Paper SAS1940-2015

SAS/STAT® 14.1: Methods for Massive, Missing, or Multifaceted Data

Maura Stokes and Statistical R&D Staff
SAS Institute Inc.

Abstract

The latest release of SAS/STAT® software brings you powerful techniques that will make a difference in your work, whether your data are massive, missing, or somewhere in the middle. New imputation software for survey data adds to an expansive array of methods in SAS/STAT for handling missing data, as does the production version of the GEE procedure, which provides the weighted generalized estimating equation approach for longitudinal studies with dropouts. An improved quadrature method in the GLIMMIX procedure means you can fit models not previously possible. The HPSPLIT procedure provides a rich set of methods for statistical modeling with classification and regression trees, including cross validation and graphical displays. The HPGENSELECT procedure adds support for LASSO model selection for generalized linear models. And new software implements generalized additive models by using an approach that handles large data easily. Other updates include key functionality for Bayesian analysis and pharmaceutical applications.

Overview

SAS/STAT 14.1 provides exciting new capabilities for its users in a number of areas. This paper describes the new features by area and also illustrates several key techniques with examples. The areas are:

- performance
- survey data analysis
- missing data analysis
- categorical data analysis
- classification and regression trees
- big data modeling
- survival analysis
- Bayesian analysis
- odds and ends

Performance doesn't just mean multithreading, and performance enhancements in this release range from solving problems that were previously intractable to providing more high-performance techniques for modeling Big Data. Missing data analysis techniques have been expanding rapidly in SAS/STAT in the past ten years, and the 14.1 release adds techniques for imputing survey data to the mix. The LOGISTIC procedure, never one for a dull moment, has extended unequal slopes models to all polytomous responses as well as providing the adjacent-category logit response function. The classification and regression trees are no longer just the purview of data miners, but are now available to SAS/STAT customers with the HPSPLIT procedure. Complete with a SAS/STAT modeling syntax, PROC HPSPLIT now produces tree plots, cost-complexity plots, and ROC curves in living color. The survival analysis tools are under constant review and development, and the techniques for analyzing interval-censored data have been enriched and a nonparametric test for analyzing competing-risk data is available in the LIFETEST procedure.Bayesian analysis is further extended with efficient sampling techniques in the MCMC procedure as well as facilities for solving general ordinary differential equations (ODEs) and fitting random process models, such as autoregressive or state-space models.

This paper describes the additions and provides examples of using the new capabilities for survey data imputation, logistic regression using adjacent-category logits, generalized additive models, and classification trees.

Performance

SAS/STAT statistical developers constantly look for ways to improve the performance of the statistical procedures in both speed and accuracy. Some of the statistical functionality has been made available for distributed computing with SAS® High-Performance Statistics software; all of its high-performance procedures are also included in SAS/STAT for single-machine use. Multithreading has long been incorporated into SAS/STAT procedures for those computational operations that might see performance benefits. Over a dozen procedures in SAS/STAT are multithreaded, and the NLMIXED procedure joins their ranks in the 14.1 release. Multithreading in PROC NLMIXED is available for the default METHOD=GAUSS method; data are allocated into different threads and the objective function is computed by accumulating values from each thread for the computed marginal log likelihood.

The QUANTREG procedure for quantile regression now supports a new alternative interior point algorithm that can be more efficient for large data. And the new FAST option in the PHREG procedure can speed up fitting of the Breslow and Efron partial likelihoods for the counting process style of input, especially when you have a large number of observations with many event times. It does so by requiring only one pass of the data to compute the likelihood, gradient, and Hessian. The new MCMC option for EXACT logistic regression in the LOGISTIC procedure implements the Markov chain Monte Carlo (MCMC) algorithm of Forster, McDonald, and Smith (2003), which can work when the asymptotic results are suspect, but network-based methods prove problematic.

And one of the most exciting performance improvements in the 14.1 release is the new FASTQUAD option in the GLIMMIX procedure, which enables you to fit multilevel models that have been computationally infeasible in the past. The following section describes this new addition.

Performance: Fitting Multilevel Models in PROC GLIMMIX with the FASTQUAD Option

In many applications, data are observed on nested units. To assess an educational program, you might collect data on the students, the classes to which they are assigned, and the schools that host these classes. To study a new drug, you might conduct a multicenter clinical trial where repeated measurements are observed on patients that are nested within centers. In a multistage survey design, you would collect data on subjects, the households to which they belong, and the counties in which the households are located.

You can use adaptive quadrature in PROC GLIMMIX to fit multilevel models to such multilevel data. The marginal distribution of a subject's data $\mathbf{y}_i$ in such a model is

$$p(\mathbf{y}_i) = \int \cdots \int p(\mathbf{y}_i|\boldsymbol{\gamma}_i, \boldsymbol{\beta}, \phi)\; p(\boldsymbol{\gamma}_i|\boldsymbol{\theta}^*)\; d\boldsymbol{\gamma}_i$$

where $\boldsymbol{\beta}$ denotes the fixed-effects, $\boldsymbol{\theta}$ and ϕ represent the parameters of the variance model, and $\boldsymbol{\gamma}_i$ represents the random effects for this subject. Adaptive quadrature is a way to approximate this multidimensional integral, with memory and computational requirements that rise exponentially with the dimension of the integral. This "curse of dimensionality" can make many multilevel models difficult or even impossible to fit currently in PROC GLIMMIX.

The FASTQUAD suboption for the METHOD=QUADRATURE option PROC GLIMMIX implements the multilevel quadrature algorithm of Pinheiro and Chao (2006). This algorithm reduces the single integral over many dimensions to a sum of integrals, each with fewer dimensions. With this option, you can now apply GLIMMIX to multilevel models that would have been too large or too slow to fit previously, and often do so in just seconds.

Consider the following example when some effect A has 4 levels:

```
proc glimmix method=quad(qpoints=5);
   class A id;
   model y = / dist=negbin;
   random A / subject=id;
run;
```

For each subject, computing the marginal log likelihood requires the numerical evaluation of a four-dimensional integral. With the number of quadrature points set to 5 by the QPOINT=5 option, this means that each marginal log likelihood evaluation requires $5^4 = 625$ conditional log likelihoods to be computed for each observation on each pass through the data. As the number of quadrature points or the number of random effects increases, this becomes a sizable computational effort. Consider adding one additional random effect B with two levels and including the interaction with A in the model:

```
proc glimmix method=quad(qpoints=5);
   class A B id;
   model y = / dist=negbin;
   random A A*B / subject=id;
run;
```

Now that single marginal likelihood calculation requires $5^{4+8} = 244,140,625$ conditional log likelihoods for each observation on each pass through the data.

You can reduce the dimension of the random effects by factoring out A from the two random effects:

```
proc glimmix method=quad(qpoints=5);
   class A B id;
   model y = / dist=negbin;
   random int B / subject=id*A;
run;
```

With random effects int and B, only $5^{(1+2)} = 125$ conditional log likelihoods need to be evaluated at each pass through the data.

This idea of reducing the dimension of random effects is the key to the multilevel Gaussian quadrature approximation algorithm described in Pinheiro and Chao (2006). By exploiting the sparseness in the random-effects design matrix Z, the multilevel quadrature algorithm reduces the dimension of the random effects to the sum of the dimensions of random effects from each level. The new FASTQUAD suboption in PROC GLIMMIX implements their method.

First, consider the following statements:

```
proc glimmix method=quad(qpoints=5);
   class A B id;
   model y = / dist=negbin;
   random A A*B B/ subject=id;
run;
```

In this case, it is not possible to factor a single SUBJECT= variable out of all the random effects. Formulated in this one-level way, a single evaluation of the marginal likelihood requires computing $5^{(4+8+2)} = 488,281,250$ conditional log likelihoods for each observation on each pass through the data.

To take advantage of the multilevel quadrature approximation, you use the FASTQUAD option and explicitly specify the two-level structure by including one RANDOM statement for each level:

```
proc glimmix method=quad(qpoints=5 fastquad);
   class A B id;
   model y = / dist=negbin;
   random B / subject=id;
   random int B / subject=id*A;
run;
```

The first RANDOM statement specifies random effect B for the level that corresponds to id; the second RANDOM statement specifies random effects int and B for the level that corresponds to id*A. With this specification, the multilevel quadrature approximation computes only $5^{(2+1+2)} = 3125$ conditional log likelihoods for each observation for each pass through the data, where $(2 + 1 + 2)$ is the sum of the number of random effects in the two RANDOM statements.

In general, consider a two-level model in which m level-2 units are nested within each level-1 unit. In this case, the one-level N_q point adaptive quadrature approximation to a marginal likelihood that is an integral over r_1 level-1 random effects and r_2 level random effects requires ${N_q}^{r_1 \times m + r_2}$ evaluations of the conditional likelihoods for each observation. The two-level adaptive quadrature approximation, however, requires only ${N_q}^{r_1 + r_2}$ evaluations of the conditional log likelihoods. By increasing exponentially with r_1 instead of with $r_1 \times m$, the multilevel quadrature algorithm reduces the computational and memory requirements significantly.

Missing Data

Missing data analysis has been a major SAS/STAT development direction in recent years. The MI and MIANALYZE procedures provide tools for producing and analyzing multiply imputed data sets; a recent update to PROC MI include the MNAR statement for facilitating sensitivity analysis by generating multiple imputations for different scenarios under the assumption that the data are missing not at random (MNAR). The weighted generalized estimating estimating equations (WGEE) approach to analyzing longitudinal data with dropouts was implemented in the GEE procedure in SAS/STAT 13.2, now updated and production in SAS/STAT 14.1.

Survey Data Analysis

The sample selection and survey data analysis software in SAS/STAT receives active development. In recent releases, plots have been added to the SURVEYREG and SURVEYMEANS procedures, kappa coefficients have been included in the SURVEYFREQ procedure, domain quantile estimates are now produced by the SURVEYMEANS procedure, and the tests for hypotheses have been updated in the SURVEYLOGISTIC procedure. In addition, the GLIMMIX procedure now fits weighted multilevel models, useful in analyzing survey data that comes from multistage sampling, where weights are often used to account for unequal sampling probabilities, nonresponse adjustments, and poststratification.

In the SAS/STAT 14.1 release, the new DF= option in the SURVEYLOGISTIC and the SURVEYPHREG procedures gives you access to a wider variety of adjusted and customized Wald tests. The replication variance estimates for quantiles produced by PROC SURVEYMEANS are now based on the Fuller (2009) approach, which produces estimates with better stability than the naive jackknife approach.

However, the major update to the survey software is the new SURVEYIMPUTE procedure for imputing missing data, as discussed below.

Missing Survey Data: Imputations

Nonresponse is a common problem in surveys. The resulting estimators suffer from nonresponse bias if the nonrespondents are different from the respondents. Estimators that use complete cases (only the observed units) might also be less precise. Imputation can reduce nonresponse bias, and by producing an imputed data set, result in consistent analyses. Imputation for survey data also means that you need to compute a variance estimator that accounts for both the sampling variance and the imputation variance.

Imputation techniques are based on either explicit or implicit models. Explicit model-based imputation techniques include multiple imputation, mean imputation, and regression imputation. Implicit techniques include hot-deck imputation, cold-deck imputation, and fractional imputation. Hot-deck imputation is the most commonly used imputation technique for survey data. A donor is selected for a recipient unit, and the observed values of the donor are imputed for the missing items of the recipient. Although the imputation method is straightforward, the variance estimator that accounts for imputation variance might not be simple and is often ignored in practice.

Fractional hot-deck imputation (Kalton and Kish 1984; Fay 1996; Kim and Fuller 2004; Fuller and Kim 2005), also known as fractional imputation (FI), is a variation of hot-deck imputation in which one missing item for a recipient is imputed from multiple donors. Each donor donates a fraction of the original weight of the recipient such that the sum of the fractional weights from all the donors is equal to the original weight of the recipient. For fully efficient fractional imputation (FEFI), all observed values in an imputation cell are used as donors for a recipient unit in that cell (Kim and Fuller 2004).

The new SURVEYIMPUTE procedure implements single and multiple hot-deck imputation and FEFI. PROC SURVEYIMPUTE imputes missing values of an item in a data set by replacing them with observed values from the same item. Available donor selection techniques include simple random selection with or without replacement, probability proportional to weights selection (Rao and Shao 1992), and approximate Bayesian bootstrap selection (Rubin and Schenker 1986). With FEFI, PROC SURVEYIMPUTE also produces imputation-adjusted replicate weights that can be used with any survey analysis procedure in SAS/STAT to estimate both the sampling variability and the imputation variability. (PROC SURVEYIMPUTE does not create imputation-adjusted replicate weights for hot-deck imputation.)

Creating an Imputed Data Set with PROC SURVEYIMPUTE

A software company conducted a survey of school personnel who use their Student Information System (SIS). A probability sample of SIS users was selected, including SIS users at middle schools and high schools in three southern

states. A two-stage stratified design was used, with a first-stage sample of schools (PSUs) selected from schools that use the SIS. The list of schools was stratified by state and by school status (new or renewal). Within the strata, schools were selected with probability proportional to size and with replacement, where the size measure was school enrollment. Five users were randomly selected with replacement from the second-stage units to complete the SIS satisfaction questionnaire.

The data set SIS_Survey_Sub contains the data and the sample design information. Variables include the school state, status of school, school, sampling weight, and department (whether the user was a teacher or administrator). The response ranges from 1 for very unsatisfied to 5 for very satisfied. Missing data occur for department, response, and both department and response in some cases.

The following statements request imputation of missing values in the SIS_Survey_Sub data set with the fully efficient fractional imputation method (FEFI):

```
proc surveyimpute data=SIS_Survey_Sub method=fefi varmethod=jackknife;
   class Department Response;
   var Department Response;
   strata State NewUser;
   cluster School;
   weight SamplingWeight;
   output out=SIS_Survey_Imputed outjkcoefs=SIS_JKCoefs;
run;
```

The METHOD=FEFI option in the SURYEYIMPUTE statement requests the fully efficient fractional imputation method, and the VARMETHOD= option requests imputation-adjusted jackknife replicate weights. The STRATA, CLUSTER, and WEIGHT statements specify the strata, clusters (PSUs), and weight variables. The VAR statement specifies the variables to be imputed (**Department** and **Response**).

By default, the variables **Department** and **Response** are imputed jointly. Therefore, the missing values for **Department** will be imputed conditionally on the observed levels of **Response**, and the missing values for **Response** will be imputed conditionally on the observed levels of **Department**. Observations that contain missing values for both **Department** and **Response** will be imputed by using the joint observed levels of **Department** and **Response**.

The OUTPUT statement names the SAS data set to which the imputed data are stored, and the OUTJKCOEFS= option in the OUTPUT statement names the SAS data set in which to save the jackknife coefficients.

Figure 1 displays summary information. There are six strata and 47 clusters. A total of 235 observations were used, which represents a population size of 6468.

Figure 1 Summary Information

Imputation Information	
Data Set	WORK.SIS_SURVEY_SUB
Weight Variable	SamplingWeight
Stratum Variables	State
	NewUser
Cluster Variable	School
Imputation Method	FEFI

Number of Observations Read	235
Number of Observations Used	235
Sum of Weights Read	6468
Sum of Weights Used	6468

Figure 1 *continued*

Design Summary	
Number of Strata	6
Number of Clusters	47

The "Missing Data Patterns" table in Figure 2 lists distinct missing data patterns along with their corresponding frequencies and weighted percentages.

Figure 2 Missing Data Patterns

Missing Data Patterns						
Group	**Department**	**Response**	**Freq**	**Sum of Weights**	**Unweighted Percent**	**Weighted Percent**
1	X	X	157	4272	66.81	66.05
2	X	.	52	1480	22.13	22.88
3	.	X	21	586	8.94	9.06
4	.	.	5	130	2.13	2.01

Missing Data Patterns							
	Group Means						
Group	**Department 0**	**Department 1**	**Response 1**	**Response 2**	**Response 3**	**Response 4**	**Response 5**
1	0.440309	0.559691	0.184457	0.206695	0.265684	0.209738	0.133427
2	0.641892	0.358108	.	.	.	.	.
3	.	.	0.261092	0.235495	0.230375	0.085324	0.187713
4	.	.	.	.	.	.	.

Five respondents have unit nonresponse (both variables in the VAR statement contain missing values), 73 respondents have item nonresponse (only one variable in the VAR statement contains a missing value), and 157 respondents have complete response (no variables in the VAR statement contain missing values). The estimated percentages in the sample for unit nonresponse, item nonresponse, and complete response are 2.1%, 31.1%, and 66.8%, respectively.

The "Imputation Summary" table in Figure 3 lists the number of nonmissing observations, missing observations, and imputed observations.

Figure 3 Imputation Summary

Imputation Summary		
Observation Status	**Number of Observations**	**Sum of Weights**
Nonmissing	157	4272
Missing	78	2196
Missing, Imputed	78	2196
Missing, Not Imputed	0	0
Missing, Partially Imputed	0	0

There are 78 observations that have missing values for at least one variable, and all 78 missing observations are imputed.

The output data set SIS_Survey_Imputed contains the observed data and the imputed values for **Department** and **Response**. In addition, this data set contains the imputation-adjusted full-sample weight (**ImpWt**), observation unit identification (**UnitId**), recipient index (**Recipient**), and imputation-adjusted jackknife replicate weights (**ImpRepWt_1**, . . . , **ImpRepWt_47**).

Suppose you want to compute frequency tables by using the imputed data set. The following statements request one-way tables for **Department** and **Response**. The analyses include the imputed values and account for both the design variance and the imputation variance.

```
proc surveyfreq data=SIS_Survey_Imputed varmethod=jackknife;
   format department Deptcode. response ResponseCode. ;
   table department response  department*response;
   weight ImpWt;
   repweights ImpRepWt: / jkcoefs=SIS_JKCoefs;
run;
```

The DATA= option in the PROC SURVEYFREQ statement specifies the input data set for analysis, SIS_Survey_Imputed, which contains the observed values and the imputed values for **Department** and **Response**. The FEFI technique uses multiple donor cells for a missing item. Therefore, the number of rows in the SIS_Survey_Imputed data set is greater than the number of rows in the observed data set, SIS_Survey_Sub. Therefore, it is very important to use only the weighted statistics from SIS_Survey_Imputed. The WEIGHT statement specifies the weight variable **ImpWt**, which is adjusted for the FEFI method. The imputation-adjusted jackknife replicate weights are saved in the variables **ImpRepWt_1**, . . . , **ImpRepWt_47** in the SIS_Survey_Imputed data set. The REPWEIGHTS statement names the replicate weight variables and the jackknife coefficients data set, SIS_JKCOEFS.

Figure 4 displays summary information. Note that the sum of weights in Figure 4 matches the sum of weights read from Figure 1, but the number of observations in Figure 4 (509) does not match the number of observations from Figure 1 (235). The FEFI technique uses multiple donor cells for a missing item. Therefore, the number of rows in the SIS_Survey_Imputed data set is greater than the number of rows in the observed data set, SIS_Survey_Sub. The sum of weights from both PROC SURVEYIMPUTE and PROC SURVEYFREQ represents the population size. The number of observations in Figure 1 represents the number of observation units, but the number of observations in Figure 4 represents the number of rows in the data set that include the observed units and the imputed rows. The number of replicates is 47, which is the same as the number of schools (PSUs).

Figure 4 One-Way Table

Data Summary	
Number of Observations	509
Sum of Weights	6468

Figure 4 *continued*

Variance Estimation	
Method	Jackknife
Replicate Weights	SIS_SURVEY_IMPUTED
Number of Replicates	47

Figure 5 displays one-way tables for **Department** and **Response**.

Figure 5 One-Way Table

Table of Department					
Department	**Frequency**	**Weighted Frequency**	**Std Err of Wgt Freq**	**Percent**	**Std Err of Percent**
Faculty	278	3200	429.52229	49.4729	6.6407
Admin/Guidance	231	3268	429.52229	50.5271	6.6407
Total	509	6468	5.3369E-11	100.000	

Table of Response					
Response	**Frequency**	**Weighted Frequency**	**Std Err of Wgt Freq**	**Percent**	**Std Err of Percent**
Very Unsatisfied	100	1256	291.92305	19.4153	4.5133
Unsatisfied	103	1371	361.02585	21.1976	5.5817
Neutral	112	1710	305.26968	26.4371	4.7197
Satisfied	100	1213	283.69298	18.7598	4.3861
Very Satisfied	94	917.82544	243.87967	14.1903	3.7706
Total	509	6468	3.2868E-11	100.000	

The Frequency column represents the number of rows in the SIS_Survey_Imputed data set, which also contains the imputed rows, and not the number of observations in the SIS_Survey_Sub data set. The Weighted Frequency, Std Err of Wgt Freq, Percent, and Std Err of Percent columns use the imputation-adjusted full-sample weight and replicate weights. You should use the weighted statistics from these columns. For example, an estimated 49.47% of SIS users are teachers, with a standard error of 6.64%. An estimate of "Very Satisfied" users is 14.19%, with a standard error of 3.77%.

Categorical Data Analysis

Categorical data analysis applies when responses are binary, ordinal, nominal, or discrete counts. Analysis strategies include both assessing the association of a response and explanatory variables and modeling that association with a statistical model such as logistic regression or Poisson regression. Many SAS/STAT procedures analyze categorical data. Recent additions include the availability of the HPGENSELECT procedure for model selection for generalized linear models (the type that can be fit with the GENMOD procedure), the addition of the UNEQUALSLOPES option in the LOGISTIC procedure so that partial proportional odds models can be fitted, score confidence limits for the odds ratio and a variety of binomial confidence limits in PROC FREQ, and the addition of the Tweedie distribution in the GENMOD procedure.

In SAS/STAT 14.1, the FREQ procedure adds exact mid-p, likelihood ratio, and Wald modified confidence limits for the odds ratio, as well as noninferiority, equivalence, and equality tests for the relative risk. The GENMOD procedure

provides a plot to assess the level of overdispersion in your data. And the GEE procedure, besides being production, is updated with the ESTIMATE, LSMEANS, and OUTPUT statement and offers alternating logistic regression analysis for both binary and ordinal data.

In the LOGISTIC procedure, the EQUALSLOPES and UNEQUALSLOPES options are available for all polytomous responses. You can now produce ROC curves and compute the AUC by using cross validated predicted probabilities, and the new ORPVALUE option displays p-values for odds ratios. Also, PROC LOGISTIC now fits the adjacent-category logit model to ordinal response data.

Fitting the Adjacent-Category Model with PROC LOGISTIC

If you have more than two response levels and these levels have an inherent ordering, PROC LOGISTIC fits a proportional odds model, which uses a cumulative logit link (CLOGIT) and a common set of slopes across all the response functions. Results from this model compare the higher response levels to the lower levels. However, you might have data for which you want to compare two consecutive response levels at a time. In SAS/STAT 14.1, the adjacent-category logit link (ALOGIT) is added to enable this analysis.

The Asbestos data set (Simonoff 2003, Section 10.2), which can be accessed from StatLib, contains data that measures worker exposure to asbestos grouped by the type of work being performed and the method of ventilation in their environment. In the following program, the response variable **Exposure** takes the values 'low' for low exposure, 'legal' for exposure near the legal limit, and 'high' for exposure above the legal limit. The classification variable **Task** takes the values 'tile' for tile removal and 'ins' for insulation removal, and the variable **Ventilation** is 'npv' for a negative pressure system and 'fan' for a general system. The **Freq** variable contains the number of workers in each group.

```
data Asbestos;
   input Task $ Ventilation $ Exposure $ Freq @@;
   datalines;
tile npv low 29   tile npv legal 1   tile npv high  1
tile fan low  3   tile fan legal 1   tile fan high  2
ins  npv low 10   ins  npv legal 1   ins  npv high  7
ins  fan low  3   ins  fan legal 3   ins  fan high 22
;
```

The default proportional odds model is fit to the the Asbestos data by using the following program:

```
proc logistic data=Asbestos;
   freq Freq;
   class Task Ventilation / param=ref;
   model Exposure = Task Ventilation;
run;
```

Figure 6 displays the response levels, and the note means that there are two response functions: the first compares 'high' to the two lower levels; the second compares 'high' and 'legal' to 'low'.

Figure 6 Response Information

Response Profile		
Ordered Value	**Exposure**	**Total Frequency**
1	high	32
2	legal	6
3	low	45

Probabilities modeled are cumulated over the lower Ordered Values.

Figure 7 shows that the common slope parameters are significant.

Figure 7 Parameter Estimates

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	**high**	1	-3.0208	0.6034	25.0599	<.0001
Intercept	**legal**	1	-2.4751	0.5633	19.3075	<.0001
Task	**ins**	1	2.2870	0.6182	13.6857	0.0002
Ventilation	**fan**	1	2.1594	0.5653	14.5921	0.0001

Because this is a cumulative logit model, it follows that the odds ratios are expressed in terms of higher exposure versus lower exposure. Figure 8 shows that a person who removes insulation has 9.8 times the odds of a higher exposure than a person removing tile, and a person using a fan has 8.7 times the odds of higher exposure than a person using a proper ventilation system. The odds ratios are the same if you compare the high level of exposure to the lower levels of exposure or the high and legal levels to the lower level.

Figure 8 Odds Ratios

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Task ins vs tile	9.845	2.931	33.071
Ventilation fan vs npv	8.666	2.862	26.243

The same main effects model is now fit for adjacent-category logits by specifying the LINK=ALOGIT option in the MODEL statement:

```
proc logistic data=Asbestos;
   freq Freq;
   class Task Ventilation / param=ref;
   model Exposure = Task Ventilation / link=alogit;
run;
```

The note in Output 9 means that there are two response functions: the first compares 'high' to 'legal'; the second compares 'legal' to 'low'.

Figure 9 Response Information

Response Profile		
Ordered Value	**Exposure**	**Total Frequency**
1	high	32
2	legal	6
3	low	45

Logits modeled use the response level with the next higher Ordered Value as the reference category.

Output 10 shows that the common slope parameters are still significant. The estimates from the cumulative logit model are larger than those from the adjacent-category logit model because the estimates refer to the entire exposure scale and not just the adjacent exposures. Agresti (2010) notes, however, that the standardized estimates (the estimates

divided by their standard error) are usually similar, and hence neither model has greater power to detect significant effects. Choice of model depends on interpretation preference.

Figure 10 Parameter Estimates

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	**high**	1	-0.1791	0.6372	0.0790	0.7786
Intercept	**legal**	1	-2.9970	0.4916	37.1614	<.0001
Task	**ins**	1	1.3481	0.3712	13.1880	0.0003
Ventilation	**fan**	1	1.2424	0.3347	13.7786	0.0002

From the odds ratios displayed in Figure 11, you can see that a person who removes insulation has 3.850 times the odds of the next higher exposure than a person removing tile, and a person using a fan has 3.464 times the odds of the next higher exposure than a person working in a properly ventilated area. The comparisons are for high to legal exposure level and for legal to low exposure level.

Figure 11 Odds Ratios

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Task ins vs tile	3.850	1.860	7.970
Ventilation fan vs npv	3.464	1.798	6.676

You might want to relax the common-slopes assumption. If you fit a cumulative logit with unequal slopes, you sometimes cannot maintain the increasing nature of the logits; even if you obtain a good fit, you may produce negative predicted probabilities when you score new data. However, adjacent-category logit models have an advantage in that these relaxed models are always valid and produce predicted probabilities that fall in [0,1].

In the following program, the UNEQUALSLOPES option is specified to remove the common slope assumption from the **Ventilation** effect. This means that slopes will be estimated for each of the adjacent logits for the ventilation effect.

```
proc logistic data=Asbestos;
   freq Freq;
   class Task Ventilation / param=ref;
   model Exposure = Task Ventilation / link=alogit unequalslopes=Ventilation;
run;
```

The odds ratios in Output 12 show that a person using a fan instead of ventilation has 9.841 times the odds of reaching the legal limit of exposure versus low exposure. However, there is no such significant effect for high versus legal exposure.

Figure 12 Odds Ratios

Odds Ratio Estimates				
Effect	**Exposure**	**Point Estimate**	**95% Wald Confidence Limits**	
Task ins vs tile		3.824	1.843	7.935
Ventilation fan vs npv	**high**	1.245	0.188	8.267
Ventilation fan vs npv	**legal**	9.841	1.417	68.324

Classification and Regression Trees

Classification and regression trees construct predictive models. These techniques are heavily used in data mining, but they are also used in standard statistical practice with one benefit being the ease of explaining the results to clients. Classification trees predict a categorical response while regression trees predict a continuous response. Tree models partition the data into segments called nodes by applying splitting rules, which assign an observation to a node based on the value of one of the predictors. The partitioning is done recursively, starting with the root node that contains all the data, continuing down to the terminal nodes, which are called leaves. The resulting tree model typically fits the training data well, but might not necessarily fit new data well. To prevent overfitting, a pruning method can be applied to find a smaller subtree that balances the goals of fitting both the training data and new data. The subtree that best accomplishes this is determined by using validation data or cross validation.

The partitioning can be represented graphically with a decision tree, which provides an accessible interpretation of the resulting model.

The HPSPLIT procedure creates a classification or regression tree model. It is a high-performance procedure, which means that it can be run in distributed mode with a SAS High-Performance Statistics product license. Otherwise, it runs in single-machine mode.

The HPSPLIT procedure provides:

- choices of algorithms for both classification and regression tree growth and pruning
- a variety of options for handling missing values
- whole and partial tree plots, cross validation plots, ROC curves, and partial tree plots

You can also produce an output data set with node and leaf assignments, predicted levels and posterior probabilities for a classification tree, and predicted response values for a regression tree

Fitting a Classification Tree with PROC HPSPLIT

To see the HPSPLIT procedure in action, consider data that are measurements of 13 chemical attributes for 178 samples of wine. Each wine is derived from one of three cultivars that are grown in the same area of Italy, and the goal of the analysis is a model that classifies samples into cultivar groups. The data are available in the UCI Irvine Machine Learning Repository; see Bache and Lichman (2013).

Figure 13 lists the first 10 observations of Wine.

Figure 13 Partial Listing of Wine

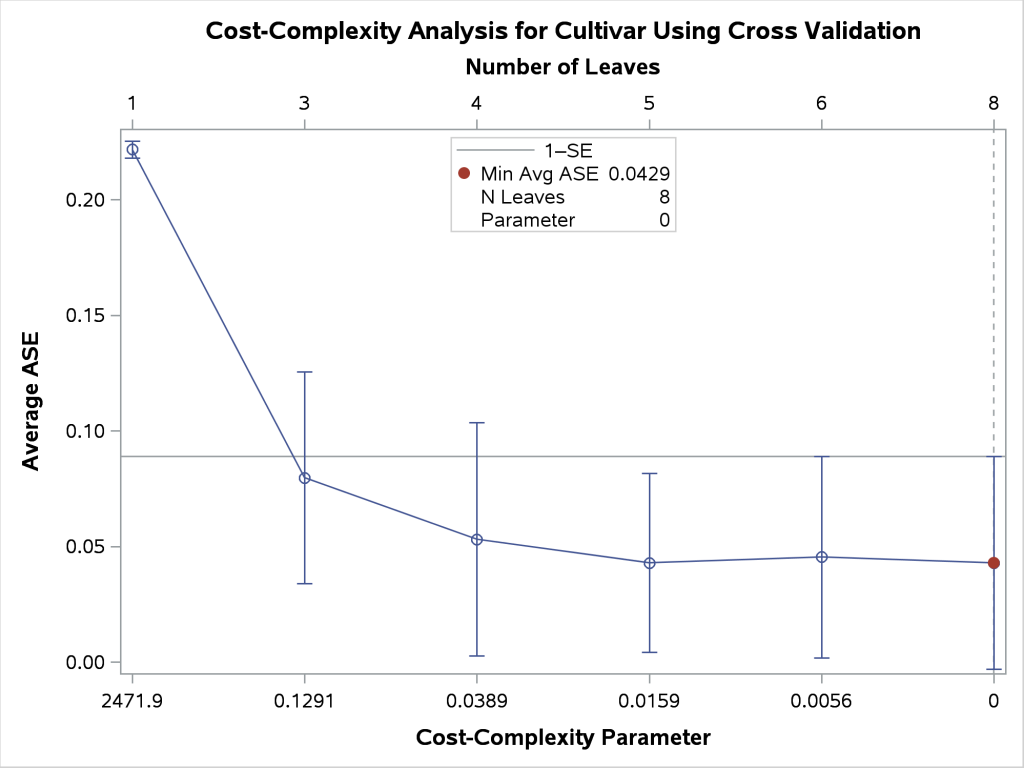
Obs	Cultivar	Alcohol	Malic	Ash	Alcan	Mg	TotPhen	Flav	NFPhen	Cyanins	Color	Hue	ODRatio	Proline
1	1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.64	1.04	3.92	1065
2	1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050
3	1	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185
4	1	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480
5	1	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82	4.32	1.04	2.93	735
6	1	14.20	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.05	2.85	1450
7	1	14.39	1.87	2.45	14.6	96	2.50	2.52	0.30	1.98	5.25	1.02	3.58	1290
8	1	14.06	2.15	2.61	17.6	121	2.60	2.51	0.31	1.25	5.05	1.06	3.58	1295
9	1	14.83	1.64	2.17	14.0	97	2.80	2.98	0.29	1.98	5.20	1.08	2.85	1045
10	1	13.86	1.35	2.27	16.0	98	2.98	3.15	0.22	1.85	7.22	1.01	3.55	1045

The variable **Cultivar** is a nominal categorical variable with levels 1, 2, and 3, and the 13 attribute variables are continuous. The following statements use the HPSPLIT procedure to create a classification tree.

```
ods graphics on;
proc hpsplit data=Wine seed=15531;
   class Cultivar;
   model Cultivar = Alcohol Malic Ash Alcan Mg TotPhen Flav
                    NFPhen Cyanins Color Hue ODRatio Proline;
   grow entropy;
   prune costcomplexity;
run;
```

The MODEL statement specifies **Cultivar** as the response variable, and the other variables are the predictor variables. The inclusion of **Cultivar** in the CLASS statement means that a classification tree model will be fit. The GROW statement specifies that the default method, entropy, is to be used for growing the tree. The PRUNE statement specifies that the default pruning method, cross-complexity, is to be used for pruning. The resulting cost-complexity plot is displayed in Figure 14.

Figure 14 Cost-complexity Plot



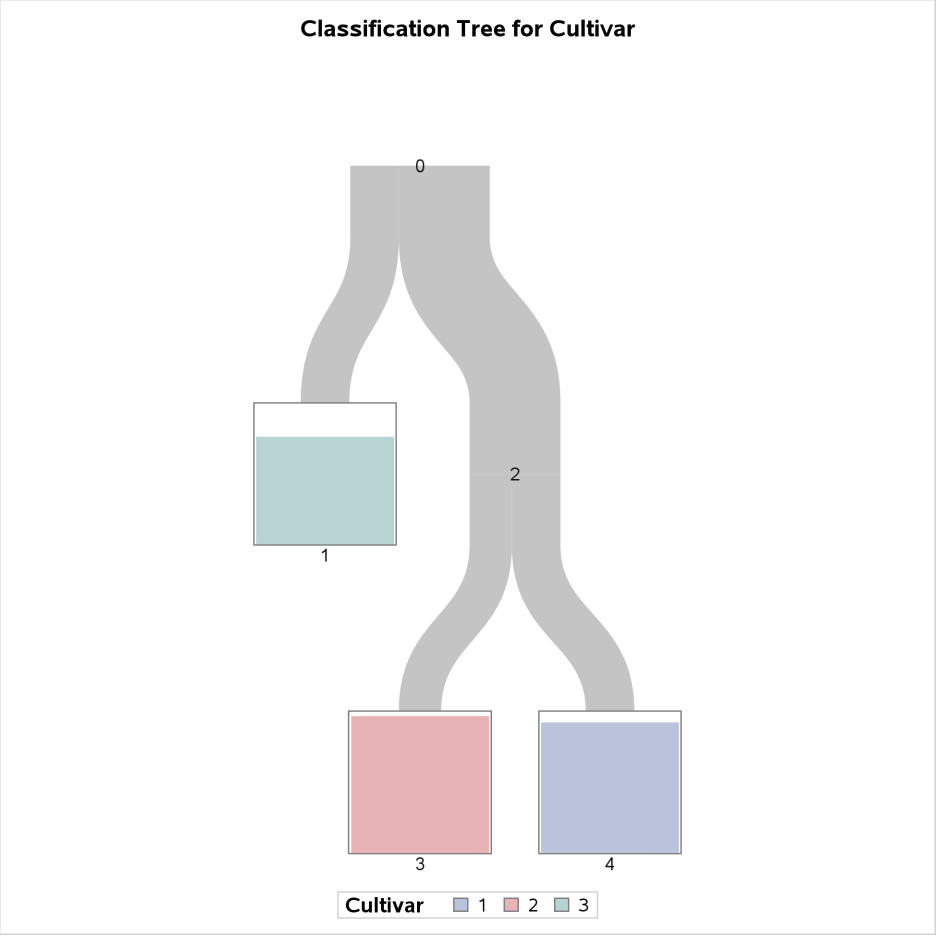
The plot shows that the procedure selects the parameter value 0, corresponding to 8 leaves. It uses Breiman's 1-SE rule, which chooses the parameter that corresponds to the smallest subtree for which the predicted error is less than one standard error above the minimum estimated ASE; see Breiman et al. (1984). However, this plot shows that the parameter value of 0.1291 is also reasonable, and it corresponds to a small tree tree with only three leaves.

The following statements build a classification tree by growing a large tree and applying cost-complexity pruning (also known as weakest link pruning) to obtain a tree with three leaves:

```
proc hpsplit data=Wine seed=15531;
   class Cultivar;
   model Cultivar = Alcohol Malic Ash Alcan Mg TotPhen Flav
                    NFPhen Cyanins Color Hue ODRatio Proline;
   prune costcomplexity(leaves=3);
run;
```

The tree diagram in Figure 15 is produced by default, and it provides an overview of the tree as a classifier.

Figure 15 Overview Diagram of Final Tree



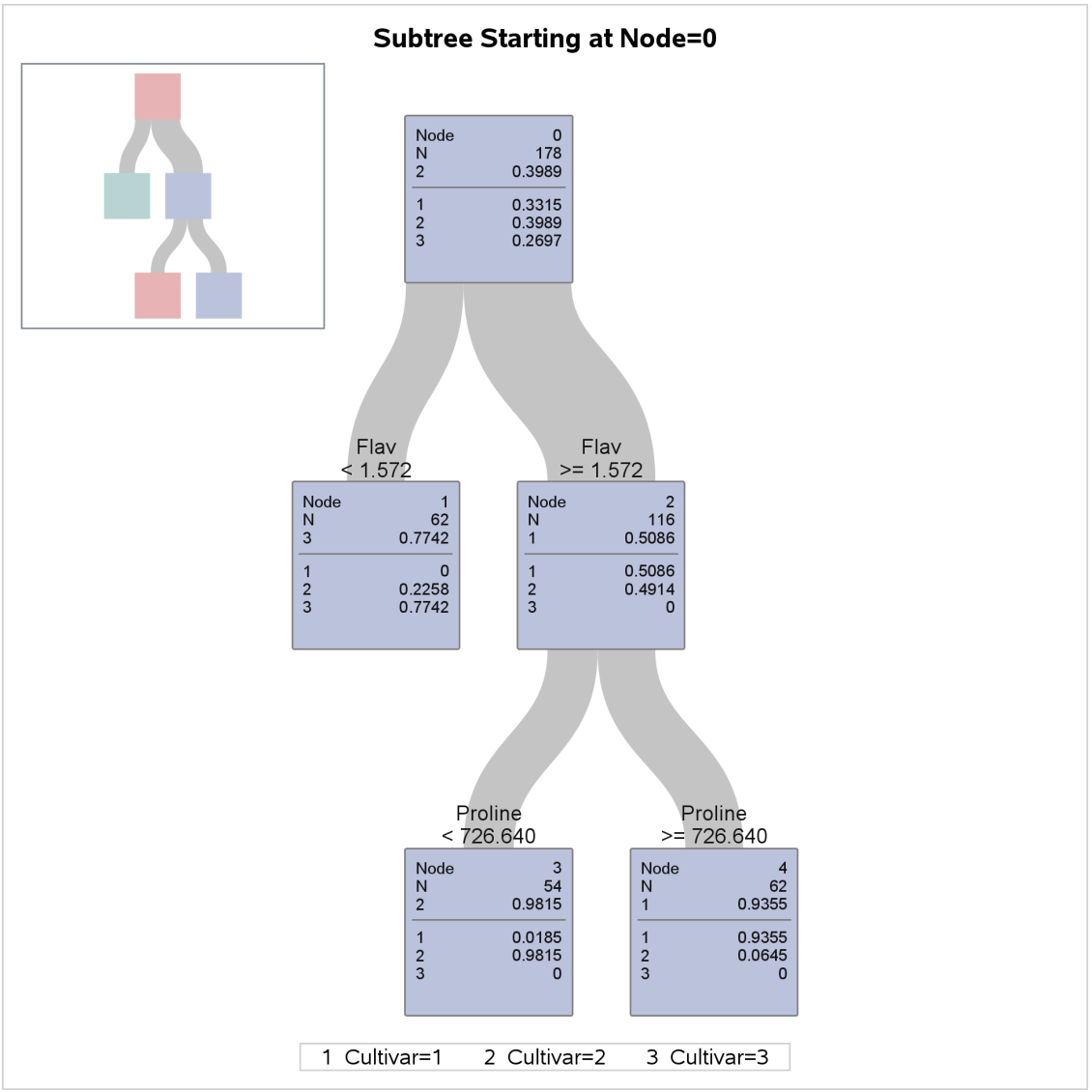
The tree is constructed starting with all of the observations in the root node (labeled 0). This node is split into a leaf node (1) and an internal node (2), which is further split into two leaf nodes (3 and 4).

The color of the bar in each leaf node indicates the most frequent level of **Cultivar** among the observations in that node; this is also the classification level assigned to all observations in that node. The height of the bar indicates the proportion of observations in the node that have the most frequent level. The width of the link between parent and child nodes is proportional to the number of observations in the child node.

The diagram reveals that splitting on just two of the attributes suffices to differentiate the three cultivars, and a tree model with only three leaves provides a high degree of accuracy for classification.

The diagram in Figure 16 provides more detail about the nodes and the splits.

Figure 16 Detailed Tree Diagram



There are 178 samples in the root node (node 0). The table below the line in the box for node 0 provides the proportion of samples for each level of **Cultivar**, and the level with the highest proportion is also given above the line. These samples are subdivided into 62 samples for which **Flav** < 1.572 (node 1) and 116 samples for which **Flav** ≥ 1.572 (node 2).

The variable **Flav** and the split point 1.572 are chosen to maximally decrease the impurity of the root node as measured by the entropy criterion. There are no samples with **Cultivar**=1 in node 1, and there are no samples with **Cultivar**=3 in node 2.

The samples in node 2 are further subdivided into 54 samples for which **Proline** < 726.640 (node 3) and 62 samples for which **Proline** ≥ 726.640 (node 4).

The classification tree yields simple rules for predicting the wine cultivar. For example, a sample for which **Flav** ≥ 1.572 and **Proline** < 726.640 is predicted to be from the third cultivar (**Cultivar**=3).

The diagram in Figure 16 happens to show the entire tree that was created by the preceding statements, but in general this diagram shows a subtree of the entire tree, which begins with the root node and has a depth of four levels. You can use the PLOTS=ZOOMEDTREE option to request diagrams that begin with other nodes and have specified depths.

The confusion matrix shown in Figure 17 evaluates the accuracy of the fitted tree for classifying the training data that were used to build the tree model.

Figure 17 Model-Based Confusion Matrix

Confusion Matrices					
		Predicted			
	Actual	1	2	3	Error Rate
Model Based	1	58	1	0	0.0169
	2	4	53	14	0.2535
	3	0	0	48	0.0000

The values in the off-diagonal entries of the matrices show how many times the fitted model misclassified a sample. For example, among the 59 samples with **Cultivar**=1, only one sample was misclassified, and it was incorrectly classified as **Cultivar**=2.

The "Tree Performance" table in Figure 18 displays various fit statistics for the tree model.

Figure 18 Fit Statistics for Tree Model

Fit Statistics for Selected Tree						
	N Leaves	ASE	Entropy	Gini	Mis-class	RSS
Model Based	3	0.0583	0.4290	0.1749	0.1067	31.1243

For example, the misclassification rate is the proportion of the 178 wine samples that were misclassified, so that is (1 + 4 + 14)/178 = 0.11.

Survival Analysis

Recent years have brought many new capabilities to the survival analysis tools in SAS/STAT software. The QUANTLIFE procedure provides two quantile regression approaches that account for right-censoring, are distribution-free, and are applicable to heteroscedastic data. The competing-risks model of Fine and Gray (1999) became available with the PHREG procedure, as well as additional tests and plots. The ICLIFETEST procedure performs nonparametric survival analysis for interval-censored data, and the ICPHREG procedure fits proportional hazards regression models to interval-censored data.

In SAS/STAT 14.1, the new FAILCODE option in the TIME statement in the PHREG procedure enables you to perform a nonparametric analysis of competing-risks data. PROC LIFETEST estimates the cumulative incidence function (CIF), which is the probability subdistribution of failure from a specific cause. If you have more than one sample of competing-risks data, PROC LIFETEST performs Gray's test (Gray 1988) to compare the CIFs of the samples. You can now fit stratified proportional hazards model with the ICPHREG procedure, as well as produce hazard plots and Lagakos and deviance residuals (Farrington 2000).

Models for Big Data

The heart of SAS/STAT has always been statistical modeling, and more recently, this has become a "Big Heart" with its attention to providing methods that are suitable for Big Data, whether that's defined as large numbers of variables or observations or both. Many procedures that are tools in the world of predictive modeling now have high-performance counterparts that work in a distributed computing environment, and additional techniques have been added, with many of them focused on modern model selection methods. The GLMSELECT procedure has been updated in recent releases with safe screening and sure independence screening methods that reduce the number of regressors to a

smaller subset from which model selection is performed; it has also been updated with the elastic net method, an extension of LASSO, as well as additional criteria for choosing models.

The HPGENSELECT procedure performs model selection for generalized linear models such as those produced by the GENMOD procedure. Introduced in SAS/STAT 13.2, PROC HPGENSELECT is a high-performance procedure that runs in single-machine mode, using the available cores, as well as in distributed mode, which requires a license for SAS High-Performance Statistics software.

In the 14.1 release, the group LASSO method becomes available with the GLMSELECT procedure. This method, a variant of LASSO, enables you to constrain groups of parameters to enter models together, such as the parameters that comprise the effect of a classification variable. Also, in this release, the HPGENSELECT procedure has been updated to include the LASSO method, support BY group processing, support the RESTRICT statement, and provide the minimum Schwarz Bayesian information criterion (SBC) for choosing a final model when you specify LASSO. As discussed previously, the HPSPLIT procedure is also a major tool for modeling Big Data, and it is updated in the 14.1 release with graphics and syntax that will appeal to both data scientists and statisticians.

The generalized additive model is another technique for modeling data with a large number of variables. STAT 14.1 introduces the new GAMPL procedure, discussed in the next section.

Model for Big Data: Generalized Additive Models

Generalized additive models are very flexible statistical models that are applicable to a number of fields. They can be thought of as extensions to generalized linear models where the response variable depends on predictors in both parametric and nonparametric ways. GAM models can handle response variables from many distributions, such as the normal, binomial, Poisson, gamma, and negative binomial distributions, and a link function transforms the mean of the response variable to the scale of the linear predictor. GAM models can include the usual linear predictors as well as nonlinear transformations of continuous variables with controlled smoothness.

The GAMPL procedure fits generalized additive models based on penalized likelihood methods (Wood 2006). Each spline term is constructed by the thin-plate regression spline technique (Wood 2003). A roughness penalty is applied to each spline term by a smoothing parameter that controls the balance between goodness of fit and the roughness of the spline curve. PROC GAMPL fits models for standard distributions in the exponential family, such as normal, Poisson, and gamma distributions. PROC GAMPL is a high-performance procedure that runs in either single-machine mode, where it exploits the available cores, or distributed mode, which requires the SAS High-Performance Statistics product license.

The GAMPL procedure:

- estimates the regression parameters of a generalized additive model that has fixed smoothing parameters by using penalized likelihood estimation
- estimates the smoothing parameters of a generalized additive model by using either the performance iteration method or the outer iteration method
- tests the total contribution of each spline term based on the Wald statistic
- enables you to construct a spline term by using multiple variables
- provides control options for constructing a spline term, such as fixed degrees of freedom, initial smoothing parameter, fixed smoothing parameter, smoothing parameter search range, user-supplied knot values, and so on
- produces graphs

Difference between the GAM and GAMPL Procedures

You might have used the GAM procedure in SAS/STAT for fitting generalized additive models. The GAMPL procedure provides a different approach to fitting generalized additive models, so you should think of these procedures as providing very different flavors of generalized additive models. The GAMPL procedure uses different approaches for constructing spline basis expansions, fitting generalized additive models, and testing smoothing components. The GAMPL procedure focuses on automatic smoothing parameter selection by using global model-evaluation criteria to determine optimal models. The GAM procedure focuses on constructing models by fitting partial residuals against each smoothing term. In general, you should not expect similar results from the two procedures.

Using PROC GAMPL to Fit a Nonparametric Logistic Regression Model

This example shows how to use the GAMPL procedure to build a nonparametric logistic regression model for a data set that contains a binary response and then use that model to classify observations. The data set used is relatively small in size, but the techniques are suitable for large data, too.

The Pima Indian Diabetes data set can be obtained from the UCI Machine Learning Repository (Asuncion and Newman 2007) and is extracted from a larger database originally owned by the National Institute of Diabetes and Digestive and Kidney Diseases. Data are for female patients who are at least 21 years old, are of Pima Indian heritage, and live near Phoenix, Arizona. The response variable is whether the subject has diabetes, and the explanatory variables represent physiological measurements and medical attributes.

The following DATA step creates the data set Pima with variables for number of pregnancies, glucose concentration, blood pressure, triceps skin fold, a diabetes pedigree function, and so on:

```
title 'Pima Indian Diabetes Study';
data Pima;
   input NPreg Glucose Pressure Triceps BMI Pedigree Age Diabetes Test@@;
   datalines;
 6  148  72  35  33.6  0.627  50  1  0    1   85  66  29  26.6  0.351  31  0  1
 1   89  66  23  28.1  0.167  21  0  0    3   78  50  32    31  0.248  26  1  0
 2  197  70  45  30.5  0.158  53  1  0    5  166  72  19  25.8  0.587  51  1  1
 0  118  84  47  45.8  0.551  31  1  0    1  103  30  38  43.3  0.183  33  0  0
 3  126  88  41  39.3  0.704  27  0  0    9  119  80  35    29  0.263  29  1  0
 1   97  66  15  23.2  0.487  22  0  1    5  109  75  26    36  0.546  60  0  0
 3   88  58  11  24.8  0.267  22  0  0   10  122  78  31  27.6  0.512  45  0  0
 4  103  60  33    24  0.966  33  0  1    9  102  76  37  32.9  0.665  46  1  1
 2   90  68  42  38.2  0.503  27  1  1    4  111  72  47  37.1   1.39  56  1  0
 3  180  64  25    34  0.271  26  0  1    7  106  92  18  22.7  0.235  48  0  0
 9  171 110  24  45.4  0.721  54  1  0    0  180  66  39    42  1.893  25  1  0
 ...
;
```

The **Test** variable splits the data set into training and test subsets. The training observations (whose **Test** value is 0) include approximately 59.4% of the data. To build a model that is based on the training data and evaluate its performance, the following DATA step generates missing responses for observations in the test subset:

```
data Pima;
   set Pima;
   Result = Diabetes;
   if Test=1 then Result=.;
run;
```

The following PROC HPLOGISTIC statements request a parametric logistic regression model on the training data and predict the test data. The PARTITION statement in the HPLOGISTIC procedure makes this straightforward. Otherwise, the statements are similar to what you might submit to PROC LOGISTIC or PROC GENMOD.

```
proc hplogistic data=Pima;
   model Diabetes(event='1') = NPreg Glucose Pressure Triceps
                               BMI Pedigree Age /dist=binomial;
   partition role=Test(test='1' train='0');
run;
```

Output 19 shows the summary statistics from the parametric logistic regression model.

Figure 19 Fit Statistics

Pima Indian Diabetes Study

Fit Statistics		
Description	**Training**	**Testing**
-2 Log Likelihood	293.11	178.54
AIC (smaller is better)	309.11	194.54
AICC (smaller is better)	309.58	195.24
BIC (smaller is better)	339.16	221.54

Output 20 shows fit statistics for both training and test subsets of the data, including the misclassification error for the test data.

Figure 20 Partition Fit Statistics

Partition Fit Statistics		
Statistic	**Training**	**Testing**
Area under the ROCC	0.8436	0.8815
Average Square Error	0.1526	0.1290
Hosmer-Lemeshow Test	0.4527	0.6312
Misclassification Error	0.2405	0.1852
R-Square	0.3111	0.3275
Max-rescaled R-Square	0.4277	0.4640
McFadden's R-Square	0.2866	0.3243
Mean Difference	0.3376	0.3836
Somers' D	0.6871	0.7630
True Negative Fraction	0.8627	0.8808
True Positive Fraction	0.5714	0.6615

The parametric logistic regression model is restricted in the sense that all variables affect the response in strictly linear fashion. If you are uncertain that a variable is an important factor and its contribution is linear in predicting the response, you might want to choose a nonparametric logistic regression model to fit the data. You can use PROC GAMPL to form a nonparametric model by including the spline transformation of each explanatory variable, as shown in the following statements:

```
proc gampl data=Pima seed=12345;
   model Result(event='1') = spline(NPreg)    spline(Glucose)
                             spline(Pressure) spline(Triceps)
                             spline(BMI)      spline(Pedigree)
                             spline(Age) / dist=binary;
run;
```

The EVENT= option specifically requests that PROC GAMPL model the probability of positive diabetes testing results. The spline transformations are specified directly in the MODEL statement. The “Response Profile” table in Figure 21 shows the frequencies of the response in both categories.

Figure 21 Response Profile

Pima Indian Diabetes Study

Response Profile		
Ordered Value	**Result**	**Total Frequency**
1	0	204
2	1	112

You are modeling the probability that Result='1'.

Figure 22 lists the summary statistics from the nonparametric logistic regression model, which include spline transformations of all variables. Compared to the parametric model, the information criteria values are generally lower.

Figure 22 Fit Statistics

Fit Statistics	
Penalized Log Likelihood	-130.14349
Roughness Penalty	3.98128
Effective Degrees of Freedom	19.96095
Effective Degrees of Freedom for Error	294.55701
AIC (smaller is better)	296.22759
AICC (smaller is better)	299.06383
BIC (smaller is better)	371.19577
UBRE (smaller is better)	-0.01689

The "Tests for Smoothing Components" table in Figure 23 shows approximate test results. Although some spline terms are significant, others are not. The null testing hypothesis is whether the total contribution from a variable is 0. So you can form a reduced model by removing those nonsignificant spline terms from the model. In this case, spline transformations for **NPreg, Pressure**, and **Triceps** are dropped from the model because their p-values are larger than the 0.1 nominal level.

Figure 23 Tests for Smoothing Components

Tests for Smoothing Components				
Component	**Effective DF**	**Effective DF for Test**	**Chi-Square**	**Pr > ChiSq**
Spline(NPreg)	1.93547	3	2.2274	0.5266
Spline(Glucose)	1.00000	1	43.5559	<.0001
Spline(Pressure)	1.00000	1	2.0574	0.1515
Spline(Triceps)	8.00000	8	4.6480	0.7944
Spline(BMI)	1.00000	1	3.1286	0.0769
Spline(Pedigree)	1.00000	1	5.2501	0.0219
Spline(Age)	5.02548	6	26.3779	0.0002

The following statements use PROC GAMPL to fit a reduced nonparametric logistic regression model. The OUTPUT statement requests predicted probabilities for both training and test data sets. The ID statement requests that the **Diabetes** and **Test** variables also be included in the output data set so that you can use them to identify test observations and compute misclassification errors.

```
proc gampl data=Pima plots seed=12345;
   model Result(event='1') = spline(Glucose)  spline(BMI)
                             spline(Pedigree) spline(Age) / dist=binary;
   output out=PimaOut;
   id Diabetes Test;
run;
```

Figure 24 shows the summary statistics from the reduced nonparametric logistic regression model.

Figure 24 Fit Statistics

Pima Indian Diabetes Study

Fit Statistics	
Penalized Log Likelihood	-135.29396
Roughness Penalty	2.88837
Effective Degrees of Freedom	8.97141
Effective Degrees of Freedom for Error	306.02206
AIC (smaller is better)	285.64236
AICC (smaller is better)	286.22700
BIC (smaller is better)	319.33665
UBRE (smaller is better)	-0.03383

In the "Estimates for Smoothing Components" table in Figure 25, PROC GAMPL reports that the effective degrees of freedom values for spline transformations of **Glucose**, **BMI**, and **Pedigree** are quite close to 1. This suggests linear forms for these three variables. For **Age**, the degrees of freedom value is nearly 5. These measures suggest nonlinear patterns in the dependency of the response on this variable.

Figure 25 Estimates for Smoothing Components

Estimates for Smoothing Components						
Component	**Effective DF**	**Smoothing Parameter**	**Roughness Penalty**	**Number of Parameters**	**Rank of Penalty Matrix**	**Number of Knots**
Spline(Glucose)	1.00000	3.903E45	3.79E-41	9	10	112
Spline(BMI)	1.00000	7.93E37	5.23E-35	9	10	186
Spline(Pedigree)	1.00000	1.048E15	5.21E-17	9	10	265
Spline(Age)	4.97141	38.2065	2.8884	9	10	42

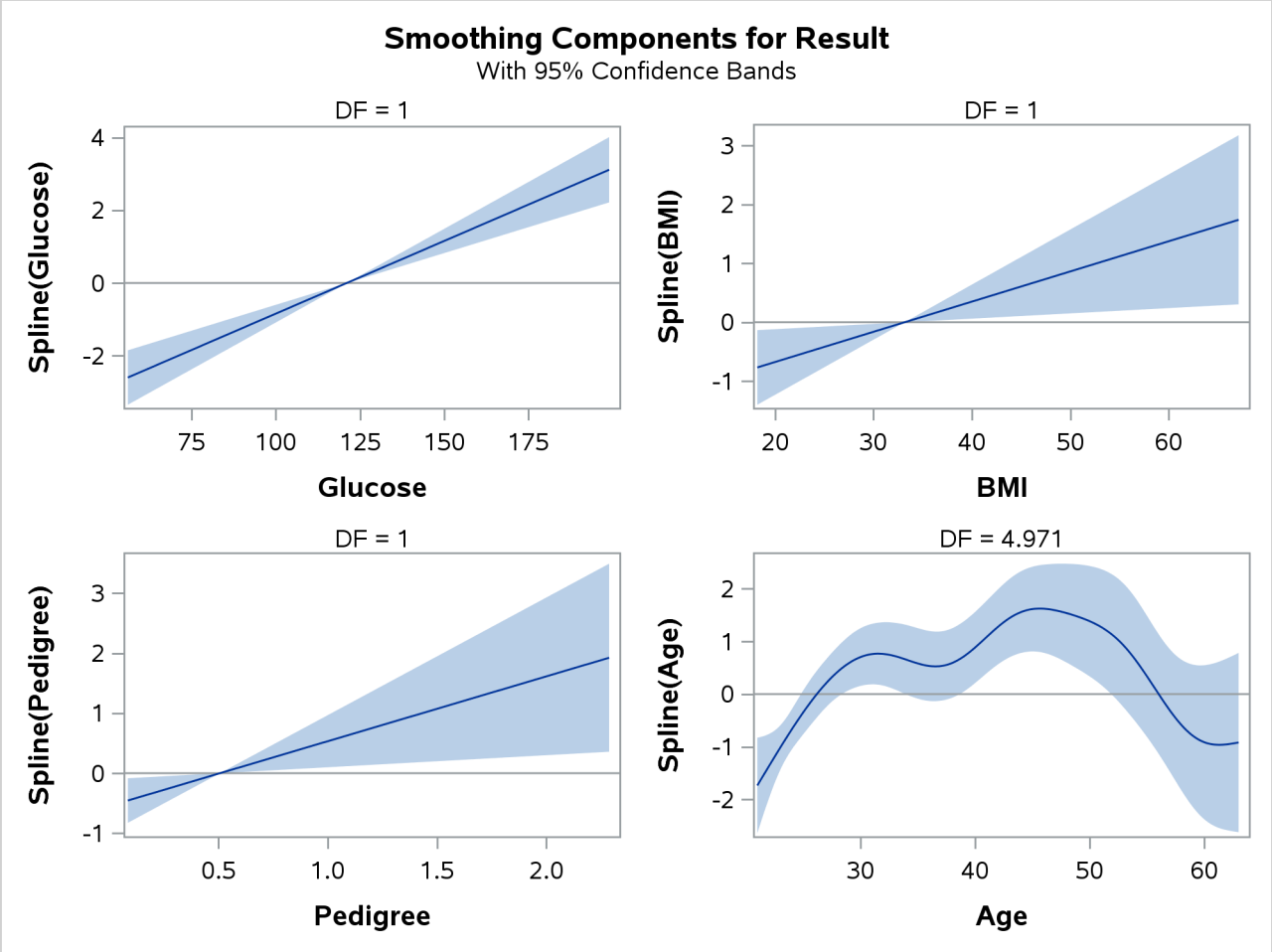
The "Tests for Smoothing Components" table in Figure 26 shows that all spline transformations are significant in predicting diabetes testing results.

Figure 26 Tests for Smoothing Components

Tests for Smoothing Components				
Component	**Effective DF**	**Effective DF for Test**	**Chi-Square**	**Pr > ChiSq**
Spline(Glucose)	1.00000	1	46.3311	<.0001
Spline(BMI)	1.00000	1	5.6408	0.0175
Spline(Pedigree)	1.00000	1	5.7940	0.0161
Spline(Age)	4.97141	6	31.6866	<.0001

The smoothing component panel, shown in Figure 27, visualizes the spline transformations for the four variables in addition to 95% Bayesian curvewise confidence bands. For **Glucose**, **BMI**, and **Pedigree**, the spline transformation is almost a straight line. For **Age**, the dependency is obviously nonlinear.

Figure 27 Smoothing Components Panel



You can then use the information in the output data set to compute the misclassification error with PROC FREQ (not shown here).

Bayesian Analysis

The Bayesian front at SAS has been growing steadily since Bayesian capabilities were provided in the GENMOD, LIFEREG, and PHREG procedures in the 9.2 release. Now, the FMM procedure for finite mixture models also includes Bayesian capabilities, the MCMC procedure provides full-service Bayesian modeling, and the BCHOICE procedure for Bayesian discrete choice models is the first SAS/STAT procedure that focuses solely on a Bayesian application. In the last several releases of SAS/STAT, the MCMC procedure became multithreaded; supports nested and non-nested hierarchical models to arbitrary depth; provides Bayesian solutions to model-based missing data analysis; supports categorical distribution for latent-variable modeling; and enables users to construct general prior distribution for random effects. The multithreaded BCHOICE procedure finetuned its sampling algorithms to make it run much faster for many models.

In the 14.1 release, the MCMC procedure has been updated with new sampling algorithms for continuous parameters: the Hamiltonian Monte Carlo (HMC) and the No-U-Turn Sampler (NUTS). These algorithms use Hamiltonian dynamics to enable distant proposal in the sampling, making them efficient in many scenarios. While their computational costs are high (gradients are required), these algorithms can lead to drastic improvements in sampling efficiency in many cases, resulting in fewer needed draws to achieve the same accuracy. PROC MCMC now support models that require lagging and leading variables, enabling you enables you to easily fit models such as dynamic linear models, state space models, and autoregressive models.

PROC MCMC also adds a general ODE solver (CALL ODE) that provides numerical solution to a system of user-defined ODEs, which can be piecewise, and a general multidimensional integration function (CALL QUAD) that finds integral of an arbitrary integrand function. These features facilitate the fitting of complex models, such as pharmacokinetic models and models that require integrated likelihood functions. The PREDDIST statement in PROC MCMC now supports prediction from a marginalized random-effects model, which enables more realistic and useful prediction from many models.

Odds and Ends

Depending on your focus area, the additional features added to SAS/STAT software can be the gem that makes your statistical life much easier. The following gems in SAS/STAT 14.1 might apply to you:

- The PLOTS=OVERDISP option in the PROC GENMOD statement produces plots that provide an idea of the overdispersion in zero-inflated models.
- The new STRATA statement in PROC NPAR1WAY provides stratified rank-based analysis of two-sample data. The following score types are available: Wilcoxon, median, normal (van der Waerden), Savage, and raw data scores. Rank-based scores can be computed by using within-stratum ranks or overall ranks; strata can be equally weighted or weighted by stratum size.
- You can use the Dirichlet-multinomial distribution as the response distribution for a mixture component in the FMM procedure.
- The generalized partial credit (GPC) model is available for ordinal items in the IRT procedure.
- The new DDFM=KENWARDROGER2 option in PROC MIXED applies the (prediction) standard error and degrees-of-freedom correction that are detailed by Kenward and Roger (2009). This correction reduces the precision estimator bias of the fixed and random effects under nonlinear covariance structures. It was first implemented in the GLIMMIX procedure.
- The METHOD=MLSB option in the PROC CALIS statement computes maximum likelihood estimates, Satorra-Bentler scaled chi-square statistics for model fit, and standard error estimates that are based on a sandwich-type formula proposed by Satorra and Bentler (1994). This estimation method is suitable when you apply the normal-theory-based maximum likelihood estimation to either normal or nonnormal data.
- The BCHOICE procedure allows varying numbers of alternatives in choice sets for logit models. The new RESTRICT statement enables you to specify boundary requirements and order constraints on fixed effects for logit models.
- The POWER procedure supports Cox proportional hazards regression models and Farrington-Manning noninferiority tests of relative risk.

Summary

SAS/STAT 14.1 provides important new capabilities in a number of areas. They include new directions such as generalized additive models based on penalized likelihood and survey data imputation, additional techniques for newer areas such as analyzing interval-censored survival data, and key updates for critical statistical modeling tools. Besides this paper, a good resource for getting more familiar with the new features in SAS/STAT 14.1 is the "What's New" chapter in the SAS/STAT documentation as well as the "Getting Started" and the "Examples" sections, from which several examples in this paper were drawn. The documentation can be found at support.sas.com/statdoc/, and various other resources are available on the Statistics and Operations Research focus area site at support.sas.com/statistics/.

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

Contributions

Robert E. Derr, Randy Tobias, Min Zhu, Pushpal Mukhopadhyay, Robert N. Rodriguez, Fang Chen and Weijie Cai contributed to this paper. Thanks go to Anne Baxter for editorial support and Tim Arnold for programming support.

Version 1.0

REFERENCES

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. 2nd ed. New York: John Wiley & Sons.

Asuncion, A., and Newman, D. J. (2007). "UCI Machine Learning Repository." http://archive.ics.uci.edu/ml/.

Bache, K., and Lichman, M. (2013). "UCI Machine Learning Repository." School of Information and Computer Sciences, University of California, Irvine. http://archive.ics.uci.edu/ml.

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

Farrington, C. P. (2000). "Residuals for Proportional Hazards Models with Interval-Censored Survival Data." *Biometrics* 56:473–482.

Fay, R. E. (1996). "Alternative Paradigms for the Analysis of Imputed Survey Data." *Journal of the American Statistical Association* 91:490–498.

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.

Forster, J. J., McDonald, J. W., and Smith, P. W. F. (2003). "Markov Chain Monte Carlo Exact Inference for Binomial and Multinomial Logistic Regression Models." *Statistics and Computing* 13:169–177.

Fuller, W. A. (2009). *Sampling Statistics*. Hoboken, NJ: John Wiley & Sons.

Fuller, W. A., and Kim, J. K. (2005). "Hot Deck Imputation for the Response Model." *Survey Methodology* 31:139–149.

Gray, R. J. (1988). "A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk." *Annals of Statistics* 16:1141–1154.

Kalton, G., and Kish, L. (1984). "Some Efficient Random Imputation Methods." *Communications in Statistics—Theory and Methods* 13:1919–1939.

Kenward, M. G., and Roger, J. H. (2009). "An Improved Approximation to the Precision of Fixed Effects from Restricted Maximum Likelihood." *Computational Statistics and Data Analysis* 53:2583–2595.

Kim, J. K., and Fuller, W. A. (2004). "Fractional Hot Deck Imputation." *Biometrika* 91:559–578.

Pinheiro, J. C., and Chao, E. C. (2006). "Efficient Laplacian and Adaptive Gaussian Quadrature Algorithms for Multilevel Generalized Linear Mixed Models." *Journal of Computational and Graphical Statistics* 15:58–81.

Rao, J. N. K., and Shao, J. (1992). "Jackknife Variance Estimation with Survey Data under Hot Deck Imputation." *Biometrika* 79:811–822.

Rubin, D. B., and Schenker, N. (1986). "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse." *Journal of the American Statistical Association* 81:366–374.

Satorra, A., and Bentler, P. M. (1994). "Corrections to Test Statistics and Standard Errors in Covariance Structure Analysis." In *Latent Variables Analysis: Applications for Developmental Research*, edited by A. von Eye, and C. C. Clogg, 399–419. Thousand Oaks, CA: Sage.

Simonoff, J. S. (2003). *Analyzing Categorical Data*. New York: Springer-Verlag.

Wood, S. (2003). "Thin Plate Regression Splines." *Journal of the Royal Statistical Society, Series B* 65:95–114.

Wood, S. (2006). *Generalized Additive Models*. Boca Raton, FL: Chapman & Hall/CRC.