medical studies, where patients are followed up until a predefined event is observed. Event times are right-censored when events are not observed because patients drop out of the study or the study terminates. Event times are interval-censored when patients are assessed infrequently and events are only known to have occurred between two consecutive assessment times.

Standard methods for survival analysis of right-censored data are not adequate for interval-censored data and can produce biased results. However, specialized methods for interval-censored data that offer valid counterparts to standard methods are increasingly available. Turnbull (1976) provided a self-consistent algorithm for estimating the survival function that is equivalent to the Kaplan-Meier estimator under right-censoring, and Finkelstein (1986) generalized the log-rank test to interval-censored data. These pioneering methods have been made feasible in practice through the introduction of computationally intensive approaches such as multiple imputation (Zhao and Sun 2004; Huang, Lee, and Yu 2008), permutation methods (Fay and Shih 2012), and the bootstrap (Sun 2001).

The ICLIFETEST procedure implements state-of-the-art nonparametric statistical methods for analyzing interval-censored data. By default, the procedure uses the efficient EMICM algorithm (Wellner and Zhan 1997) to estimate the survival function. It supports a variety of weight functions (Fay 1999) to perform the generalized log-rank test for equality of survival functions. To the extent that specialized methods are available for interval-censored data, the ICLIFETEST procedure addresses many of the same issues that the LIFETEST procedure covers for right-censored data. For a comprehensive introduction to the ICLIFETEST procedure, see Guo, So, and Johnston (2014).

Introductory Example

The breast cancer data set of Finkelstein and Wolfe (1985) is an example of interval-censored data. Data were collected on 94 patients in a study that compared the risks of cosmetic breast deterioration after tumorectomy. Participants received one of two treatments: radiation therapy (RT) or radiation plus chemotherapy (RCT). Deterioration times were not directly observed; they were interval-censored because deterioration events were only known to have occurred between visits that patients made to a clinic. For 38 patients, the times were right-censored because they did not experience an event prior to the end of the study.

The following statements save the observations in a data set named BCS:

```
data BCS;
   input lTime rTime @@;
   if _N_<=46 then trt = 'RT    ';
   else trt = 'RT+RCT';
   datalines;
45  .  25 37  37  .
 6 10  46  .   0  5

   ... more lines ...

44 48  22 32  11 20
14 17  10 35  48  .
;
```

The next statements use the ICLIFETEST procedure to estimate the survival functions for both treatment groups and to perform the generalized log-rank test:

```
ods graphics on;
proc iclifetest data=BCS impute(seed=123);
   time (lTime, rTime);
   test trt;
run;
```

Figure 1 displays the nonparametric survival estimates for the radiation group (**trt**=RT). A complication of interval censoring is that estimates of failure probability and survival probability are available only for a set of nonoverlapping intervals and are constant within each interval. This is evident in Figure 2, which plots the survival probabilities (dashed lines are plotted across intervals for which the estimates are not defined).

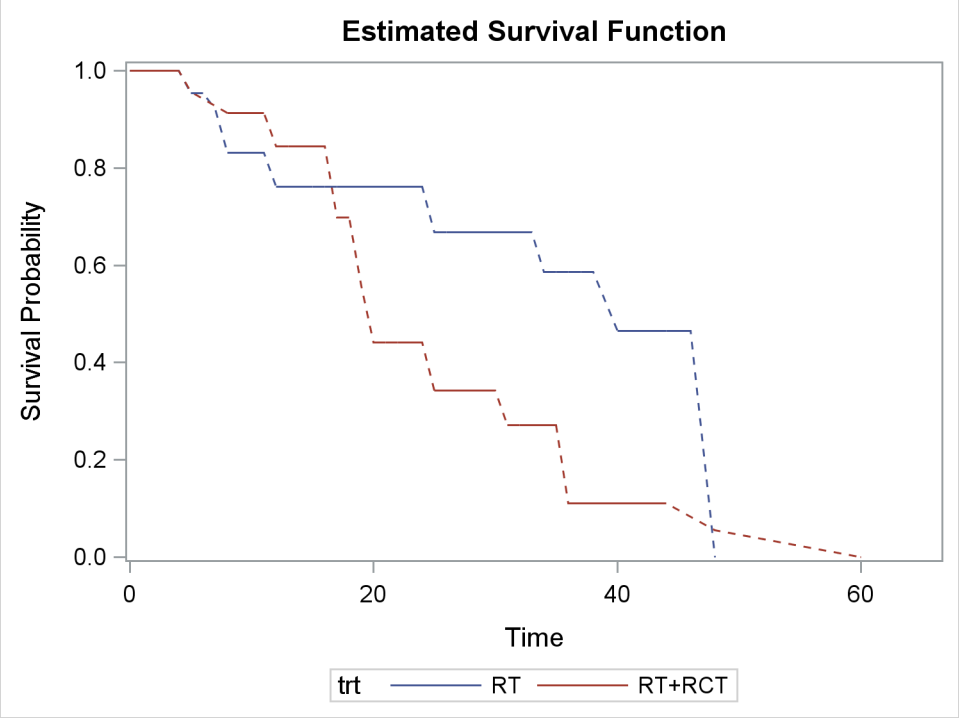
Figure 1 Nonparametric Survival Estimates

Group 1: trt = RT

Nonparametric Survival Estimates

Time Interval		Probability Estimate Failure	Survival	Imputation Standard Error
0	4	0.0000	1.0000	0.0000
5	6	0.0463	0.9537	0.0356
7	7	0.0797	0.9203	0.0457
8	11	0.1684	0.8316	0.0584
12	24	0.2391	0.7609	0.0629
25	33	0.3318	0.6682	0.0706
34	38	0.4136	0.5864	0.0740
40	46	0.5344	0.4656	0.0759
48	Inf	1.0000	0.0000	0.0000

Figure 2 Nonparametric Survival Estimates by Treatment



The group that received radiation alone tended to survive longer before experiencing deterioration than the group that received both treatments. This is confirmed by the generalized log-rank test presented in Figure 3.

Figure 3 Generalized Log-Rank Test

Test of Equality over Group

Weight	Chi-Square	DF	Pr > Chi-Square
SUN	7.1907	1	0.0073

Analysis of Competing Risks with the PHREG Procedure

Standard survival analysis methods apply to an event of interest that is either known to have occurred or is censored. However, in many situations, such as clinical cancer studies, a patient is subject to more than one failure event, and the occurrence of one event impedes the occurrence of others. For example, a leukemia patient's relapse might be unobservable because the patient dies before relapse is diagnosed.

In the presence of competing risks, the usual product-limit estimate for the time-to-event distribution function is biased because you can no longer assume that a subject will experience the event of interest if the follow-up period is long enough. Instead, the analysis often focuses on inference about the cumulative incidence function (CIF), which is the marginal failure subdistribution of a particular cause.

To model the cumulative incidence, Fine and Gray (1999) defined the subdistribution hazard as the hazard of the CIF and imposed a proportional hazards assumption on the subdistribution hazards. This model is referred to as the proportional subdistribution hazards (PSH) model.

You can use PROC PHREG to fit the PSH model by specifying the EVENTCODE= option in the MODEL statement to indicate the event of interest. Maximum likelihood estimates of the regression coefficients are obtained by the Newton-Raphson algorithm, and the covariance matrix of the parameter estimator is computed as a sandwich estimate. For details, see the chapter "The PHREG Procedure" in *SAS/STAT User's Guide*.

Introductory Example

Bone marrow transplant (BMT) is a standard treatment for acute leukemia. Klein and Moeschberger (1997) present a set of BMT data for 137 patients, who were grouped into three risk categories based on their status at the time of transplantation: acute lymphoblastic leukemia (ALL), acute myelocytic leukemia (AML) low-risk, and AML high-risk. During the follow-up period, some patients experienced a relapse, whereas others died while in remission.

In this example, the PSH model is used to compare disease-free survival for the risk categories in the BMT data. Here, relapse is the event of interest and death is a competing risk.

The following statements provide the data:

```
proc format;
   value DiseaseGroup 1='ALL'
                      2='AML-Low Risk'
                      3='AML-High Risk';
data Bmt;
   input Disease T Status @@;
   label T='Disease-Free Survival in Days';
   format Disease DiseaseGroup.;
   datalines;
1   2081   0   1   1602   0   1   1496   0   1   1462   0   1   1433   0
1   1377   0   1   1330   0   1    996   0   1    226   0   1   1199   0

   ... more lines ...

3    625   1   3     48   1   3    273   1   3     63   2   3     76   1
3    113   1   3    363   2
;
```

The variable **Disease** designates the risk group of a patient, which is either ALL, AML–Low Risk, or AML–High Risk. The variable **T** is the disease-free survival time in days, which is either the time to censoring, the time to relapse, or the time to death. The indicator variable **Status** has three values: 0 for censored times, 1 for relapsed patients, and 2 for patients who died before experiencing a relapse.

The next statements create a data set that provides the PHREG procedure with covariate patterns for creating CIF plots:

```
data Risk;
   Disease=1; output;
   Disease=2; output;
   Disease=3; output;
   format Disease DiseaseGroup.;
run;
```

The following statements use the PHREG procedure to fit the PSH model:

```
proc phreg data=Bmt plots(overlay=stratum)=cif;
   class Disease (order=internal ref=first);
   model T*Status(0)=Disease / eventcode=1;
   hazardratio 'Pairwise' Disease / diff=pairwise;
   baseline covariates=Risk;
run;
```

To designate relapse (Status=1) as the event of interest, you specify EVENTCODE=1 in the MODEL statement. The HAZARDRATIO statement requests subdistribution hazard ratios for pairs of disease groups. The COVARIATES= option specifies a data set that contains covariate settings for predicting cumulative incidence functions or for displaying cumulative incidence curves, which are requested here with the PLOTS=CIF option.

Figure 4 shows that of the events for the 137 transplant patients, 42 were relapses (the event of interest), 41 were deaths without relapse, and 54 events were censored.

Figure 4 Distribution of Events and Censored Observations

Summary of Failure Outcomes

Total	Event of Interest	Competing Event	Censored
137	42	41	54

Hazard ratio estimates for one disease group relative to another are displayed in Figure 5.

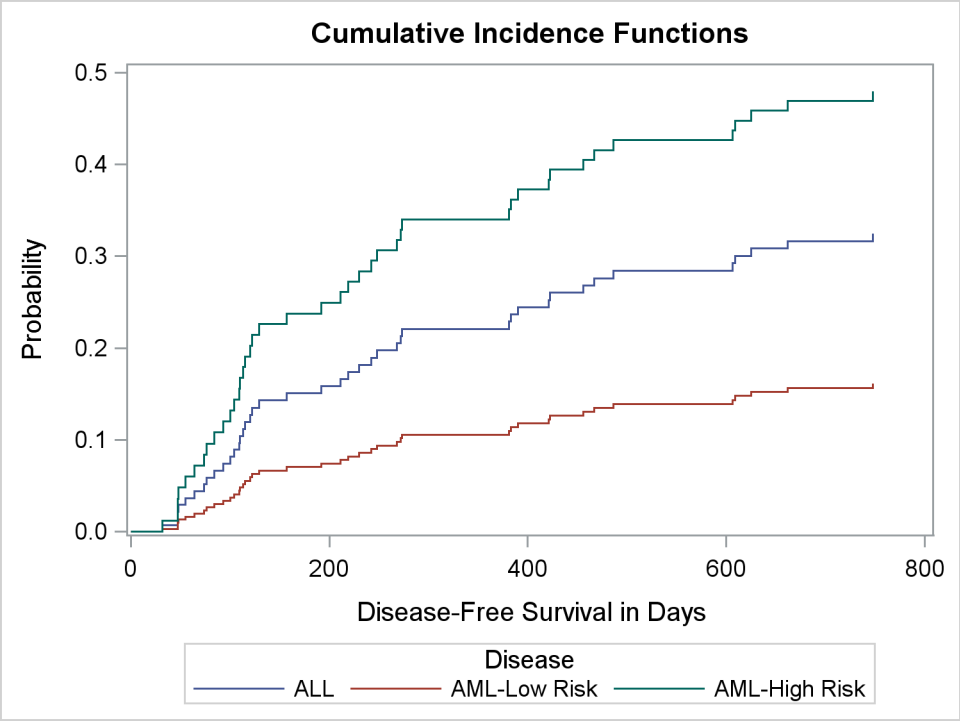
Figure 5 Pairwise Comparisons of Disease Groups

Pairwise: Hazard Ratios for Disease

Description	Point Estimate	95% Wald Confidence Limits	
Disease ALL vs AML-Low Risk	2.233	0.964	5.171
Disease AML-Low Risk vs ALL	0.448	0.193	1.037
Disease ALL vs AML-High Risk	0.601	0.293	1.233
Disease AML-High Risk vs ALL	1.663	0.811	3.408
Disease AML-Low Risk vs AML-High Risk	0.269	0.127	0.573
Disease AML-High Risk vs AML-Low Risk	3.713	1.745	7.900

The hazard of relapse for the ALL patients is 2.2 times that for the AML low-risk patients, and the hazard for the AML high-risk patients is 1.7 times that for the ALL patients. It is expected that at any particular time after the transplant, an AML high-risk patient is more likely to relapse than an ALL patient, and an ALL patient is more likely to relapse than an AML low-risk patient. This ordering is evident in the cumulative incidence functions plotted in Figure 6.

Figure 6 CIFs for Disease Groups



Sensitivity Analysis in Multiple Imputation with the MI Procedure

Missing values are a problem in many statistical analyses. Excluding observations that have missing values ignores the possibility of systematic differences between complete cases and incomplete cases. So the inference you make might not apply to the entire population, especially if the number of complete cases is small.

Multiple imputation is a strategy for making valid inferences; it replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. The MI procedure creates multiply imputed data sets for incomplete p-dimensional multivariate data. You can individually analyze these data sets by using a complete-case method and then using the MIANALYZE procedure to combine the results and obtain appropriate p-values and confidence intervals.

Multiple imputation usually assumes that the data are missing at random (MAR). That is, for a variable **Y**, the probability that a value is missing depends only on the observed values of other variables, not on the unobserved values of **Y**.

The MAR assumption cannot be verified from the data. For a study that assumes MAR, the sensitivity of inferences to departures from this assumption should be examined (National Research Council 2010, p. 111). If it is plausible that the missing data are not MAR, you can perform a sensitivity analysis under the assumption that the data are missing not at random (MNAR). That is, you impute missing values under a plausible MNAR scenario, and you examine the results. If the conclusion differs from the conclusion under MAR, then you should question the MAR assumption.

You can use the new MNAR statement in the MI procedure to impute missing values by assuming MNAR. You can specify a subset of observations from which imputation models are to be derived for specified variables, and you can specify an imputed variable and adjustment parameters (such as shift and scale) for adjusting the imputed variable values for a specified subset of observations. For a detailed explanation of how the MI procedure performs sensitivity analysis, see Yuan (2014).

Introductory Example

A clinical trial is conducted to test the efficacy of a new drug and involves two groups of patients: a treatment group that receives the new drug, and a control group that receives a placebo. Efficacy scores for the patients are saved in a data set named Mono1. The variable **Trt** is an indicator for treatment. The variable **Y0** is the baseline efficacy score, and the variable **Y1** is the efficacy score at a follow-up visit. For some patients, the values of **Y1** are missing.

The following statements use the MI procedure to impute missing values for **Y1** under the MAR assumption. The results of a regression analysis for each of the ten imputed data sets are combined by using the MIANALYZE procedure.

```
proc mi data=Mono1 seed=14823 nimpute=10 out=Out1;
   class Trt;
   monotone reg( y1);
   var Trt y0 y1;
run;
proc reg data=Out1;
   model y1= Trt y0;
   by  _Imputation_;
   ods output parameterestimates=Parm1;
run;
proc mianalyze parms=Parm1;
   modeleffects Trt;
run;
```

The table in Figure 7 shows that the treatment effect is significant at a level of 0.05.

Figure 7 Parameter Estimates

Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
Trt	0.893577	0.265276	0.366563	1.420591	90.029	0.624115	1.121445	0	3.37	0.0011

You can analyze the sensitivity of this conclusion to the MAR assumption by using a control-based pattern imputation under the MNAR assumption, as described by Ratitch and O'Kelly (2011). An imputation model for missing observations in the treatment group is constructed not from the observed data in the treatment group, but rather from the observed data in the control group. This model is also used to impute missing observations in the control group.

```
proc mi data=Mono1 seed=14823 nimpute=10 out=Out1X;
   class Trt;
   monotone reg(y1);
   mnar model( y1 / modelobs=(Trt='0'));
   var y0 y1;
run;
```

The MNAR statement imputes missing values for scenarios under the MNAR assumption. The MODEL option specifies that only observations for which TRT=0 are used to derive the imputation model for the variable **Y1**. Thus, **Y0** and **Y1** (but not **Trt**) are specified in the VAR statement.

As in the analysis under the MAR assumption, the REG and MIANALYZE procedures (not shown) are used to perform regression tests on the imputed data sets and to combine the results. The table in Figure 8 shows that the significance of the treatment effect is not reversed under control-based pattern imputation.

Figure 8 Parameter Estimates

Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
Trt	0.664712	0.297378	0.069701	1.259724	59.197	0.329363	0.892285	0	2.24	0.0292

Weighted GEE Method for Missing Data with the GEE Procedure

Missing observations are common in longitudinal studies, where they result from dropouts or skipped visits. To draw valid inference when data are missing, you can use different approaches, such as maximum likelihood, multiple imputation, fully Bayesian analysis, and inverse probability weighting methods (Little and Rubin 2002; National Research Council 2010). The GEE procedure, introduced in SAS/STAT 13.2, provides a weighted generalized estimating equation (GEE) method for analyzing longitudinal data that have missing observations. This approach extends the usual GEE approach (Liang and Zeger 1986).

When none of the data are missing, the weighted GEE method is identical to the usual GEE method, which is available in the GENMOD procedure. The usual method, which is based on complete cases, is valid if the data are missing completely at random (MCAR), but it can lead to biased results if the data are missing at random (MAR). In contrast, the GEE procedure implements inverse probability weighting to account for dropouts under the MAR assumption (Robins and Rotnitzky 1995; Preisser, Lohman, and Rathouz 2002). Observation-specific and subject-specific weighted methods are available for estimating regression parameters when dropouts occur. Both methods provide consistent estimates if the data are MAR.

For an in-depth introduction to the GEE procedure, see Lin and Rodriguez (2014).

Introductory Example

This example is taken from a longitudinal study of women who used contraception during four consecutive months (Fitzmaurice, Laird, and Ware 2011). In this study 1,151 women were randomly assigned to one of two treatments: 100 mg or 150 mg of depot-medroxyprogesterone acetate (DMPA). The response variable is their amenorrhea status in each of the four months. The example follows the analysis done by Fitzmaurice, Laird, and Ware (2011), where the question is whether the treatment had an effect on the rate of amenorrhea over time.

The following statements create the data set Amenorrhea:

```
data Amenorrhea;
 input id dose time y@@;
 datalines;
   1     0       1     0
   1     0       2     .
   1     0       3     .

   ... more lines ...

1151     1       4     1
;
```

Two additional variables, **prevy** and **ctime**, are created for use in the analysis:

```
data Amenorrhea;
  set Amenorrhea;
  by id;
  prevy=lag(y);
  if first.id then prevy=.;
  time=time-1;
  ctime=time;
run;
```

The variables are as follows:

- **id**: patient ID
- **y**: indicator of amenorrhea symptoms (1 for amenorrhea; 0 otherwise)
- **prevy**: indicator of amenorrhea symptoms in previous month
- **time**: month (0, 1, 2, or 3)
- **ctime**: a copy of **time** that is used as a class variable in the missingness model
- **dose**: administered dose (0 for 100 mg, 1 for 150 mg)

Suppose that y_{ij} denotes the amenorrhea status of woman i at the jth visit, $j = 1, \ldots, 4$, and let $\mu_{ij} = \mathrm{P}(y_{ij} = 1)$ denote the average rate of this status. To explore whether the treatment has an effect on the rate of amenorrhea over time, the following marginal response model is proposed:

$$\mathrm{logit}(\mu_{ij}) = \beta_0 + \beta_1 \mathrm{time}_{ij} + \beta_2 \mathrm{time}_{ij}^2 + \beta_3 \mathrm{dose}_i + \beta_4 \mathrm{dose}_i \times \mathrm{time} + \beta_5 \mathrm{dose}_i \times \mathrm{time}^2$$

A logistic regression model for missingness is proposed to obtain the weights for the weighted GEE method:

$$\begin{aligned}\mathrm{logit}\, p(r_{ij} = 1 | r_{ij-1} = 1, \mathrm{dose}_i, \mathrm{time}_{ij}, y_{ij-1}) =& \alpha_0 + \alpha_1 I(\mathrm{time}_{ij} = 2) + \alpha_2 I(\mathrm{time}_{ij} = 3) \\ &+ \alpha_3 \mathrm{dose}_i + \alpha_4 y_{ij-1} + \alpha_5 \mathrm{dose}_i \times y_{ij-1}\end{aligned}$$

The next statements use the observation-specific weighted GEE method to analyze the data:

```
proc gee data=Amenorrhea desc;
  class id ctime;
  missmodel ctime prevy dose dose*prevy / type=obslevel; /* missingness model */
  model y=time dose time*time dose*time dose*time*time / dist=bin; /* marginal model */
  repeated subject=id/within=ctime corr=cs;
run;
```

The MODEL statement specifies the regression model for the mean response, and the MISSMODEL statement specifies the logistic regression model for missingness. The REPEATED statement requests the GEE approach for correlated data analysis.

Figure 9 shows the parameter estimates for the missingness model. The estimate of α_4 is –0.4514 with a p-value of 0.0053, which indicates that the probability of dropout is related to previous amenorrhea status. This result suggests that it is more appropriate to assume that the data are MAR than to assume that the data are MCAR.

Figure 9 Parameter Estimates for Missingness Model

Maximum Likelihood Parameter Estimates for Missingness Model

Parameter		Estimate	Standard Error	Wald 95% Confidence Limits		Z	Pr > |Z|
Intercept		2.3967	0.1438	2.1149	2.6785	16.67	<.0001
ctime	1	-0.7286	0.1439	-1.0106	-0.4466	-5.06	<.0001
ctime	2	-0.5919	0.1469	-0.8798	-0.3040	-4.03	<.0001
ctime	3	0.0000	0.0000	0.0000	0.0000	.	.
prevy		-0.4514	0.1619	-0.7687	-0.1341	-2.79	0.0053
dose		0.0680	0.1313	-0.1893	0.3253	0.52	0.6046
prevy*dose		-0.2381	0.2196	-0.6685	0.1923	-1.08	0.2782

Figure 10 shows the parameter estimates for the marginal model. The interactions **dose*time** and **dose*time*time** are significant, indicating that the change of rate over time depends on the dose.

Figure 10 Parameter Estimates for Marginal Model

Analysis of GEE Parameter Estimates

Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-1.4965	0.1072	-1.7067	-1.2863	-13.95	<.0001
time	0.5379	0.1334	0.2764	0.7994	4.03	<.0001
dose	0.1061	0.1491	-0.1861	0.3983	0.71	0.4767
time*time	-0.0037	0.0405	-0.0831	0.0757	-0.09	0.9275
dose*time	0.4092	0.1903	0.0362	0.7823	2.15	0.0315
dose*time*time	-0.1264	0.0577	-0.2395	-0.0134	-2.19	0.0284

Bayesian Analysis of Discrete Choice Models with the BCHOICE Procedure

The BCHOICE procedure performs Bayesian analysis for discrete choice models. These models are used in modern marketing research and related areas, where it is important to model the process that a consumer follows to choose products or reach decisions when faced with multiple alternatives. The rising popularity of discrete choice models coincides with the growth of Bayesian approaches, which offer convenience for modeling and computation that would otherwise be difficult.

Discrete choice models are derived under the assumption of utility maximization. The utility function for a decision maker consists of a component that is observed by the researcher and an error component that is not observed. Different specifications for the distribution of the error component result in different choice models: logit, nested logit, and probit. The BCHOICE procedure supports these three models, and it can include random effects for estimating individual-level parameters, which enable you to infer heterogeneity in product preferences and price sensitivity.

To use the BCHOICE procedure, you specify a choice model and define the choice sets. You can supply prior distributions for the model parameters as alternatives to default noninformative priors. PROC BCHOICE uses Markov chain Monte Carlo (MCMC) methods to sample from the posterior distributions, and sampling algorithms are multithreaded for faster performance.

Introductory Example

This example discusses a logit model that has random effects and thereby enables you to characterize and estimate heterogeneity among individuals. The Bayesian approach provides a ready solution, unlike frequentist methods that optimize the likelihood function for this model and are numerically difficult.

A study estimates the market demand for kitchen trash cans (Rossi 2013). The cans have four attributes, each of which has two levels: touchless opening (1=Yes, 0=No), steel material (1=Yes, 0=No), automatic bag replacement (1=Yes, 0=No), and price (1=USD80, 0=USD40). Data were obtained by enrolling 104 people and assigning 10 choice tasks (choice sets) to each participant. For each task, the participant stated a preference between two types of trash cans. Figure 11 shows the data for the first three tasks that were assigned to the first participant.

Figure 11 Data for First Three Choice Tasks

Obs	ID	Task	Choice	Touchless	Steel	AutoBag	Price80
1	1	1	1	0	1	1	0
2	1	1	0	1	1	0	0
3	1	2	0	0	0	0	0
4	1	2	1	1	1	1	1
5	1	3	0	0	0	0	0
6	1	3	1	1	1	0	0

In the data, **ID** is the individual's ID, and **Task** indexes the tasks. The response is **Choice**, which records the individual's choice (1=chosen, 0=not chosen). The variables **Touchless**, **Steel**, **AutoBag**, and **Price80** are indicators for the attributes.

The following statements fit a logit model that has random effects:

```
proc bchoice data=Trashcan nmc=20000 seed=123 nthread=4;
   class ID Task;
   model Choice =  / choiceset=(ID Task);
   random Touchless Steel AutoBag Price80 / sub=ID remean type=un;
run;
```

The CLASS statement names the classification variables. The MODEL statement specifies the dependent variable and fixed effects (there are none here). The CHOICESET= option specifies how a choice set is defined. The RANDOM statement specifies **Touchless**, **Steel**, **AutoBag**, and **Price80** as random effects; the REMEAN option estimates the population means of the random effects. The SUB= option defines **ID** as a subject index for the random-effects grouping. The TYPE= option specifies an unstructured covariance matrix for the random effects.

Figure 12 shows posterior summary statistics and 95% highest posterior density (HPD) intervals for the parameters.

Figure 12 Posterior Summary Statistics

Posterior Summaries and Intervals

Parameter	N	Mean	Standard Deviation	95% HPD Interval	
REMean Touchless	20000	1.7244	0.2768	1.1945	2.2701
REMean Steel	20000	1.0501	0.2594	0.5651	1.5916
REMean AutoBag	20000	2.2016	0.3665	1.5012	2.9281
REMean Price80	20000	-4.6801	0.6745	-6.0233	-3.3977
RECov Touchless, Touchless	20000	3.1543	1.1459	1.2923	5.2574
RECov Steel, Touchless	20000	-0.5430	0.8639	-2.3815	0.9947
RECov Steel, Steel	20000	2.7095	1.0287	1.0180	4.7272
RECov AutoBag, Touchless	20000	-0.6107	0.9366	-2.3945	1.2510
RECov AutoBag, Steel	20000	0.00844	0.7357	-1.5000	1.4219
RECov AutoBag, AutoBag	20000	3.6673	1.5595	1.1457	6.7011
RECov Price80, Touchless	20000	-1.3213	1.2650	-3.7587	0.7990
RECov Price80, Steel	20000	-1.5920	1.2508	-4.1760	0.7193
RECov Price80, AutoBag	20000	-2.3216	1.7357	-5.8465	0.6625
RECov Price80, Price80	20000	7.8238	3.3671	2.4007	14.2544

The means of the random effects (**Touchless**, **Steel**, **AutoBag**, and **Price80**) are shown in the first four rows of Figure 12. Across all respondents, the average part-worths for touchless opening, steel material, and automatic bag replacement are all positive, indicating that most people favor those features. The average part-worth for having to pay more is negative, which seems intuitive.

The estimated covariance matrix for the random effects is given by the means for parameters whose label begins with “RECov”:

$$\hat{\boldsymbol{\Omega}}_{\boldsymbol{\gamma}} = \begin{pmatrix} 3.15 & . & . & . \\ -0.54 & 2.71 & . & . \\ -0.61 & 0.01 & 3.67 & . \\ -1.32 & -1.60 & -2.32 & 7.82 \end{pmatrix}$$

The covariance matrix characterizes the variability of part-worths across respondents. For example, the variance for price (labeled “RECov Price80, Price80”) is quite large, indicating a substantial difference in response to price. Off-diagonal elements of the matrix correspond to pairs of attribute levels that tend to be evaluated similarly (positive covariance) or differently (negative covariance) across all respondents.

The posterior means of the covariances between **Price80** and each of the attributes **Touchless**, **Steel**, and **AutoBag** are all negative. Although the 95% HPD intervals include 0, about 90% of the posterior samples for each of the three covariances are less than 0. This implies that respondents who prefer some of the new features are also unwilling to pay a higher price for the trash can. Offering a discounted price might be an effective method for introducing the new features to customers.

Overall, this analysis reveals that individuals are diverse in their preferences for features. Heterogeneity among individuals in the market can be an important aspect of choice models, and ignoring it can lead to incorrect inferences.

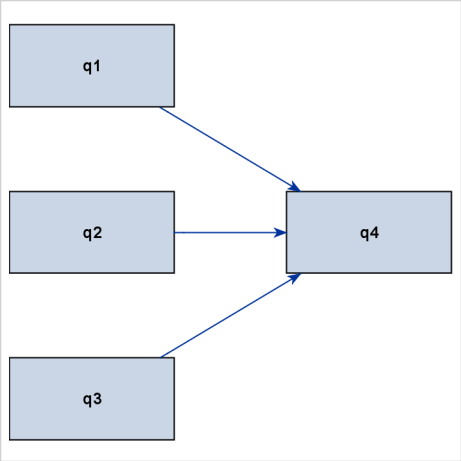
Creating Path Diagrams with the CALIS Procedure

The CALIS procedure fits structural equation models, which have a wide range of applications in education, psychology, sociology, health science, and marketing. In a structural equation model, you hypothesize functional or causal relationships among observed and latent variables.

You can use the CALIS procedure to specify complex relationships and estimate model parameters by maximum likelihood or other estimation methods. In SAS/STAT 13.1, the CALIS procedure produces path diagrams that represent models graphically. You can request and customize path diagrams for initial, standardized, and unstandardized models.

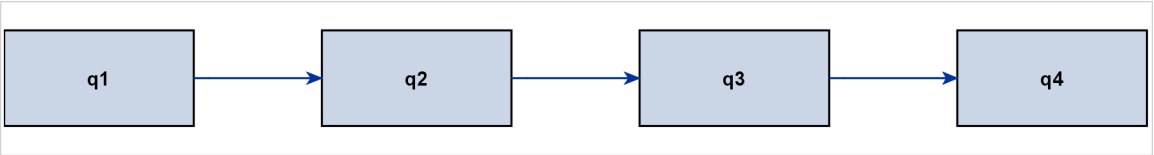
Path diagrams provide visually compelling representations of theoretical models and statistical results. To see how a path diagrams works, it is convenient to start with a basic regression model, which is a special case of a structural equation model. Consider the sales (in millions) of a company in four quarters, **q1**, **q2**, **q3**, and **q4**. You can use a regression model to predict **q4** from **q1**, **q2**, and **q3**. The path diagram in Figure 13 represents this model.

Figure 13 Regression Model for Q4 Sales



You can specify more complex models with the CALIS procedure. For example, the path diagram in Figure 14 represents a model that assumes only sequential functional relationships between sales in adjacent quarters.

Figure 14 Sequential Functional Relationships for Sales



Here, in contrast to the basic regression model, the variables **q1** and **q2** do not directly predict fourth quarter sales—that is, there are no direct arrows from sales in the first two quarters to **q4**. Instead, **q1** and **q2** influence **q4** only through their direct and indirect effects on **q3**. In general, path diagrams not only provide intuitive ways to present complex variable relationships in models, but they are also useful for visually contrasting different types of models.

Introductory Example

To show how PROC CALIS creates path diagrams from model input, consider a data set from Wheaton et al. (1977). The variables **Anomie67** and **Powerless67** are measures of alienation in 1967, which is a latent construct denoted as **Alien67**. Similarly, variables **Anomie71** and **Powerless71** are measures of alienation in 1971, which is a latent construct denoted as **Alien71**. Variables **Education** and **SEI** are measures of socioeconomic status in 1967, which is a latent construct denoted as **SES**.

In the following statements, the PATH statement in the CALIS procedure specifies the relationships among the observed variables and the latent constructs. The values in the PATH statement are fixed values for corresponding path effects. Otherwise, the effects are free to be estimated.

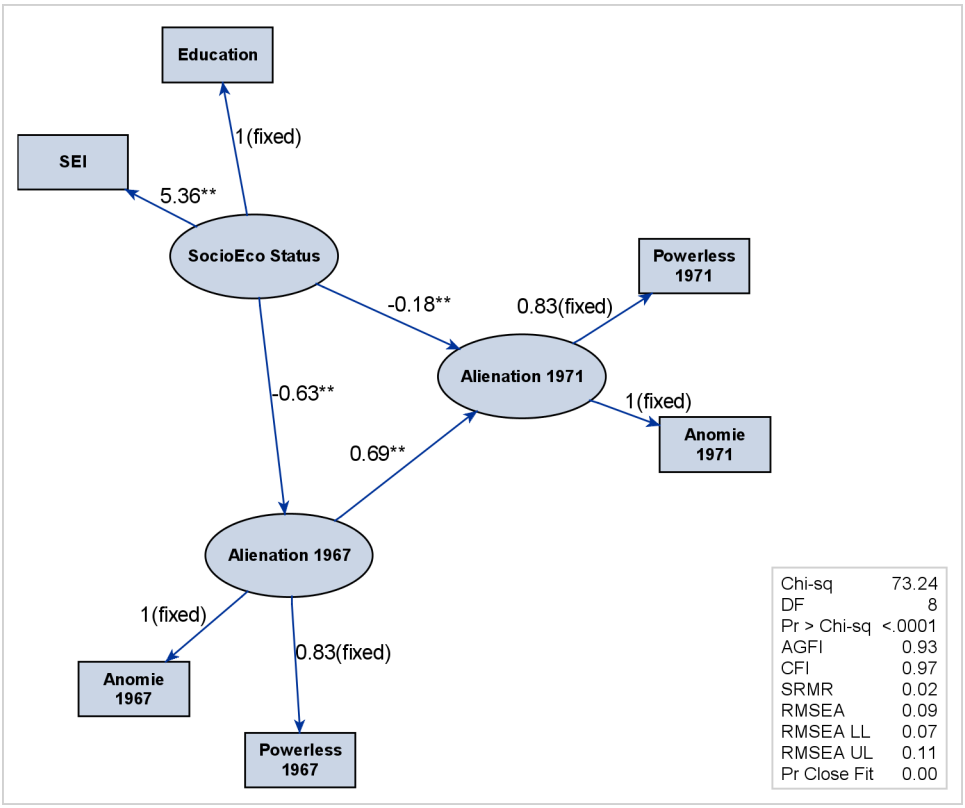
```
proc calis nobs=932 data=Wheaton;
   path
      Anomie67   Powerless67  <===  Alien67   = 1.0  0.833,
      Anomie71   Powerless71  <===  Alien71   = 1.0  0.833,
      Education  SEI          <===  SES       = 1.0  ,
      Alien67    Alien71      <===  SES       ,
      Alien71                 <===  Alien67    ;
   pathdiagram
      label=[Powerless71="Powerless 1971" Powerless67="Powerless 1967"
             Alien71="Alienation 1971" Alien67="Alienation 1967"
             Anomie71="Anomie 1971" Anomie67="Anomie 1967" Education="Education"
             SES="SocioEco Status"] notitle novariance;
run;
```

Although you can request a default path diagram simply by specifying PLOT=PATHDIAGRAM in the PROC CALIS statement, this example uses the PATHDIAGRAM statement, which enables you to customize the path diagram. The LABEL= option specifies labels that are to be displayed in place of variable names. The NOTITLE option suppresses the default title, and the NOVARIANCE option suppresses the display of variance parameters.

Figure 15 shows the path diagram of the unstandardized solution. Estimates that are significant at the 0.01 α-level are flagged by two asterisks. Estimates that are significant at the 0.05 α-level are flagged by a single asterisk. Fit statistics are displayed in an inset.

The CALIS procedure provides many more options for customizing path diagram output. For additional information, see the chapter "The CALIS Procedure" in *SAS/STAT User's Guide*.

Figure 15 Stability of Alienation



Analysis of Item Response Models with the IRT Procedure

Item response theory (IRT) was first proposed in the field of psychometrics for assessing ability. It is widely used in education to calibrate and evaluate items in tests, questionnaires, and other instruments, and to score subjects on their abilities, attitudes, and other latent traits.

During the past several decades, educational assessment has increasingly used IRT-based techniques to develop tests. Today, all major educational tests, such as the Scholastic Aptitude Test (SAT) and the Graduate Record Examination (GRE), are developed with the aid of item response theory because it improves measurement accuracy and reliability while reducing assessment time and effort, especially for computerized adaptive testing.

In recent years, IRT-based models have also become increasingly popular in studies of health outcomes and quality of life, and in clinical research (Hays, Morales, and Reise 2000; Edelen and Reeve 2007; Holman, Glas, and de Haan 2003; Reise and Waller 2009).

Early IRT models, such as the Rasch model and two-parameter model, concentrated on analyzing dichotomous responses that have a single latent trait. Extensions of these models that offer more flexibility have been developed. These extensions are not mutually exclusive, and they can be combined to address the complexity of data and to test substantive theory in practical applications.

The IRT procedure handles most of these extended models, including multidimensional exploratory models, multidimensional confirmatory models, and multigroup analysis. For more information about various item response models, see De Ayala (2009) and Embretson and Reise (2000). For a comprehensive introduction to the IRT procedure, see An and Yung (2014).

Introductory Example

This example illustrates the use of the IRT procedure with data from the Law School Admission Test (LSAT). The data include responses from 1,000 subjects to five binary items. The following statements create the data set IrtLsat:

```
data IrtLsat;
   input item1-item5 @@;
   datalines;
0 0 0 0 0
0 0 0 0 0
0 0 0 0 0

   ... more lines ...

1 1 1 1 1
1 1 1 1 1
;
```

The next statements use the IRT procedure to fit a default item response model:

```
proc irt data=IrtLsat plots=icc;
   var item1-item5;
run;
```

By default, PROC IRT fits a unidimensional two-parameter model for binary responses. This model assumes that the correlation among the items can be explained by a single latent factor. You can check this assumption by examining the eigenvalues (not shown here). For this example, the first eigenvalue is much larger than the others and accounts for almost 40% of the common variance, which suggests that a unidimensional model is reasonable for this example.

The parameter estimates that are produced by the IRT procedure are shown in Figure 16.

Figure 16 Parameter Estimates for Two-Parameter Model

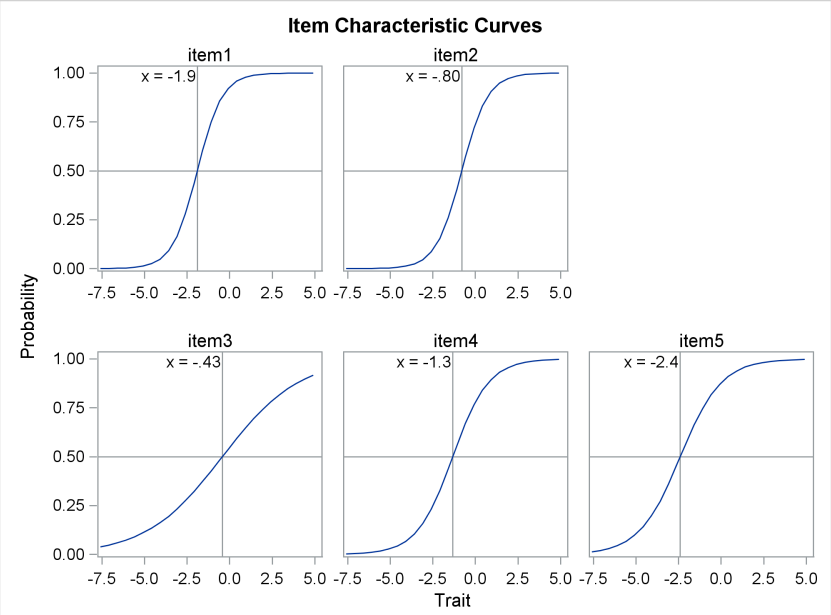
Item Parameter Estimates

Item	Parameter	Estimate	Standard Error	Pr > |t|
item1	**Threshold**	-2.59087	0.22115	<.0001
	Slope	1.36225	0.25067	<.0001
item2	**Threshold**	-1.05859	0.12506	<.0001
	Slope	1.32388	0.27282	<.0001
item3	**Threshold**	-0.19313	0.06667	0.0019
	Slope	0.44845	0.11478	<.0001
item4	**Threshold**	-1.26496	0.10733	<.0001
	Slope	0.95289	0.18798	<.0001
item5	**Threshold**	-1.98140	0.12915	<.0001
	Slope	0.82665	0.17174	<.0001

The threshold parameter estimates range from –2.59 to –0.19, corresponding respectively to **item1**, which is the easiest item, and **item3**, which is the most difficult item. The fact that all the threshold parameters are less than 0 suggests that all the items are relatively easy and are most useful in discriminating subjects who have lower abilities.

The PLOTS=ICC option requests plots of item characteristic curves (ICCs), which are shown in Figure 17. The reference line indicates the difficulty parameter value for the item. This parameter applies only to binary responses and is a transformation of the threshold parameter. The probability of correct response is 0.5 for any subject whose ability is equal to the value of the difficulty parameter.

Figure 17 Item Characteristic Curves



The slope parameter is a measure of the differential capability of an item. A high slope parameter suggests that the item provides a high degree of differentiation among subjects who have similar abilities. The slope parameter estimates in Figure 16 range from 0.45 to 1.36. By comparing the ICCs in Figure 17, you can observe how the slope parameter affects the shape of the ICC. The ICC for **item1** is the steepest, and the ICC for **item3** is flattest.

Analysis of Weighted Multilevel Models with the GLIMMIX Procedure

Multilevel modeling is appropriate for data that arise from a multistage sampling design. In such designs, you first select a sample of primary sampling units (PSU) and then, from each sampled PSU, you select a sample of secondary sampling units (SSU). You continue selecting smaller clusters or units until you reach the individual units for which survey responses are measured.

Each stage in a multistage sampling design maps directly to a level in a multilevel model. The nested clusters in the design correspond to hierarchical random effects in the model, and the characteristics of the units at each stage become the explanatory variables at the level that corresponds to that stage.

A multistage sampling design often involves unequal sampling probabilities, rational oversampling, poststratification, and nonresponse; these are used to derive survey weights. In order to draw valid inferences about the population of interest, you need to take the weights into account. Because information is collected on a sample of the population, you cannot compute the likelihood for the entire population. Instead, you can use the weighted likelihood as an estimate for the population likelihood.

In SAS/STAT 13.1, the GLIMMIX procedure provides two new weighting options that enable you to accommodate the weighting scheme in a multistage sample:

- The OBSWEIGHT= option in the MODEL statement specifies observational level weights.
- The WEIGHT= option in the RANDOM statement specifies weights at higher levels.

For inference about fixed effects and variance parameters that are estimated by the weighted likelihood method, you can use empirical (sandwich) variance estimators. Zhu (2014) discusses the new weighting options in detail.

Introductory Example

This example uses data about reading proficiency of 15-year-old American students from the Programme for International Student Assessment (PISA) study to show how you can use PROC GLIMMIX to fit a weighted multilevel model.

The PISA study is based on a three-stage sampling design. Geographic areas (PSUs) are sampled at stage 1, schools within areas are sampled at stage 2, and students within schools are sampled at stage 3. Altogether, 2,069 students from 148 schools in 46 PSUs are included in the analysis. The student-level weight (**StuWt**) and the school-level weight (**SchWt**) are inverse sampling probabilities that are further adjusted for nonresponse and noninclusion. To reduce the bias in variance parameter estimates, scaled student-level weights (**ScaledStuW**) are computed using Method 1 in Pfeffermann et al. (1998) and Rabe-Hesketh and Skrondal (2006).

The outcome considered here is the binary variable **Passread**. This variable takes the value 1 when the reading proficiency is at the top two levels. The variable **Idschool** identifies the schools. The explanatory variables are as follows:

- **Female** indicates whether the student is female.
- **ISEI** indicates the student's international socioeconomic index (ISEI).
- **Highschool** indicates whether the highest education level by either parent is high school.
- **College** indicates whether the highest education level by either parent is college.
- **Testlang** indicates whether the test language (English) is spoken at home.
- **Onefor** indicates whether one parent is foreign-born.
- **Bothfor** indicates whether both parents are foreign-born.
- **MISEI** indicates the school mean ISEI.

Of these eight variables, **MISEI** is a school-level covariate and the rest are student-level covariates. The following statements create the data set:

```
data pisa;
   input female isei schwt highschool college onefor bothfor
         testlang passread idschool misei scaledstuw;
   datalines;
1 37 146.02000427 0 1 1 0 1 1 1 48.25 1
1 77 146.02000427 1 0 0 0 1 1 1 48.25 1
1 53 146.02000427 0 1 0 0 1 0 1 48.25 1

   ... more lines ...

0 23 140.91999817 1 0 0 0 1 0 151 40.5 0.9620853081
0 23 140.91999817 0 1 0 0 1 0 151 40.5 0.9620853081
;
```

The following statements fit a weighted two-level random-intercept logistic model:

```
proc glimmix data=pisa method=quadrature empirical=classical;
   model passread = isei female highschool college onefor
                    bothfor testlang misei
                    / dist=binomial link=logit obsweight=scaledstuw solution;
   random intercept / subject=idschool weight=schwt;
run;
```

When fitting a weighted multilevel model, you should use the METHOD=QUADRATURE estimation option. The EMPIRICAL=CLASSICAL option requests empirical (sandwich) variance estimators, which are recommended for inference on fixed effects and variance parameters in weighted multilevel models.

The OBSWEIGHT= option specifies the weight variable **ScaledStuW** for the student level. The WEIGHT= option specifies the weight variable **SchWt** for the school level, which is identified by the SUBJECT=**Idschool** option.

The estimates of the fixed effects are shown in Figure 18.

Figure 18 Fixed-Effects Solutions

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-5.8743	0.9495	147	-6.19	<.0001
isei	0.01821	0.004789	1913	3.80	0.0001
female	0.6222	0.1534	1913	4.05	<.0001
highschool	0.1030	0.4755	1913	0.22	0.8285
college	0.4532	0.5036	1913	0.90	0.3683
onefor	-0.1091	0.2730	1913	-0.40	0.6895
bothfor	-0.2802	0.3252	1913	-0.86	0.3890
testlang	0.6251	0.3808	1913	1.64	0.1008
misei	0.06819	0.01635	1913	4.17	<.0001

These estimates have implications that agree with prior educational studies. For example, it is expected that female students tend to be more proficient in reading than male students. Also, a student's international socioeconomic index is believed to have an impact on a student's reading ability.

Bootstrap Methods for Inference with the NLIN Procedure

By default, the NLIN procedure provides standard confidence intervals and estimates of standard errors, which assume asymptotic normality of parameter estimates. However, even for a correctly specified nonlinear regression model, the reliability of standard inference methods can be questionable. Prior to SAS/STAT 9.3, the only feature that the procedure offered for checking this reliability was Hougaard's skewness measure.

Now, PROC NLIN offers many such features, together with alternative methods of inference when standard methods fail. Global measures of curvature, Box's bias, and graphical diagnostics were added in SAS/STAT 9.3; the PROFILE statement for parameter profiling was added in SAS/STAT 12.1; and the BOOTSTRAP statement for bootstrap estimation was added in SAS/STAT 13.1. For a discussion of these features, see Gebremariam (2014).

Introductory Example

This example demonstrates the utility of these features by fitting a nonlinear regression model based on the Mitscherlich equation and data from Clarke (1987):

```
data Clarke1987;
   input x y @@;
datalines;
1 3.183  2 3.059  3 2.871  4 2.622  5 2.541  6 2.184  7 2.110
8 2.075  9 2.018 10 1.903 11 1.770 12 1.762 13 1.550; run;

proc nlin data=Clarke1987 bias hougaard nlinmeasures;
  parms theta1=-0.15 theta2=2.0 theta3=0.80;
  model y = theta3 + theta2*exp(theta1*x);
  profile theta1 / range = -4 to 4 by 0.2 all;
  profile theta2 theta3 / range = -2 to 2 by 0.2 all;
  bootstrap / seed=123 nsamples=5000 bootci bootplots(all);
run;
```

Global measures of nonlinearity, shown in Figure 19, indicate a far-from-linear model.

Figure 19 Global Measures of Nonlinearity

Global Nonlinearity Measures	
Max Intrinsic Curvature	0.0935
RMS Intrinsic Curvature	0.0418
Max Parameter-Effects Curvature	7.0183
RMS Parameter-Effects Curvature	3.1301
Curvature Critical Value	0.5193
Raw Residual Variance	0.0053
Projected Residual Variance	0.0059

The effect of this nonlinearity on the parameter estimates is gauged by Box's bias and Hougaard's skewness measures (Bates and Watts 1988; Ratkowsky 1983). Figure 20 shows that the estimates of θ_2 and θ_3 are biased and have skewed distributions.

Figure 20 Parameter Estimates and Standard Confidence Intervals

			Approximate 95%			Bootstrap Bias-Corrected 95%				
Parameter	Estimate	Approx Std Error	Confidence Limits		Bootstrap Std Dev	Confidence Limits		Skewness	Bias	Percent Bias
theta1	-0.1031	0.0255	-0.1599	-0.0462	0.0250	-0.1508	-0.0535	-0.0619	-0.00026	0.26
theta2	2.5190	0.2658	1.9268	3.1112	0.4123	2.1838	3.5162	1.5428	0.0802	3.18
theta3	0.9631	0.3216	0.2466	1.6797	0.4608	-0.2531	1.3731	-1.3956	-0.0772	-8.02

These results indicate that the estimates of θ_2 and θ_3 are strongly affected by nonlinearity, whereas the estimate of θ_1 is minimally affected. This is contrary to what you might expect based on the model expression.

Further support for this conclusion comes from the confidence curves (Figure 21 and Figure 23) and histograms of the bootstrap parameter estimates (Figure 22 and Figure 24), which were requested with the BOOTPLOTS option. The likelihood-based confidence intervals for θ_3 deviate from the corresponding standard confidence intervals. In contrast, these two intervals are practically identical for θ_1. Furthermore, the distribution of bootstrap estimates of θ_3 in Figure 24 has a long left tail, which exhibits a large departure from normality, unlike the bootstrap estimates of θ_1 in Figure 22 which are normally distributed.

Figure 21 Confidence Curve of θ_1

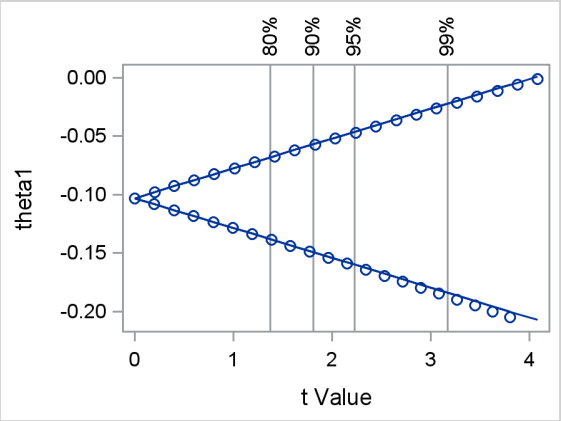


Figure 22 Distribution of Bootstrap Estimates of θ_1

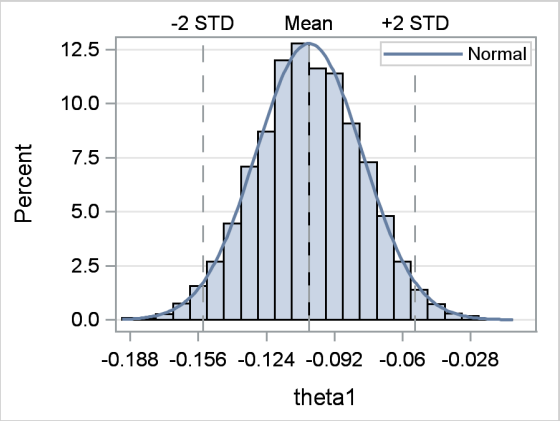


Figure 23 Confidence Curve of θ_3

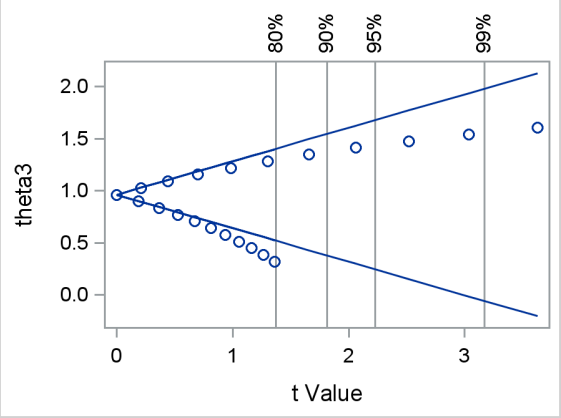
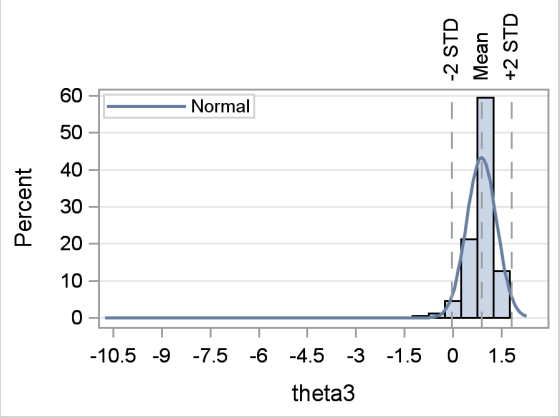


Figure 24 Distribution of Bootstrap Estimates of θ_3



If a nonlinear regression model shows negligible effects of nonlinearity, standard inference can be used and offers ease of verification and interpretation. For the model in this example, standard inference seems to be highly unreliable. Such cases call for inferential techniques (such as the bootstrap) that have better finite sample size properties or reparameterization of the model (DiCiccio and Efron 1996; Ratkowsky 1990). Figure 20 shows that only for θ_1 does the standard confidence interval tightly cover the bias-corrected bootstrap confidence interval.

For unbiased prediction with this model, reparameterization of the model is recommended. Useful hints for reparameterization that can result in a close-to-linear model can be drawn from histograms of the bootstrap estimates (Ratkowsky 1990).

For more information about approaches for diagnosing and reparameterizing a far-from-linear regression model, see Gebremariam (2014).

High-Performance Statistical Procedures

Beginning with the 12.3 release, SAS/STAT includes a family of high-performance procedures that are designed for predictive modeling and other large-data tasks. What sets these procedures apart from traditional SAS/STAT procedures is that you can run them in two ways:

- You can run a high-performance statistical procedure in single-machine mode on the server where SAS is installed, just like other SAS/STAT procedures. No additional license is required. High-performance procedures are multithreaded; they take advantage of multiple cores when run in single-machine mode.
- You can also run a high-performance statistical procedure in distributed mode on a cluster of machines that distribute the data and the computations. Because each node in the cluster does a slice of the work, the procedure takes full advantage of the computing power of the cluster to fit large models that have massive amounts of data. To run in distributed mode, you need to license SAS® High-Performance Statistics. For more information about distributed mode, see Cohen and Rodriguez (2013).

The high-performance procedures in SAS/STAT 13.1 are as follows:

- the new HPCANDISC procedure, which performs canonical discriminant analysis
- the new HPFMM procedure, which performs analysis of finite mixture models
- the HPGENSELECT procedure, which performs model selection for generalized linear models
- the HPLMIXED procedure, which fits mixed linear models
- the HPLOGISTIC procedure, which performs model selection for logistic regression models
- the HPNLMOD procedure, which fits nonlinear regression models
- the new HPPRINCOMP procedure, which performs principal component analysis
- the HPREG procedure, which performs model selection for linear regression models
- the HPSPLIT procedure, which creates decision tree models

These procedures are documented in *SAS/STAT User's Guide: High-Performance Procedures*.

High-performance statistical procedures do not have a one-to-one correspondence to traditional SAS/STAT procedures. The introduction of high-performance procedures affords the opportunity to consolidate similar functionality from several traditional SAS/STAT procedures into a single high-performance modeling procedure. For example, PROC HPREG has features drawn from the GLM, GLMSELECT, and REG procedures.

Furthermore, because high-performance statistical procedures are designed for predictive modeling and other large-data tasks, they do not implement all of the inferential features that are available in corresponding traditional procedures. For example, PROC HPLOGISTIC does not have all of the exact inferential methods that PROC LOGISTIC provides for small-to-moderate data.

Benefits of High-Performance Statistical Procedures

Although the size of your current data might not necessitate that you run high-performance procedures in distributed mode, they are all multithreaded and often perform faster than other procedures, depending on your data and your model. For example, if you are fitting a logistic model that involves a large amount of data, the HPLOGISTIC procedure could give you better performance than the LOGISTIC procedure, which is not multithreaded. However, multithreaded operations are also available in many traditional SAS/STAT procedures, including the ADAPTIVEREG, BCHOICE, FMM, IRT, GLM, GLMSELECT, LOESS, MCMC, REG, QUANTLIFE, QUANTREG, and QUANTSELECT procedures. For an explanation of the speedups that you can expect from thread-enabled procedures, see Cohen (2002).

A second benefit of high-performance statistical procedures is that they can offer functionality that is not available in traditional SAS/STAT procedures. An example is the HPGENSELECT procedure, which was introduced in SAS/STAT 12.3.

Selection of Generalized Linear Models with the HPGENSELECT Procedure

The HPGENSELECT procedure provides both model fitting and model building for generalized linear models. Like the GENMOD procedure, it fits generalized linear models using maximum likelihood. In addition, the HPGENSELECT procedure provides variable selection methods, including forward, backward, and stepwise selection, for building models. The HPGENSELECT procedure supports standard distributions and link functions, and it provides specialized models for zero-inflated count data, ordinal data, and unordered multinomial data.

In contrast to the GENMOD procedure, which offers a rich set of methods for statistical inference, the HPGENSELECT procedure is designed for predictive modeling and other large-data tasks.

Introductory Example

This example uses a simulated data set called GLM, which has one million observations. The data set contains a response variable **yPoisson**, which has a simulated Poisson distribution. The log transform of its mean depends linearly on five continuous regressors, named **xIn1** through **xIn5**, and on two classification regressors, named **cIn1** and **cIn2**. The data set also contains 85 noise variables (**xOut1** through **xOut80**, and **cOut1** through **cOut5**) that are not in the model (Cohen and Rodriguez 2013).

The following statements use the HPGENSELECT procedure, running in single-machine mode, to select a Poisson regression model from all the variables:

```
proc hpgenselect data=GLM;
   class c:;
   model yPoisson = x: c: / dist=Poisson;
   selection method=stepwise(choose=sbc);
   performance details;
run;
```

The SELECTION statement requests the stepwise selection method and specifies the Schwarz Bayesian criterion for choosing from the sequence of models that is generated in a stepwise fashion. The PERFORMANCE statement requests a table that provides details about where the computing time was spent.

The "Selection Summary" table in Figure 25 shows details about each step in the selection process.

Figure 25 Selection Summary

Selection Summary

Step	Effect Entered	Number Effects In	SBC	p Value
0	Intercept	1	2520370.52	.
1	cIn2	2	2514356.23	<.0001
2	xIn5	3	2509634.52	<.0001
3	xIn4	4	2506400.84	<.0001
4	xIn3	5	2504537.53	<.0001
5	xIn2	6	2503805.59	<.0001
6	cIn1	7	2503217.62	<.0001
7	xIn1	8	2503026.40*	<.0001
8	xOut5	9	2503032.21	0.0047
9	xSubtle	10	2503040.56	0.0194
10	xOut71	11	2503049.87	0.0338
11	xOut64	12	2503059.32	0.0368
12	xOut6	13	2503068.96	0.0411

*** Optimal Value of Criterion**

The selection process correctly selected the seven regressors whose names begin with the prefix **In**.

The timing table in Figure 26 indicates that the procedure took slightly more than three minutes. Most of this time was spent evaluating and refitting candidate models.

Figure 26 Timing

Procedure Task Timing

Task	Seconds	Percent
Reading and Levelizing Data	8.59	4.59%
Candidate evaluation	103.09	55.08%
Candidate model fit	69.37	37.07%
Final model fit	6.11	3.26%

Figure 27 shows the parameter estimates for the selected model.

Figure 27 Parameter Estimates

Parameter Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-0.091178	0.004568	398.4309	<.0001
xln1	1	-0.051745	0.003614	205.0205	<.0001
xln2	1	0.098711	0.003614	746.0052	<.0001
xln3	1	-0.156815	0.003617	1879.3093	<.0001
xln4	1	0.206283	0.003620	3247.0598	<.0001
xln5	1	-0.248897	0.003620	4726.2445	<.0001
cln1 1	1	-0.051178	0.002088	600.9542	<.0001
cln1 2	0	0	.	.	.
cln2 1	1	0.197963	0.002566	5953.5470	<.0001
cln2 2	1	0.096953	0.002627	1361.9474	<.0001
cln2 3	0	0	.	.	.

Summary

Enhancements in SAS/STAT 13.1 and SAS/STAT 13.2 deliver the following benefits:

- Powerful extensions of widely used methods: The new ICLIFETEST procedure provides nonparametric techniques for survival analysis of interval-censored lifetime data. The PHREG procedure provides the proportional subdistribution hazards model for survival analysis in the presence of competing risks. The new GEE procedure provides the weighted generalized estimating equation approach, which extends the usual GEE approach to longitudinal data that have missing values. The GLIMMIX procedure provides new weighting options for analysis of data that arise from a multistage sampling design.
- Additional model-building methods: The GLMSELECT procedure provides the elastic net method for selecting effects in general linear models. The HPGENSELECT procedure (added in SAS/STAT 12.3) provides effect selection for generalized linear models.
- More ways to assess assumptions: The MI procedure provides sensitivity analysis for the MAR assumption in multiple imputation. The NLIN procedure provides a rich set of methods for assessing and improving the reliability of estimates in nonlinear regression.
- New directions: The new BCHOICE procedure provides Bayesian analysis for discrete choice models, which are used in market research. The new IRT procedure fits item response models which are the basis for educational testing and are also applied in studies of health outcomes and clinical research.
- Even more graphical displays: Graphs are as much a part of output as tables in new development, and graphs continue to be added to existing procedures. The CALIS procedure provides extensive functionality for creating path diagrams that visualize structural equation models.

- Greater computational performance: The MCMC procedure improves the performance of Markov chain Monte Carlo sampling through multithreading. High-performance statistical procedures provide capability for predictive modeling and other large-data tasks, and they take advantage of multiple cores when run in single-machine mode.

Keeping Up with New Releases of SAS/STAT

The best place to find out about enhancements is the chapter "What's New in SAS/STAT" in the online documentation at http://support.sas.com/documentation/onlinedoc/stat/.

Also, be sure to visit the Statistics and Operations Research focus area, whose portal is http://support.sas.com/statistics. There you can watch helpful videos, download overview papers, and subscribe to the quarterly e-newsletter.

REFERENCES

An, X. and Yung, Y.-F. (2014), "Item Response Theory: What It Is and How You Can Use the IRT Procedure to Apply It," in *Proceedings of the SAS Global Forum 2014 Conference*.
URL http://support.sas.com/resources/papers/proceedings14/SAS364-2014.pdf

Bates, D. M. and Watts, D. G. (1988), *Nonlinear Regression Analysis and Its Applications*, New York: John Wiley & Sons.

Castelloe, J. (2014), "Power and Sample Size for MANOVA and Repeated Measures with the GLMPOWER Procedure," in *Proceedings of the SAS Global Forum 2014 Conference*, Cary, NC: SAS Institute Inc.
URL http://support.sas.com/resources/papers/proceedings14/SAS030-2014.pdf

Clarke, G. P. Y. (1987), "Approximate Confidence Limits for a Parameter Function in Nonlinear Regression," *Journal of the American Statistical Association*, 82, 221–230.

Cohen, R. (2002), "SAS Meets Big Iron: High Performance Computing in SAS Analytical Procedures," in *Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.

Cohen, R. and Rodriguez, R. N. (2013), "High-Performance Statistical Modeling," in *Proceedings of the SAS Global Forum 2013 Conference*, Cary, NC: SAS Institute Inc.

De Ayala, R. J. (2009), *The Theory and Practice of Item Response Theory*, New York: Guilford Press.

DiCiccio, T. J. and Efron, B. (1996), "Bootstrap Confidence Intervals," *Statistical Science*, 11, 189–212.

Edelen, M. O. and Reeve, B. B. (2007), "Applying Item Response Theory (IRT) Modeling to Questionnaire Development, Evaluation, and Refinement," *Quality of Life Research*, 16, 5–18.

Embretson, S. E. and Reise, S. P. (2000), *Item Response Theory for Psychologists*, Mahwah, NJ: Lawrence Erlbaum Associates.

Fay, M. P. (1999), "Comparing Several Score Tests for Interval Censored Data," *Statistics in Medicine*, 18, 273–285.

Fay, M. P. and Shih, J. H. (2012), "Weighted Logrank Tests for Interval Censored Data When Assessment Times Depend on Treatment," *Statistics in Medicine*, 31, 3760–3772.

Fine, J. P. and Gray, R. J. (1999), "A Proportional Hazards Model for the Subdistribution of a Competing Risk," *Journal of the American Statistical Association*, 94, 496–509.

Finkelstein, D. M. (1986), "A Proportional Hazards Model for Interval-Censored Failure Time Data," *Biometrics*, 42, 845–854.

Finkelstein, D. M. and Wolfe, R. A. (1985), "A Semiparametric Model for Regression Analysis of Interval-Censored Failure Time Data," *Biometrics*, 41, 933–945.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2011), *Applied Longitudinal Analysis*, Hoboken, NJ: John Wiley & Sons.

Gebremariam, B. (2014), "Is Nonlinear Regression Throwing You a Curve? New Diagnostic and Inference Tools in the NLIN Procedure," in *Proceedings of the SAS Global Forum 2014 Conference*.
URL http://support.sas.com/resources/papers/proceedings14/SAS384-2014.pdf

Guo, C., So, Y., and Johnston, G. (2014), "Analyzing Interval-Censored Data," in *Proceedings of the SAS Global Forum 2014 Conference*.
URL http://support.sas.com/resources/papers/proceedings14/SAS279-2014.pdf

Hays, R. D., Morales, L. S., and Reise, S. P. (2000), "Item Response Theory and Health Outcomes Measurement in the Twenty-First Century," *Medical Care*, 38, Suppl. 9, 1128–1142.

Holman, R., Glas, C. A. W., and de Haan, R. J. (2003), "Power Analysis in Randomized Clinical Trials Based on Item Response Theory," *Controlled Clinical Trials*, 24, 390–410.

Huang, J., Lee, C., and Yu, Q. (2008), "A Generalized Log-Rank Test for Interval-Censored Failure Time Data via Multiple Imputation," *Statistics in Medicine*, 27, 3217–3226.

Klein, J. P. and Moeschberger, M. L. (1997), *Survival Analysis: Techniques for Censored and Truncated Data*, New York: Springer-Verlag.

Liang, K.-Y. and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Lin, G. and Rodriguez, R. N. (2014), "Weighted Methods for Analyzing Missing Data with the GEE and CAUSALTRT Procedures," in *Proceedings of the SAS Global Forum 2014 Conference*.
URL http://support.sas.com/resources/papers/proceedings14/SAS166-2014.pdf

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 2nd Edition, Hoboken, NJ: John Wiley & Sons.

National Research Council (2010), *The Prevention and Treatment of Missing Data in Clinical Trials*, Panel on Handling Missing Data in Clinical Trials, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, Washington, DC: National Academies Press.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998), "Weighting for Unequal Selection Probabilities in Multilevel Models," *Journal of the Royal Statistical Society, Series B*, 60, 23–40.

Preisser, J. S., Lohman, K. K., and Rathouz, P. J. (2002), "Performance of Weighted Estimating Equations for Longitudinal Binary Data with Drop-Outs Missing at Random," *Statistics in Medicine*, 21, 3035–3054.

Rabe-Hesketh, S. and Skrondal, A. (2006), "Multilevel Modelling of Complex Survey Data," *Journal of the Royal Statistical Society, Series A*, 169, 805–827.

Ratitch, B. and O'Kelly, M. (2011), "Implementation of Pattern-Mixture Models Using Standard SAS/STAT Procedures," in *Proceedings of PharmaSUG 2011 (Pharmaceutical Industry SAS Users Group)*, SP04, Nashville.

Ratkowsky, D. (1983), *Nonlinear Regression Modeling*, New York: Marcel Dekker.

Ratkowsky, D. (1990), *Handbook of Nonlinear Regression Models*, New York: Marcel Dekker.

Reise, S. P. and Waller, N. G. (2009), "Item Response Theory and Clinical Measurement," *Annual Review of Clinical Psychology*, 5, 27–48.

Robins, J. M. and Rotnitzky, A. (1995), "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90, 122–129.

Rossi, P. E. (2013), personal communication.

Stokes, M. (2013), "Current Directions in SAS/STAT Software Development," in *Proceedings of the SAS Global Forum 2013 Conference*, Cary, NC: SAS Institute Inc.

Stokes, M., Chen, F., Yuan, Y., and Cai, W. (2012), "Look Out: After SAS/STAT 9.3 Comes SAS/STAT 12.1," in *Proceedings of the SAS Global Forum 2012 Conference*, Cary, NC: SAS Institute Inc.

Sun, J. (2001), "Variance Estimation of a Survival Function for Interval-Censored Survival Data," *Statistics in Medicine*, 20, 1249–1257.

Turnbull, B. W. (1976), "The Empirical Distribution Function with Arbitrarily Grouped, Censored, and Truncated Data," *Journal of the Royal Statistical Society, Series B*, 38, 290–295.

Wellner, J. A. and Zhan, Y. (1997), "A Hybrid Algorithm for Computation of the Nonparametric Maximum Likelihood Estimator from Censored Data," *Journal of the American Statistical Association*, 92, 945–959.

Wheaton, B., Muthén, B. O., Alwin, D. F., and Summers, G. F. (1977), "Assessing Reliability and Stability in Panel Models," in D. R. Heise, ed., *Sociological Methodology*, San Francisco: Jossey-Bass.

Yuan, Y. (2014), "Sensitivity Analysis in Multiple Imputation for Missing Data," in *Proceedings of the SAS Global Forum 2014 Conference*.
URL http://support.sas.com/resources/papers/proceedings14/SAS270-2014.pdf

Zhao, Q. and Sun, J. (2004), "Generalized Log-Rank Test for Mixed Interval-Censored Failure Time Data," *Statistics in Medicine*, 23, 1621–1629.

Zhu, M. (2014), "Analyzing Multilevel Models with the GLIMMIX Procedure," in *Proceedings of the SAS Global Forum 2014 Conference*.
URL http://support.sas.com/resources/papers/proceedings14/SAS026-2014.pdf

Acknowledgments

The following SAS/STAT developers contributed significantly to this paper: Xinming An, Bob Derr, Biruk Gebremariam, Changbin Guo, Warren Kuhfeld, Guixian Lin, Amy Shi, Ying So, Yang Yuan, Yiu-Fai Yung, and Min Zhu. The authors also thank Anne Baxter and Ed Huddleston for editorial assistance.

Contact Information

Your comments and questions are valued and encouraged. You can contact the authors at the following address:

Robert N. Rodriguez	Maura E. Stokes
SAS Institute Inc.	SAS Institute Inc.
SAS Campus Drive	SAS Campus Drive
Cary, NC 27513	Cary, NC 27513
Bob.Rodriguez@sas.com	Maura.Stokes@sas.com