Paper 432-2013

Current Directions in SAS/STAT® Software Development

Maura Stokes
SAS Institute Inc.

Abstract

Recent years brought you SAS/STAT® releases in rapid succession, and another release is targeted for 2013. Which new software features will make a difference in your work? What new statistical trends should you know about? This paper describes recent areas of development focus, such as Bayesian analysis, missing data analysis, postfitting inference, quantile modeling, finite mixture models, specialized survival analysis, and structural equation modeling. This paper introduces you to the concepts and illustrates them with practical examples.

More Frequent Releases of SAS/STAT Software

In previous years, SAS/STAT software was updated only when Base SAS® software was released, but SAS/STAT is now released independently of Base SAS along with other SAS analytical products. This means these products can be released when enhancements are ready, and the goal is to update SAS/STAT every 12 to 18 months. To mark this newfound independence, the release numbering scheme for SAS analytical products has also changed; the current production release is SAS/STAT 12.1.

With the new release paradigm, SAS provides you with new statistical techniques faster. However, it may be difficult to keep track of all of the new development. This paper provides a round-up of the current directions in SAS/STAT development, with emphasis on the areas that have received the most attention.

Developing SAS/STAT Software

SAS/STAT software has a long and rich legacy. Originally, statistical functionality was included in Base SAS for mainframes, but SAS/STAT became its own product in 1976. Originally written in the PL/I and Fortran programming languages, it was rewritten in the C language and made portable so it could run on a variety of platforms (multivendor architecture) including PCs. The Output Delivery System (ODS) was then incorporated, giving you complete control over the format of the tabular results and making reporting quite flexible. More recently, the development of ODS Statistical Graphics has resulted in the integration of graphics into the statistical analyses.

What hasn't changed over the years has been the goal of providing rich, up-to-date statistical techniques to SAS customers. SAS/STAT 12.1 includes over 75 procedures, soon to expand again with the 12.3 and 13.1 releases. New development directions are determined in a number of ways: customer input, company directives, the appearance of important new methodologies, and the drive to constantly refine and update existing software. Technical support is one channel from customers to development which fosters greater understanding of customer use and often leads to software enhancements. Other channels include professional contacts and customer meetings. Often, statistical development directions are influenced by overall SAS company directives, such as the current initiative to excel in high-performance computing.

Of course, statistical developers, specialized PhD statisticians with computing backgrounds, keep up with current methodology in the usual ways: journals, professional conferences, expert contacts, and workshops. Developers attend statistical conferences and SAS user groups, both sources of feedback and suggestions. When R&D engages in a new area, it often begins with an invited seminar by a renowned expert and then sustains the contact for feedback as development proceeds.

SAS development focuses on statistical methods that work in practice. SAS customers deal with real-life data, with its size and pitfalls, and the promising methodology that looked great for a few well-designed test cases often needs more work before it's ready for the real world. An important objective is consistency in syntax so that customers can build on what they already know. The goal is always clear and consistent

output, including graphics, and a major milestone for each new project is clearing a peer review of tabular and graphical content and layout. The documentation is written by the developers and undergoes substantial technical review and editing before it is finalized.

Numerical accuracy and computational performance are essential. Numerical analysts develop the mathematical routines that are incorporated into the statistical procedures. Existing procedures are continously evaluated to see if they can benefit from cutting-edge algorithms. SAS avoids changing default methods because the software is used so often in production jobs, but new methods are added regularly and the documentation states when a newer method has become the predominant practice.

The following sections outline some recent development directions, including model building, Bayesian analysis, linear model enrichment, analyzing time-to-event data, missing data methods, complex data, and high-performance computing.

Model Building

Model building has only grown more important in recent times as statisticians and data analysts face the world of Big Data and its inherent increases in data dimensionality. The value of predictive analytics depends on finding good models, so model selection is paramount. SAS/STAT model building took a step forward some years ago with the introduction of the GLMSELECT procedure for linear models selection. PROC GLMSELECT provides modern methods such as LARS and LASSO, and it also includes a rich array of selection criteria and numerous graphical displays. More recently, model building in SAS/STAT has expanded to include model selection for quantile regression and generalized linear models.

New QUANTSELECT Procedure

Quantile regression is a distribution-free method that determines a linear predictor for the conditional quantile. It is especially useful with data that are heterogeneous such that the tails and central location of the conditional distributions vary with the covariates. The QUANTREG procedure provides quantile regression in SAS/STAT software. Beginning with SAS/STAT 12.1, you can also perform model selection for quantile regression with the new QUANTSELECT procedure. This procedure provides capabilities similar to those offered by the GLMSELECT procedure, including:

- forward, backward, stepwise, and LASSO selection methods
- variable selection criteria: AIC, SBC, AICC, and so on
- variable selection for both quantiles and the quantile process
- the EFFECT statement for constructed model effects (splines)

PROC QUANTSELECT is multithreaded so that it can take advantage of multiple processors. It is very efficient and can handle hundreds of variables and thousands of observations. After you have selected a model with the QUANTSELECT procedure, you can proceed to use the QUANTREG procedure for final model analysis.

The following example illustrates the use of the QUANTSELECT procedure with baseball data from the 1986 season. The goal is to predict player salary. You can request model selection for any number of quantiles, and if you do so, you will find that different models are selected. If you are interested only in the model for those players making the most money, you can base the model on the 90th quantile, which is the analysis performed here.

The following statements input the baseball data:

```
data baseball;
   length name $ 18;
   length team $ 12;
   input name $ 1-18 nAtBat nHits nHome nRuns nRBI nBB
         yrMajor crAtBat crHits crHome crRuns crRbi crBB
         league $ division $ team $ position $ nOuts nAssts
         nError salary;
```

```
datalines;
Allanson, Andy        293    66     1    30    29    14
   1   293    66     1    30    29    14
American East Cleveland C 446 33 20 .
Ashby, Alan           315    81     7    24    38    39
  14  3449   835    69   321   414   375
National West Houston C 632 43 10 475
.....
.....
```

The following statements invoke the QUANTSELECT procedure. The variable SALARY is the response variable, and a number of explanatory variables are available for selection. The adaptive LASSO method is used for model selection, with AIC as the stopping criterion. The plot requested is the coefficient panel.

```
proc quantselect data=baseball plots=(coef);
  class league division;
  model Salary = nAtBat nHits nHome nRuns nRBI nBB
        yrMajor crAtBat crHits crHome crRuns crRbi
        crBB league division nOuts nAssts nError /
           selection=lasso (adaptive stop=aic)
  quantile=.9;
run;
```

Figure 1 displays the selection summary information. You can see the values of AIC and AICC change as variables are added to the model. The optimal value of AIC is 2099.23 at the fifth step, which corresponds to a model with four variables: number of hits, career hits, career home runs, and career RBIs. These explanatory variables are the main factors in determining salary for the 90th percentile.

Figure 1 Selection Summary

Quantile Level = 0.9

Selection Summary						
Step	Effect Entered	Number Effects In	AIC	AICC	SBC	Adjusted R1
0	Intercept	1	2436.7289	2436.7442	2440.3011	0.0000
1	crHits	2	2197.4349	2197.4811	2204.5792	0.3655
2	crRbi	3	2183.6148	2183.7075	2194.3313	0.3819
3	nHits	4	2113.2757	2113.4308	2127.5643	0.4593
4	crHome	5	2099.2203*	2099.4538*	2117.0811*	0.4735*
* Optimal Value Of Criterion						

Figure 2 displays the coefficient panel, which shows the progression of the standardized coefficients and the SBC throughout the selection process.

Figure 2 Coefficient Panel

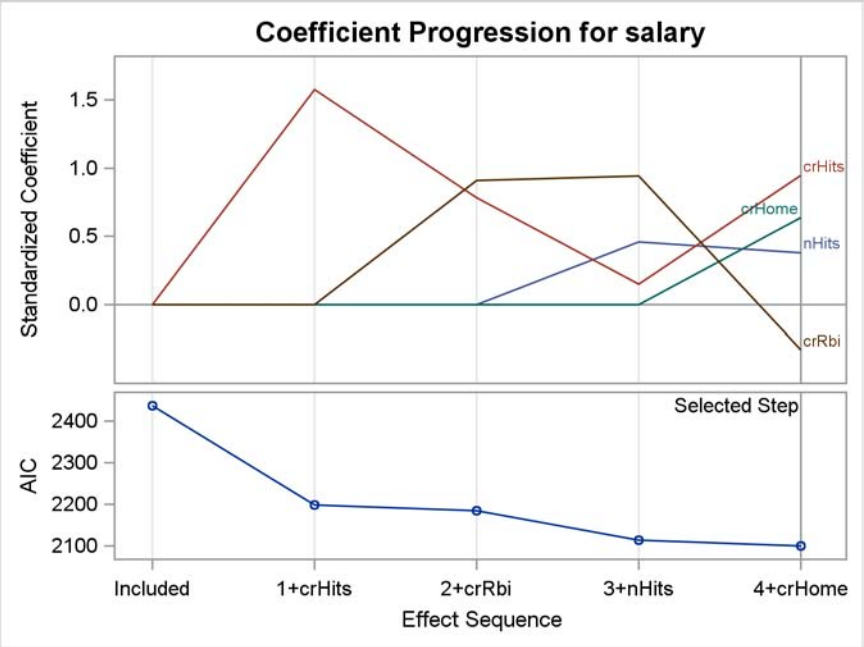


Figure 3 contains the parameter estimates and their standardized versions.

Figure 3 Parameter Estimates

Parameter Estimates			
Parameter	**DF**	**Estimate**	**Standardized Estimate**
Intercept	1	-100.381451	0
nHits	1	3.895310	0.379872
crHits	1	0.660121	0.949536
crHome	1	3.340492	0.637123
crRbi	1	-0.453248	-0.329973

Additional inference can be performed by using this model in the QUANTREG procedure.

New HPGENSELECT Procedure

The GENMOD procedure fits generalized linear models in SAS; these models include normal regression, logistic regression, Poisson regression, and other analyses where you specify a link function and a distribution that belongs to the exponential family. The new HPGENSELECT procedure, available with SAS/STAT 12.3 (which runs on Base 9.4), performs model selection for generalized linear models (GLMs). This means that you can now perform model selection for analyses such as Poisson regression, negative binomial regression, and any other GLM. Designed for the distributed computing of SAS® High-Performance Statistics, PROC HPGENSELECT also works in single-machine mode. It provides forward, backward, and stepwise selection (LASSO-type methods are still a research topic), and includes the AIC, SBC, and AICC selection criteria. It does not produce graphs with this version, but they will surface in a future release.

Recall the baseball data. You can treat the number of home runs hit during the year as counts that follow the Poisson distribution, and thus you can employ Poisson regression to model these counts. The following statements illustrate how you would request model selection for Poisson regression with the HPGENSELECT procedure.

```
proc hpgenselect data=baseball;
   class league division;
   model nHome = nAtBat nHits nRuns nRBI nBB
                 yrMajor crAtBat crHits crHome crRuns crRbi
                 crBB league division nOuts nAssts nError
              / distribution=poisson link=log;
   selection  method=forward details=all;
run;
```

You specify exactly the same MODEL statement as you would specify with the GENMOD procedure. You specify the selection method with the SELECTION statement, a new statement used by the high-performance procedures. Forward stepwise selection is specified with the METHOD=FORWARD option.

Figure 4 displays information about the execution mode. Two threads were employed on a single machine.

Figure 4 Performance Information

Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

Effects are added to the model if they produce improvement as judged by comparing the p-value of a score test to the entry significance level (SLE), which is 0.05 by default. The forward selection ends when no additional effect meets this criterion.

Figure 5 provides the final effects that entered the model and the details of effect selection.

Figure 5 Selection Details

Selection Details							
Step	**Description**	**Effects In Model**	**Chi-Square**	**Pr > ChiSq**	**-2 LogL**	**AIC**	**AICC**
0	Initial Model	1			3419.513	3421.513	3421.525
1	nRBI entered	2	1595.3565	<.0001	1994.778	1998.778	1998.815
2	nAssts entered	3	48.6818	<.0001	1943.370	1949.370	1949.445
3	nHits entered	4	20.2656	<.0001	1922.611	1930.611	1930.737
4	nRuns entered	5	43.3903	<.0001	1880.196	1890.196	1890.386
5	crHome entered	6	6.3802	0.0115	1873.922	1885.922	1886.189
6	crRbi entered	7	27.3765	<.0001	1845.695	1859.695	1860.052
7	crAtBat entered	8	6.9953	0.0082	1838.769	1854.769	1855.229
8	crRuns entered	9	13.0622	0.0003	1825.626	1843.626	1844.203
9	crHits entered	10	15.2906	<.0001	1810.307	1830.307	1831.014
10	crBB entered	11	4.7299	0.0296	1805.576	1827.576	1828.428

Selection Details	
Step	**BIC**
0	3425.287
1	2006.327
2	1960.693
3	1945.709
4	1909.069
5	1908.569
6	1886.117
7	1884.965
8	1877.597
9	1868.052
10	1869.096

Figure 6 contains fit statistics for the selected model. Note that the values of the −2 log likelihood and the various information criteria AIC, BIC, and AICC are smaller than the corresponding values for the full model (Initial Model) in the Selection Details table, which indicates that the more parsimonious model provides a better fit. However, note that the value of Pearson chi-square divided by its degrees of freedom is 1.60. This is an indication of overdispersion, which suggests that Poission regression is not the best technique to apply to these data.

Figure 6 Fit Statistics

Fit Statistics	
-2 Log Likelihood	1805.57637
AIC (smaller is better)	1827.57637
AICC (smaller is better)	1828.42799
BIC (smaller is better)	1869.09644
Pearson Chi-Square	496.91515
Pearson Chi-Square/DF	1.59780

Negative binomial regression is often an alternative in this situation; it allows the variance to be larger than the mean, unlike the assumption of equivalence in Poisson regression. The HPGENSELECT procedure also provides negative binomial regression, and that is requested with the DIST=NB option:

```
proc hpgenselect data=baseball;
   class league division;
   model nHome = nAtBat nHits nRuns nRBI nBB
               yrMajor crAtBat crHits crHome crRuns crRbi
               crBB league division nOuts nAssts nError
             / dist=nb link=log;
   selection  method=forward details=all;
run;
```

Figure 7 lists the seven effects that entered the model for negative binomial regression.

Figure 7 Selection Summary

Selection Summary			
Step	**Effect Entered**	**Number Effects In**	**p Value**
0	**Intercept**	1	.
1	**nRBI**	2	<.0001
2	**nAssts**	3	<.0001
3	**nHits**	4	0.0019
4	**nRuns**	5	<.0001
5	**crHome**	6	0.0203
6	**crRbi**	7	<.0001
7	**yrMajor**	8	0.0464

Figure 8 displays the corresponding fit statistics for the final model; these values are somewhat smaller than the fit statistics reported for the final Poisson regression model. Note, however, that this is a different analysis and these measures cannot be directly compared. The Pearson chi-square/degrees of freedom ratio takes the value 1.01 for this analysis, which indicates no evidence of overdispersion. Thus, this model is deemed to be satisfactory.

Figure 8 Fit Statistics

Fit Statistics	
-2 Log Likelihood	1785.67309
AIC (smaller is better)	1803.67309
AICC (smaller is better)	1804.25001
BIC (smaller is better)	1837.64405
Pearson Chi-Square	318.66241
Pearson Chi-Square/DF	1.01485

Figure 9 contains the parameter estimates for this model.

Figure 9 Parameter Estimates

Parameter Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**
Intercept	1	1.110918	0.093513	141.1309	<.0001
nHits	1	-0.004750	0.001497	10.0700	0.0015
nRuns	1	0.007781	0.002272	11.7265	0.0006
nRBI	1	0.024132	0.001756	188.8612	<.0001
yrMajor	1	0.024209	0.012201	3.9369	0.0472
crHome	1	0.004504	0.000875	26.4625	<.0001
crRbi	1	-0.001383	0.000326	17.9730	<.0001
nAssts	1	-0.000636	0.000191	11.1047	0.0009
Dispersion	1	0.066758	0.014310	.	.

Bayesian Analysis

While the founding principle of Bayesian analysis goes back to the theorem developed by Sir Thomas Bayes in 1763, modern Bayesian analysis didn't catch fire until the computing advances of the 20th century. Bayesian analysis is based on the idea that inferences are based on measures of uncertainty through probability distributions, and prior knowledge can impact the inferences. In Bayesian analysis, parameters are random and you must estimate their posterior distributions. Summary statistics produced are posterior modes and credible intervals as opposed to the point estimates and confidence intervals in frequentist statistical analysis.

The Bayesian framework provides a straightforward framework for addressing scientific questions. For example, you can estimate the probability of an interval containing a value, versus the in-the-long run definition of a confidence interval which confuses so many clients. Thus, the framework is attractive for all types of situations, not just those in which you have prior information. While every Bayesian analysis incorporates a prior distribution, you can have noninformative prior distributions which do not influence the likelihood but do allow you to perform the Bayesian analysis. While there are analytic solutions for a few analyses, usually a closed form does not exist and simulation methods are required, such as Markov chain Monte Carlo methods.

SAS provides two avenues to Bayesian analysis: built-in Bayesian analysis in certain modeling procedures and the MCMC procedure for general-purpose modeling. Adding the BAYES statement generates Bayesian

analyses without the need to program priors and likelihoods for the GENMOD, PHREG, LIFEREG, and FMM procedures. Thus, you can obtain Bayesian results for:

- standard regression
- Poisson regression
- logistic regression
- loglinear models
- accelerated failure time models
- Cox proportional models
- piecewise exponential models
- frailty models
- finite mixture models

These procedures are ideal for users beginning to use Bayesian methods and will suffice for many analysis objectives.

The MCMC procedure is a general-purpose procedure for fitting Bayesian models. It uses a variety of sampling algorithms to draw from the posterior distribution. It produces the same convergence diagnostics and posterior summaries that you would find by using the BAYES statement in the modeling procedures. However, the MCMC procedure allows any likelihood, prior, or hyperprior that can be programmed with the SAS language. It supports multivariate distributions as well. If you are familiar with the NLMIXED procedure, you are familiar with the type of programming statements that the MCMC procedure requires.

The RANDOM statement in the MCMC procedure facilitates the specification of random effects in linear or nonlinear models. You can build nested or nonnested hierarchical models to artbitrary depth. Using the RANDOM statement can result in reduced simulation time and improved convergence for models that have a large number of subjects. The MCMC procedure also handles missing data for both responses and covariates. See papers by Fang Chen in the online conference proceedings for additional information.

Richer Linear Models

Hidden in the SAS/STAT 9.22 release (which had stealth qualities in its own right) is the new PLM procedure. This new procedure was the end result of a re-architecturing effort which put all of the post-fitting analysis statements–CONTRAST, LSMEANS, LSMESTIMATES,ESTIMATE–into a common framework. This meant that:

- 30 postfitting statements added to existing procedures
- faster new procedure development
- more efficient maintenance
- new features get to all relevant procedures faster

In addition, the PLM procedure performs postfitting inference with model fit information saved from a number of SAS/STAT modeling procedures. These procedures are equipped with the new STORE statement, which saves model information as a SAS *item store*. An item store is a special SAS binary file that is used to store and re-store information that has a hierarchical structure. Ten SAS/STAT procedures now provide the STORE statement: GENMOD, GLIMMIX, GLM, LOGISTIC, MIXED, ORTHOREG, PHREG, SURVEYLOGISTIC, SURVEYPHREG, and SURVEYREG.

The PLM procedure takes these item stores as input and performs tasks such as testing hypotheses, producing effect plots, and scoring a new data set. These tasks are specified through the usual complement of postfitting statements such as the TEST, LSMEANS, and new EFFECTPLOT and SCORE statements.

Any procedure that offers the STORE statement can produce the item stores that are necessary for postfitting processing with the PLM procedure. This allows you to perform additional postfitting inference at a later time without having to refit your model, which is especially convenient for those models that are computationally expensive. In addition, with growing concerns for data confidentiality, storing and using intermediate results for remaining analyses might become a requirement in some organizations.

The EFFECT statement provides a convenient way to add additional modeling terms to your linear model. These effects can include classification beyond simple grouping (multimember and lag effects), continuous modeling beyond simple polynomials (polynomial and spline effects), and general terms that you define yourself (collection effects). These constructed effects are viewed as any other model effect, which means you can use them in any postfitting analysis that is based on your model. The EFFECT statement also works well together with the new EFFECTPLOT statement, which provides a way to visualize the impact of the effects on the response variable. The EFFECT statement is available with a number of SAS/STAT linear modeling procedures.

Analyzing Time-To-Event Data

Data that measure time until an event, also known as lifetime, failure time, or survival data, occur frequently. They are often seen in clinical trials for medical treatments, such as survival time for heart transplant patients, but they also occur in many other settings; consider the lifetime of pedometers or the time until a new real estate agent makes his first sale. Time-to-event data require special attention. Not only are you measuring the failure time as a dependent variable, and possibly some covariates so that you can form a statistical model, but you often have to deal with censoring as well.

Censoring occurs when the event of interest hasn't occurred by the time data collection ends. It may happen because patients in a study withdraw, or drop out, before the study concludes. Or the experiment may simply be terminated before the event has occurred for some experimental units. In any event, you only know the lower bound of the failure time; and the observations are right-censored. You can have data that are left-censored, or only known to be smaller than a given value, and you can also have interval-censored data, or failure times that are only known to fall within a certain interval. You have to take censoring into account because, for example, in the clinical trial, the longer-lived individuals are more likely to be right-censored.

SAS/STAT has provided tools for the analysis of time-to-event data for years. They include the LIFETEST procedure for estimating the survivor function and comparing the survival curves of various groups. In addition, the LIFEREG procedure provides parametric regression methods for modeling the distribution of survival times with a set of covariates, and the PHREG procedure provides proportional hazards regression (Cox regression).

With recent releases, the survival analysis tools have been extended in a number of ways, including

- SURVEYPHREG procedure provides Cox regression for data collected from a complex survey
- Bayesian methods are available with the BAYES statement in the PHREG and LIFEREG procedures
- QUANTLIFE procedure provides quantile regression for right-censored data
- macro for interval censoring
- piecewise exponential regression available via PROC PHREG
- frailty analysis available via PROC PHREG

The SURVEYPHREG procedure is useful for data collected in large government surveys; for example, you might use Cox regression to model the time until the onset of depression in a national mental health care survey with a number of socioeconomic covariates. Bayesian methods are widely used in survival data analysis, and the BAYES statement had been added to the PHREG and LIFEREG procedures so that you can now perform Bayesian analysis for Cox regression and accelerated lifetime models. Note that the PIECEWISE= option has been added to the BAYES statement in the PHREG procedure so that you can now perform piecewise exponential modeling easily with SAS (both the Bayesian analysis and the maximum likelihood analysis are produced).

When experimental units are clustered, the failure times of those units within a cluster tend to be correlated. You need to account for the within-cluster correlation, and one way of doing that is the shared frailty model, in which the cluster effects are incorporated in the model as normally distributed random variables. Stokes, Chen, and So (2011) describe the new PHREG functionality to fit shared frailty models via the specification of a RANDOM statement in the SAS/STAT 9.3 release. The penalized partial likelihood approach is used, and that first implementation assumed that the frailties were distributed as lognormal. With SAS/STAT 12.1, the frailties can also be assumed to be distributed as gamma. PROC PHREG also provides Bayesian frailty analysis.

Competing risks develop when subjects are exposed to more than one cause of failure: for example, the cause of death in a bone marrow transplant could be relapse, death during remission, or death due to another cause. In that case, the cumulative incidence function is more appropriate than the standard Kaplan-Meier method of survival analysis. The SAS macro %CIF implements nonparametric methods for implementing this method and also provides Gray's method for testing differences between these functions in multiple groups. See Lin et al. (2012) for more information about this method and this macro.

Quantile regression provides an alternative and flexible technique for the analysis of survival data. You can apply this technique to right-censored responses, which allows you to explore the covariate effects on the quantiles of interest. Two approaches are implemented: one is based on the idea of the Kaplan-Meier estimator, and the other is based on the Nelson-Aalen estimator of the cumulative hazard function. The new QUANTLIFE procedure provides interior point algorithms for estimation, Wald tests for the parameter estimates, survival plots, conditional quantile plots, and quantile process plots. It also supports the EFFECT statement so that it can fit regression quantile spline curves, and it is multithreaded to take advantage of multiple processors when they are available.

Consider a study of primary biliary cirrhosis disease discussed in Lin, Wei, and Ying (1993). Prognostic factors studied included age, edema, bilirubin, albumin, and prothrombin. Researchers followed 418 patients between 1974 and 1984. The patients had a median follow-up time of 4.74 years and a censoring rate of 61.5%. The following SAS statements create the SAS data set PBC:

```
data pbc;
   input Time Status Age Albumin Bilirubin Edema Protime @@;
   label Time="Follow-up Time in Days";
   logAlbumin   =log(Albumin);
   logBilirubin =log(Bilirubin);
   logProtime   =log(Protime);
   datalines;
  400 1 58.7652 2.60 14.5 1.0 12.2 4500 0 56.4463 4.14  1.1 0.0 10.6
 1012 1 70.0726 3.48  1.4 0.5 12.0 1925 1 54.7406 2.54  1.8 0.5 10.3
 1504 0 38.1054 3.53  3.4 0.0 10.9 2503 1 66.2587 3.98  0.8 0.0 11.0
 1832 0 55.5346 4.09  1.0 0.0  9.7 2466 1 53.0568 4.00  0.3 0.0 11.0
 2400 1 42.5079 3.08  3.2 0.0 11.0   51 1 70.5599 2.74 12.6 1.0 11.5
 3762 1 53.7139 4.16  1.4 0.0 12.0  304 1 59.1376 3.52  3.6 0.0 13.6
 ...
 ...
```

The syntax for the MODEL statement for the QUANTLIFE procedure is similar to that used in other SAS survival procedures. The LOG option requests that the log response values be analyzed, the METHOD=NA option specifies the Nelson-Aalen method, and the PLOT=(QUANTPLOT SURVIVAL QUANTILE) option requests the estimated parameter by quantiles plot, the survival plot, and the predicted quantiles plot. The QUANTILE=(.1 .4 .5 .85) option requests that those quantiles be modeled.

```
ods graphics on;
proc quantlife data=pbc  LOG  method=na plot=(quantplot survival quantile) seed=1268;
   model Time*Status(0)=logBilirubin logProtime logAlbumin Age Edema
                       /quantile=(.1 .4 .5 .85);
run;
ods graphics off;
```

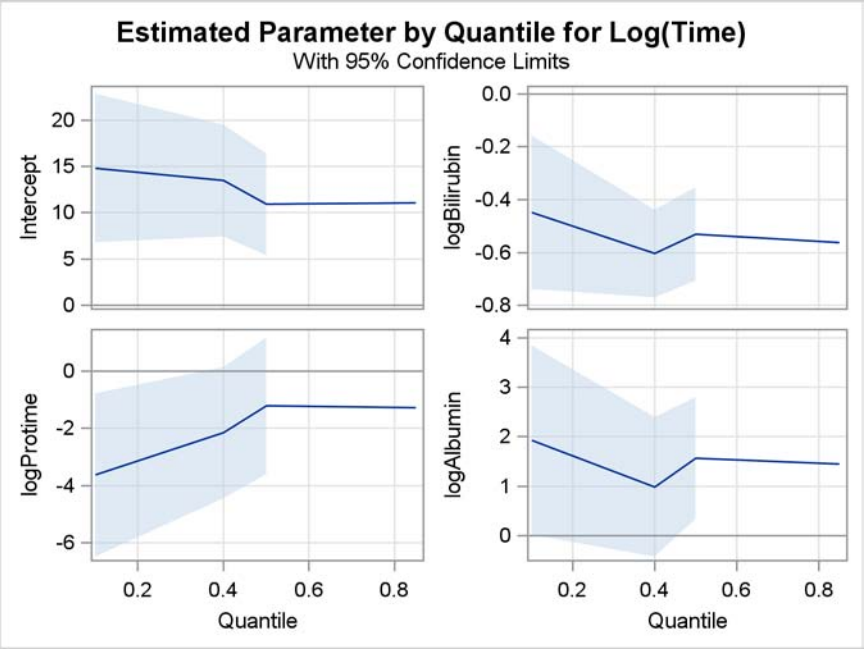
Figure 10 contains the parameter estimates. Each of the requested quantiles has its own set of parameter estimates. The confidence limits are computed by resampling methods.

Figure 10 Parameter Estimates

Parameter Estimates								
Quantile	Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
0.1000	**Intercept**	1	14.8030	4.0967	6.7736	22.8325	3.61	0.0003
	logBilirubin	1	-0.4488	0.1485	-0.7398	-0.1578	-3.02	0.0027
	logProtime	1	-3.6378	1.4560	-6.4915	-0.7841	-2.50	0.0129
	logAlbumin	1	1.9286	0.9756	0.0165	3.8408	1.98	0.0487
	Age	1	-0.0244	0.0107	-0.0455	-0.00334	-2.27	0.0237
	Edema	1	-1.0712	0.6688	-2.3820	0.2396	-1.60	0.1100
0.4000	**Intercept**	1	13.4716	3.0874	7.4204	19.5228	4.36	<.0001
	logBilirubin	1	-0.6047	0.0846	-0.7705	-0.4389	-7.15	<.0001
	logProtime	1	-2.1632	1.1726	-4.4615	0.1351	-1.84	0.0658
	logAlbumin	1	0.9819	0.7191	-0.4274	2.3912	1.37	0.1728
	Age	1	-0.0255	0.00681	-0.0389	-0.0122	-3.74	0.0002
	Edema	1	-1.0589	0.3104	-1.6672	-0.4506	-3.41	0.0007
0.5000	**Intercept**	1	10.9205	2.8047	5.4235	16.4175	3.89	0.0001
	logBilirubin	1	-0.5315	0.0904	-0.7087	-0.3543	-5.88	<.0001
	logProtime	1	-1.2222	1.2142	-3.6020	1.1577	-1.01	0.3148
	logAlbumin	1	1.5700	0.6284	0.3383	2.8016	2.50	0.0129
	Age	1	-0.0318	0.00883	-0.0491	-0.0145	-3.60	0.0004
	Edema	1	-0.7316	0.3743	-1.4653	0.00202	-1.95	0.0513
0.8500	**Intercept**	1	11.0778	.	.	.	.	.
	logBilirubin	1	-0.5615	.	.	.	.	.
	logProtime	1	-1.2711	.	.	.	.	.
	logAlbumin	1	1.4471	.	.	.	.	.
	Age	1	-0.0144	.	.	.	.	.
	Edema	1	-0.4339	.	.	.	.	.

This behavior of the covariate coefficients is illustrated in the plot in Figure 11. This is a scatter plot of the estimated regression parameter against the quantiles. In the plot for logPROTIME, the parameter estimate grows smaller from its value of –3.6456 for the 0.1 quantile and levels off around –1.0 for the 0.5 and higher quantiles.

Figure 11 Estimated Parameter by Quantiles Plot



See Lin and Rodriguez (2013) for more information about the QUANTLIFE procedure.

Missing Data Methods

While most of the data sets analyzed in classroom settings (or displayed in SAS documentation) are chock full of data, practicing statisticians rarely find this situation in the wild. Standard practice in software packages is to analyze complete case data by default, also known as casewise deletion, and SAS follows this practice. However, unless the missing observations magically reflect a random sample from the complete observations, the resulting inference is almost sure to be biased. Thus, managing missing data is an important aspect of statistical analysis.

One strategy for handling missing data is to impute the missing value, or substitute a value, and then analyze the data as if they were complete. Single imputation substitutes a single value, but the resulting results are biased toward zero since they don't reflect the uncertainty about the predictions of the unknown missing values (Rubin 1987). Multiple imputation does incorporate that uncertainty by replacing each missing value with a set of plausible values. You generate m complete data sets with m replacement values, and then you combine the results. This produces unbiased results.

The MI procedure in SAS/STAT produces multiply imputed data sets for incomplete multivariate data sets. The imputation method depends on the pattern of missingness and the type of the imputed variable. You then use standard SAS procedures to analyze the m imputed data sets, and you use the MIANALYZE procedure to combine the results and generate valid statistical inference.

The pattern of missing data determines the imputation method. There are many choices when you have monotone missing data: for missing continuous variables, you can use a regression method, predictive mean matching, or a propensity scoring method. For categorical missing variables, you can apply a logistic regression method or a discriminant function method. When you have arbitrary missing data patterns, you can use an MCMC method that assumes multivariate normality or a fully functional specification method (FCS) that assumes the existence of a joint distribution for all variables. The FCS method is a recent addition to the MI procedure, and it offers additional flexibility.

SAS/STAT offers other techniques for managing missing data. The CALIS procedure for fitting structural equations model now provides full information maximum likelihood (FIML), which is an estimation method that uses information from both the incomplete and the complete observations. Besides FIML estimation, PROC CALIS also provides features for analyzing data coverage and missing patterns. Since structural equations modeling includes measurement error models, regression models, and factor analyses as subset

analyses, the FIML method it includes has a wide range of applications. See Yung and Zhang (2011) for more information.

One of the benefits of the Bayesian approach is that it can handle missing data in a straightforward manner. It treats the missing values as unknown parameters and estimates their posterior distributions. It is a model-based solution, and the additional parameters don't add that much additional complexity to the analysis; they are simply sampled sequentially in the MCMC simulation. This approach does take into account the uncertainty about the missing values so you can estimate the posterior marginal distributions of the parameters of interest conditional on the observed and partially observed data. The MCMC procedure automatically samples all missing values and incorporates them in the Markov chain for the parameters. You can use PROC MCMC to handle various types of missing data, including data that are missing at random (MAR) and missing not at random (MNAR). PROC MCMC can also perform joint modeling of missing responses and covariates. See Chen (2013) for more information.

Managing missing data continues to be an important area of research and application, and providing additional techniques is a major focus of current SAS/STAT research and development.

High-Performance Computing

SAS/STAT software has taken advantage of multithreading algorithms to improve performance in several procedures, including the REG, GLM, LOESS, and GLMSELECT procedures, for years. See Cohen (2002) for information on how this works and how the data configuration affects the resulting computing performance. Recent new procedures in SAS/STAT come equipped with multithreading if it would benefit their performance. These procedures include the FMM, QUANTSELECT, QUANTLIFE, and ADAPTIVEREG procedures.

Most recently, SAS has focused on meeting the challenges of Big Data with more attention to high-performance computing. New software products designed specifically for distributed computing include the high-performance analytical products, which operate on data stored in databases such as Teradata, Greenplum, and Hadoop and uses multiple parallel processing techniques across a grid of servers. New statistical procedures were designed for this software, including those that perform logistic regression, linear regression, mixed models, and model selection for both linear models and generalized linear models.

Beginning with Base SAS 9.4/SAS/STAT 12.3, released in the summer of 2013, the SAS/STAT product includes the high-performance procedures for use in single-machine mode. These procedures were designed to provide specific functionality required for Big Data analysis in a distributed environment, such as predictive modeling. Not all features of the traditional procedure are included or are relevant. The high-performance procedures are evolving and additional functionality such as graphics and BY-group processing will surface in later releases. However, these procedures may provide benefit for the typical SAS/STAT user:

- New functionality is included in some of these procedures. For example, model selection for generalized linear models is available with the HPGENSELECT procedure.
- Depending on the characteristics of the data and the complexity of the analysis, users may find performance gains in single-machine mode with these procedures. However, note that if you compare a high-performance procedure with a traditional procedure that is multithreaded (for example, PROC HPREG compared to PROC REG), you are unlikely to see performance differences in the single-machine mode of execution.
- Users with Big Data who would benefit from using the high-performance analytics products in a distributed environment can exercise the procedures in single-machine mode and assess their functionality. When the customer needs to process Big Data (more observations and larger numbers of variables), she can license the high-performance analytics products and execute the same procedures in a distributed, in-memory environment.

The high-performance procedures are in the first stages of development. While production software, they will be enhanced as time goes on and will include some additional features, such as ODS graphics and BY-group processing. These procedures are documented in *SAS/STAT 12.3 User's Guide: High-Performance Procedures*. See Cohen (2013) for more detail.

Complex Data

Data today are increasingly complex, or perhaps that complexity is being acknowledged because there are more sophisticated methods available to analyze them. A major growth area for SAS/STAT has always been statistical techniques that address data complexity. Recently, additional methods for handling complex data include quantile regression, finite mixture models, and adaptive regression splines.

As previously discussed, quantile regression extends the basic regression model to the relationship between the conditional quantiles of a response variable with one or more covariates. It makes no distributional assumptions about the error term, and so it offers model robustness. It is a semiparametric method that can provide a more complete picture of your data based on these conditional distributions. Linear programming algorithms are used to produce the quantile regression estimates. See Koenker (2005) for further detail. SAS/STAT provides the QUANTREG procedures for quantile regression analysis, the QUANTSELECT procedure for model selection for quantile regression, and the QUANTLIFE procedure for right-censored data.

New ADAPTIVEREG Procedure

SAS/STAT software provides various tools for nonparametric regression, including the LOESS, TPSPLINE, and GAM procedures. Typical nonparametric regression methods involve a large number of parameters to capture nonlinear trends, so the model space is fairly large. The sparsity of data in high dimensions is another issue, often resulting in slow convergence or even failure for many nonparametric regression methods.

The LOESS and TPSPLINE procedures are limited to problems in low dimensions. The GAM procedure fits generalized additive models with the assumption of additivity. It can handle data sets, but the computation time for its local scoring algorithm (Hastie and Tibshirani 1990) to converge increases quickly with the size of the data set.

The new ADAPTIVEREG procedure provides a nonparametric modeling approach for high-dimensional data. PROC ADAPTIVEREG fits multivariate adaptive regression splines as introduced by Friedman (1991b). The method is a nonparametric regression technique that combines both regression splines and model selection methods. It does not assume parametric model forms, and it does not require knot values for constructing regression spline terms. Instead, it constructs spline basis functions in an adaptive way by automatically selecting appropriate knot values for different variables; it performs model reduction by applying model selection techniques. Thus, the ADAPTIVEREG procedure is both a nonparametric regression procedure and a predictive modeling procedure.

The ADAPTIVEREG procedure:

- supports classification variables with different ordering options
- enables you to force effects into the final model or restrict variables in linear forms
- supports options for fast forward selection
- supports partitioning of data into training, validation, and testing roles
- provides leave-one-out and k-fold cross validation
- produces graphical representations of the selection process, model fit, functional components and fit diagnostics

For more detail, see Kuhfeld and Cai (2013).

The following example illustrates the use of the ADAPTIVEREG procedure. Researchers collected data on city-cycle fuel efficiency and automobile characteristics for 361 vehicle models manufactured from 1970 to 1982. The data can be downloaded from the UCI Machine Learning Repository (Asuncion and Newman 2007). The following DATA step creates the data set AUTOMPG:

```
title 'Automobile MPG study';
data autompg;
   input mpg cylinders displacement horsepower weight
         acceleration year origin name $35.;
   datalines;
18.0   8   307.0   130.0   3504   12.0   70   1   chevrolet chevelle malibu
15.0   8   350.0   165.0   3693   11.5   70   1   buick skylark 320
18.0   8   318.0   150.0   3436   11.0   70   1   plymouth satellite
16.0   8   304.0   150.0   3433   12.0   70   1   amc rebel sst
17.0   8   302.0   140.0   3449   10.5   70   1   ford torino
...
...
;
```

There are nine variables in the data set. The response variable MPG is city-cycle mileage per gallon (mpg). Seven predictor variables (number of cylinders, displacement, weight, acceleration, horsepower, year and origin) are created. The variables for number of cylinders, year, and origin are categorical.

The dependency of vehicle fuel efficiency on these factors might be nonlinear. Dependency structures within the predictors might also mean that some of the predictors are redundant. For example, a model with more cylinders is likely to have more horsepower. The object of this analysis is to explore the nonlinear dependency structure and to find a parsimonious model that does not overfit the data. A more parsimonious model has better predictive ability.

The following PROC ADAPTIVEREG statements fit an additive model with linear spline terms of continuous predictors. The variable transformations and the model selection based on the transformed terms are performed in an adaptive and automatic way. If the ADDITIVE option is not supplied, PROC ADAPTIVEREG will fit a model with both main effects and two-way interaction terms.

```
ods graphics on;
proc adaptivereg data=autompg plots=all;
   class cylinders year origin;
   model mpg = cylinders displacement horsepower
               weight acceleration year origin / additive;
run;
ods graphics off;
```

Figure 12 displays information on how the bases are constructed.

Figure 12 Bases

Automobile MPG Study

Basis Information	
Name	**Transformation**
Basis0	None
Basis1	Basis0*MAX(Weight - 3139,0)
Basis2	Basis0*MAX(3139 - Weight,0)
Basis3	Basis0*NOT(MISSING(HorsePower))
Basis4	Basis0*MISSING(HorsePower)
Basis5	Basis3*MAX(HorsePower - 158,0)
Basis6	Basis3*MAX(158 - HorsePower,0)
Basis7	Basis3*(Year = 80 OR Year = 82 OR Year = 81 OR Year = 79 OR Year = 78 OR Year = 77 OR Year = 73)
Basis8	Basis3*NOT(Year = 80 OR Year = 82 OR Year = 81 OR Year = 79 OR Year = 78 OR Year = 77 OR Year = 73)
Basis9	Basis0*MAX(Acceleration - 21,0)
Basis10	Basis0*MAX(21 - Acceleration,0)
Basis11	Basis0*(Cylinders = 3 OR Cylinders = 6)
Basis12	Basis0*NOT(Cylinders = 3 OR Cylinders = 6)
Basis13	Basis4*(Origin = 1)
Basis14	Basis4*NOT(Origin = 1)
Basis15	Basis0*(Origin = 3)
Basis16	Basis0*NOT(Origin = 3)
Basis17	Basis0*(Cylinders = 6)
Basis18	Basis0*NOT(Cylinders = 6)
Basis19	Basis0*(Year = 73 OR Year = 80 OR Year = 82 OR Year = 81 OR Year = 79)
Basis20	Basis0*NOT(Year = 73 OR Year = 80 OR Year = 82 OR Year = 81 OR Year = 79)

The “Parameter Estimates” table in Figure 13 displays parameter estimates for constructed basis functions in addition to each function’s construction component. For example, BASIS1 has an estimate of –0.003242. It is constructed from a parent basis function BASIS0 (intercept) and a linear spline function of WEIGHT with a single knot placed at 3139. BASIS3 is constructed from a parent basis function BASIS0 and an indicator function of YEAR. The indicator is set to 1 when a class level of YEAR falls into the subset of levels listed in the “Levels” column and set to 0 otherwise.

Figure 13 Parameter Estimates

Regression Spline Model after Backward Selection					
Name	**Coefficient**	**Parent**	**Variable**	**Knot**	**Levels**
Basis0	29.4394		Intercept		
Basis2	0.004412	Basis0	Weight	3139.00	
Basis3	-21.2899	Basis0	HorsePower	.	
Basis6	0.1534	Basis3	HorsePower	158.00	
Basis7	2.3920	Basis3	Year		10 12 11 9 8 7 3
Basis9	1.6658	Basis0	Acceleration	21.0000	
Basis10	0.4672	Basis0	Acceleration	21.0000	
Basis11	-8.1766	Basis0	Cylinders		0 3
Basis13	-10.0976	Basis4	Origin		0
Basis15	2.1354	Basis0	Origin		2
Basis17	6.7675	Basis0	Cylinders		3
Basis19	1.4987	Basis0	Year		3 10 12 11 9

During the model construction and selection process, some basis function terms are removed.

Variable importance is another criterion that focuses on the contribution of each individual. Variable importance is defined to be the square root of the GCV value of a submodel with all basis functions that involve a removed variable, minus the square root of the GCV value of the selected model, then scaled to have the largest importance value of 100. Figure 14 lists importance values for four variables that comprise the selected model. Similar to the ANOVA decomposition results, WEIGHT and YEAR are two dominant factors that determine vehicle mpg values, while DISPLACEMENT and ACCELERATION are less important.

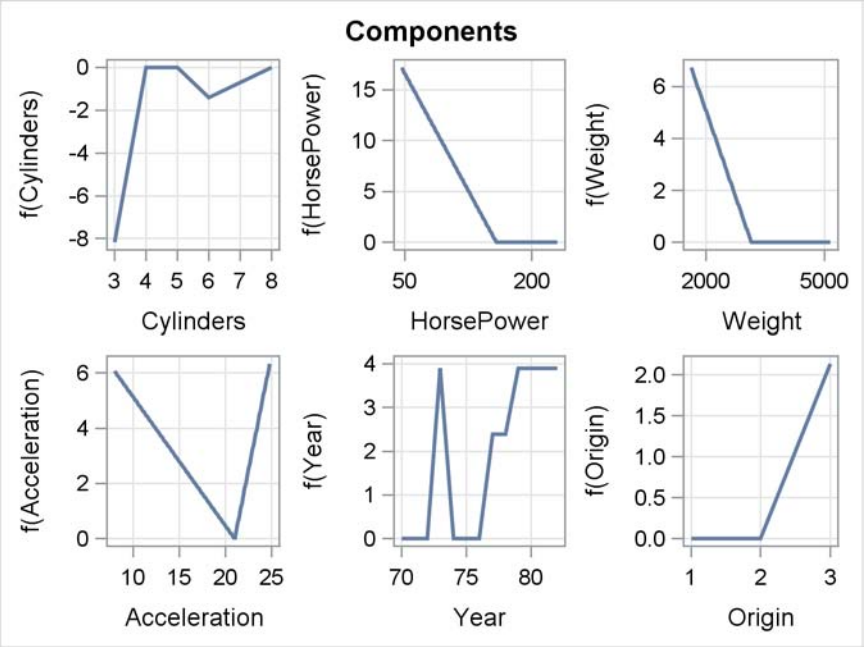
Figure 14 Variable Importance

Variable Importance		
Variable	**Number of Bases**	**Importance**
HorsePower	1	100.00
Year	2	85.46
Weight	1	21.10
Cylinders	2	19.08
Origin	2	18.67
Acceleration	2	16.38

The component panel in Figure 15 displays the fitted functional components against their forming variables. When a vehicle model's displacement is less than 85, its mpg value increases with its displacement. The displacement does not matter much once it exceeds 85. The shape of the functional component strongly suggests a logarithm transformation. The component of WEIGHT shows that vehicle weight has negative impact on its mpg value. The trend suggests a possible reciprocal transformation. When a model's acceleration value is larger than 20.7, it affects the mpg value in a positive manner. It does not matter much if it is less than 20.7. Although YEAR is treated as a classification variable, its values are ordinal. The general

trend is quite clear: more recent models tend to have higher mpg values. Automobile companies apparently paid more attention to improving vehicle fuel efficiency after 1976.

Figure 15 Component Panel



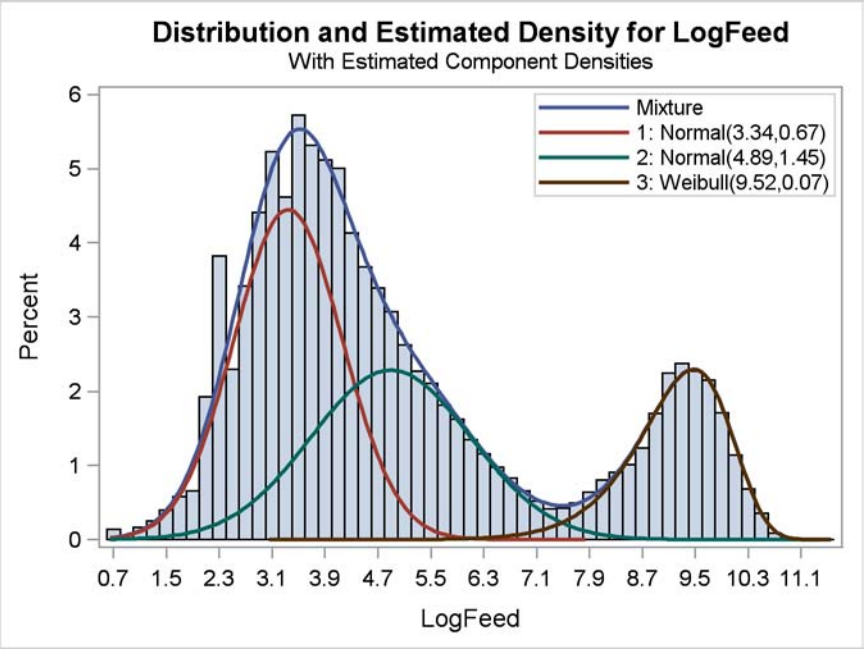
Finite Mixture Models

Another type of complexity occurs when data can be viewed as coming from a mixture of different distributions. Finite mixture models enable you to fit statistical models to data when the distribution of the response is a finite mixture of univariate distributions. These models are useful for applications such as estimating multimodal or heavy-tailed densities, fitting zero-inflated or hurdle models to count data with excess zeros, modeling overdispersed data, and fitting regression models with complex error distributions. Many well-known statistical models for dealing with overdispersed data are members of the finite mixture model family (for example, zero-inflated Poisson models and zero-inflated negative binomial models.)

PROC FMM performs maximum likelihood estimation for all models, and it provides Markov chain Monte Carlo estimation for many models, including zero-inflated Poisson models. The procedure includes many built-in link and distribution functions, including the beta, shifted, Weibull, beta-binomial, and generalized Poisson distributions, as well as standard members of the exponential family of distributions. In addition, several specialized built-in mixture models are provided, such as the binomial cluster model (Morel and Nagaraj, 1993).

The results of a finite mixture models analysis is displayed in Figure 16. The FMM procedure was used to fit a three-component model–two normal components and a Weibull component– to log feeding time for cattle. The plot shows the observed and estimated distributions for the response.

Figure 16 Density Plot



Summary

Of course, recent SAS/STAT releases include numerous other updates. The STDRATE procedure computes direct and indirect standardized rates and proportions, measures key in epidemiology. The survey data analysis procedures continue to be a major development focus, with post-stratification estimation now available with the SURVEYMEANS procedure and Poisson sampling included in PROC SURVEYSELECT. Other new features include:

- WEIGHT statement in PROC LIFETEST
- partial R-square for relative importance of parameters in PROC LOGISTIC
- Miettinen-Nurminen confidence limits for the difference of proportions in PROC FREQ
- group sequential design with nonbinding acceptance boundary in the SEQDESIGN and SEQTEST procedures
- REF= option added to the CLASS statement for GLM, MIXED, GLIMMIX, and ORTHOREG procedures

SAS/STAT software provides its customers with a rich array of current statistical techniques. Released every 12-18 months, it provides modern statistical methodology via software that is geared towards user expectations and shaped for today's data.

For Further Information

A good place to start for further information is the "What's New in SAS/STAT 12.1" chapter in the online documentation. In addition, the Statistics and Operations Focus Area includes substantial information about the statistical products, and you can find it at `support.sas.com/statistics/`. The quarterly e-newsletter for that site is available on its home page. And of course, complete information is available in the online documentation located here: `support.sas.com/documentation/onlinedoc/stat/`.

References

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

Bellman, R. E. (1961), *Adaptive Control Processes,* Princeton University Press.

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

Buja, A., Duffy, D., Hastie, T., and Tibshirani, R. (1991), "Discussion: Multivariate Adaptive Regression Splines," *The Annals of Statistics,* 19, 93–99.

Chen, F. (2009), "Bayesian Modeling Using the MCMC Procedure," in *Proceedings of the SAS Global Forum 2008 Conference,* Cary NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings09/257-2009.pdf.

Chen, F. (2011), "The RANDOM Statement and More: Moving on with PROC MCMC," in *Proceedings of the SAS Global Forum 2011 Conference,* Cary NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings11/334-2011.pdf.

Collier Books (1987), "The 1978 Baseball Encyclopedia Update," New York: Macmillan.

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

Cohen, R. and Rodriguez, R. (2013) "High Performance Statistical Modeling," Available at http://support.sas.com/statistics/papers/

Derr, R. (2013) "Ordinal Response Modeling with the LOGISTIC Procedure," *Proceedings of the SAS Global Forum 2013 Conference,* Cary, NC: SAS Institute Inc. Available at http:/support.sas.com/resources/papers/proceedings13/446-2013.pdf.

Friedman, J. (1991a), "Estimating Functions of Mixed Ordinal and Categorical Variables Using Adaptive Splines," Technical report, Stanford University.

Friedman, J. (1991b), "Multivariate Adaptive Regression Splines," *The Annals of Statistics,* 19, 1–141.

Friedman, J. (1993), "Fast MARS," Technical report, Stanford University.

Florida Department of Health, "Florida Vital Statistics Annual Report 2000." Available at http://www.flpublichealth.com/VSBOOK/pdf/2000/Population.pdf. Accessed February 2012.

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

Gibbs, P., Tobias, R., Kiernan, K., and Tao, J. 2013) "Having an EFFECT: More General Linear Modeling and Analysis with the New EFFECT Statement in SAS/STAT® Software," *Proceedings of the SAS Global Forum 2013 Conference,* Cary, NC: SAS Institute Inc. Available at http:/support.sas.com/resources/papers/proceedings13/437-2013.pdf.

Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models,* New York: Chapman & Hall.

Koenker, R. (2005), *Quantile Regression,* New York: Cambridge University Press.

Kuhfeld, W., and Cai, W. (2013) "Introducing the New ADAPTIVEREG Procedure for Adaptive Regression," *Proceedings of the SAS Global Forum 2013 Conference,* Cary, NC: SAS Institute Inc. Available at http:/support.sas.com/resources/papers/proceedings13/457-2013.pdf.

Lin, G., So, Y., and Johnston, G. (2012) "Analyzing Survival Data with Competing Risks Using SAS Software," *Proceedings of the SAS Global Forum 2012 Conference,* Cary, NC: SAS Institute Inc. Available at http:/support.sas.com/resources/papers/proceedings11/344-2012.pdf.

Lin, G. (2013) "Using the QUANTLIFE Procedure for Survival Analysis," *Proceedings of the SAS Global Forum 2013 Conference,* Cary, NC: SAS Institute Inc. Available at http:/support.sas.com/resources/

papers/proceedings11/421-2013.pdf.

Lin, D. Y., Wei, L. J., and Ying, Z. (1993), "Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals," *Biometrika*, 80, 557–572.

Little, R.J.A. and Rubin, D.B. (2002), *Statistical Analysis with Missing Data,* Second Edition, New York: John Wiley & Sons.

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

Peng L. and Huang Y. (2008), "Survival Analysis with Quantile Regression Models," *Journal of the American Statistical Association*, 103, 637–649

Portnoy S. (2003). "Censored Quantile Regression,"" *Journal of American Statistical Association*, 98, 1001–1012.

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

Stokes, M., Rodriguez, R. and Cohen, R. (2010), "SAS/STAT 9.22: The Next Generation," in *Proceedings of the SAS Global Forum 2011 Conference,* Cary NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings10/264-2010.pdf.

Stokes, M., Chen, F., and So, Y. (2011), "On Deck: SAS/STAT 9.3," *Proceedings of the SAS Global Forum 2011 Conference,* Cary, NC: SAS Institute Inc. Available at http:/support.sas.com/resources/papers/proceedings11/331-2011.pdf.

Stokes, M., (2012), "Look Out: After SAS/STAT® 9.3 Comes SAS/STAT 12.1!," *Proceedings of the SAS Global Forum 2012 Conference,* Cary, NC: SAS Institute Inc. Available at http:/support.sas.com/resources/papers/proceedings12/313-2012.pdf.

Yang, Y. (2013) "Computing Direct and Indirect Standardized Rates and Risks with the STDRATE Procedure," *Proceedings of the SAS Global Forum 2013 Conference,* Cary, NC: SAS Institute Inc. Available at http:/support.sas.com/resources/papers/proceedings11/423-2013.pdf.

U.S. Bureau of Census (2011), "Age and Sex Composition: 2010." Available at http://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf. Accessed February 2012.

Acknowledgments

The authors are grateful to Fang Chen for his contributions to the manuscript.

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

Version

1.0