Paper 313-2012

Look Out: After SAS/STAT® 9.3 Comes SAS/STAT 12.1!

Maura Stokes, Fang Chen, Yang Yuan, and Weijie Cai
SAS Institute, Inc. Cary NC

Abstract

Heralded by a new release-numbering scheme, SAS/STAT 12.1 comes loaded with new statistical capabilities. New development areas include model selection for quantile regression, quantile regression for censored data, and multivariate adaptive regression splines. Epidemiologists will like the STDRATE procedure for computing direct and indirect standardized rates and risks for study populations. The FMM procedure becomes production and includes new features such as additional distributions. Other notable enhancements include modeling missing covariates with the MCMC procedure and fitting Bayesian frailty models with PROC PHREG. This paper reviews highlights from earlier releases and describes highlights of SAS/STAT 12.1, slated for release during 2012.

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 the 'mother ship' along with other SAS analytical products. This means that these products can be released to customers 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 is changing with the next release; they will be numbered '12.1.' This numbering scheme will be maintained when new versions of Base SAS and SAS/STAT ship at the same time. For example, when Base SAS 9.4 is released, SAS/STAT 13.1 will be released.

To keep informed about SAS/STAT releases, see support.sas.com/stat/ for product news and see support.sas.com/statistics/ for in-depth information and a link to the e-newsletter.

Overview of Recent and Future Updates

SAS/STAT 9.22 made available a full complement of postfitting capabilities in many linear modeling procedures. This release also introduced the PLM procedure, which enables you to take stored model information and use it to perform additional inference and scoring without refitting the original model. The SURVEYPHREG procedure provides survival analysis, in the form of Cox proportional hazards regression, for sample survey data. More powerful and customizable structural equation modeling, first implemented with the experimental TCALIS procedure in SAS/STAT 9.2, was rolled into the CALIS procedure. Other enhancements included exact Poisson regression, zero-inflated negative binomial models, model-averaging, and improvements to the spatial analysis procedures. See Stokes, Rodriguez, and Cohen (2010) for more information.

SAS/STAT 9.3 became available in 2011, and it introduced the experimental FMM procedure, which fits statistical models to data where the distribution of the response is a finite mixture of univariate distributions. The MI procedure added the FCS statement, which specifies a multivariate imputation by fully conditional specification (FCS) methods. The NLIN procedure was updated with features for diagnosing the nonlinear model fit. The SURVEYPHREG procedure became production and now handles time-dependent covariates. The MCMC procedure added a RANDOM statement, which simplifies the specification of hierarchical random-effects models and significantly reduces simulation time while improving convergence. See Stokes, Chen, and So (2011) for more information.

The upcoming 12.1 release of SAS/STAT emphasizes modern regression methods. The new QUANTSELECT procedure for quantile regression model selection works similarly to the GLMSELECT procedure, and the new QUANTLIFE procedure performs quantile regression for censored data. The new ADAPTIVEREG procedure provides flexible regression model for high-dimensional data. In addition, epidemiologists will benefit from the new STDRATE procedure, which computes direct and indirect standardized rates and risks for study populations. The FMM procedure for finite mixture models becomes production, and Bayesian analysis capabilities are also updated.

This paper reviews the highlights of the new release and illustrates them with practical examples. It draws heavily from the documentation. See sas.com/statistics/papers/ for any update of this paper at release time.

New STDRATE Procedure

Epidemiologists constantly deal with confounders that can bias a measure of the association between an exposure and an event outcome. If confounding is not taken into account, the overall event rate estimated might not be meaningful so you employ stratification to control potential confounding. You first subdivide a population into constituent subpopulations according to certain criteria for confounding variables, such as age and gender. Then, you estimate the effect of the exposure within each stratum and you combine the stratum-specific effect estimates into an overall estimate that is presumably free of bias.

The STDRATE procedure computes direct and indirect standardized rates and risks for study populations. Direct standardization computes the weighted average of stratum-specific estimates in the study population, using weights such as population-time from a standard or reference population. For two study populations with the same reference population, the procedure compares directly standardized rates or risks. In addition, the procedure also computes Mantel-Haenszel effect estimates, such as the rate difference, from two study populations without a reference population.

Indirect standardization computes the weighted average of stratum-specific estimates in the reference population, using weights from the study population. The ratio of the overall rate or risk in the study population and the corresponding weighted estimate in the reference population, which is also the ratio of the observed number of events and the expected number of events in the study population, is the standardized morbidity or mortality ratio (SMR). The SMR compares rates or risks in the study and reference populations. The indirect standardized rate estimate is the product of the SMR and the crude rate estimate for the reference population.

The following example illustrates the use of the STDRATE procedure to compute standardized mortality ratios to compare the death rates of skin cancer between Florida and the United States as a whole. Indirect standardization is used.

The FLORIDA_43 data set contains stratum-specific mortality information for skin cancer during 2000 from the Department of Health in Florida. The variable AGE is the grouping variable that determines the strata for the standardization; variables EVENT and PYEAR represent the number of events and total person-years, respectively. The COMMA11. format is used to input numbers that contain commas.

```
data Florida_C43;
input Age $1-5 Event PYear comma11.;
datalines;
00-04   0     953,785
05-14   0   1,997,935
15-24   4   1,885,014
25-34  14   1,957,573
35-44  43   2,356,649
45-54  72   2,088,000
55-64  70   1,548,371
65-74 126   1,447,432
75-84 136   1,087,524
85+    73     335,944
;
```

The US_C43 data set contains comparable mortality information for the United States for the year 2000 (from the Centers for Disease Control and Prevention, 2002; U.S. Bureau of Census 2011). The same variables are created as in the previous DATA step.

```
data US_C43;
input Age $ 1-5 Event comma7. PYear comma12.;
datalines;
00-04     0  19,175,798
05-14     1  41,077,577
15-24    41  39,183,891
25-34   186  39,892,024
35-44   626  45,148,527
45-54 1,199  37,677,952
55-64 1,303  24,274,684
65-74 1,637  18,390,986
75-84 1,624  12,361,180
85+     803   4,239,587
;
```

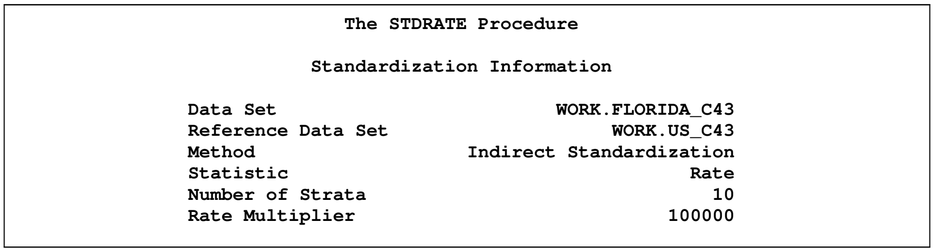
The following statements invoke the STDRATE procedure and request indirect standardization to compare the mortality rates between Florida and the United States. The DATA= option specifies the study data set, and the REFDATA= option specifies the reference data set. You request indirect standardization with the METHOD=INDIRECT option. Specifying STAT=RATE requests the rate as the frequency measure for standardization, and specifying MULT=100000 (default) displays the deaths per 100,000 person-years in the results. The PLOTS=ALL option requests a plot of the resulting standardized mortality rates.

```
ods graphics on;
proc stdrate data=Florida_C43 refdata=US_C43
             method=indirect
             stat=rate(mult=100000)
             plots=all
             ;
   population event=Event total=PYear;
   reference  event=Event total=PYear;
   strata Age / info(cl=none) smr;
run;
ods graphics off;
```

The EVENT= and TOTAL= options in the POPULATION statement specify variables for the number of events and person-years in the study population, and the same options specify these variables in the REFERENCE statement. You list the stratification variable AGE in the STRATA statement. The INFO option requests stratum-specific statistics such as rates, and the SMR option requests stratum-specific SMR estimates.

Figure 1 contains the standardization information.

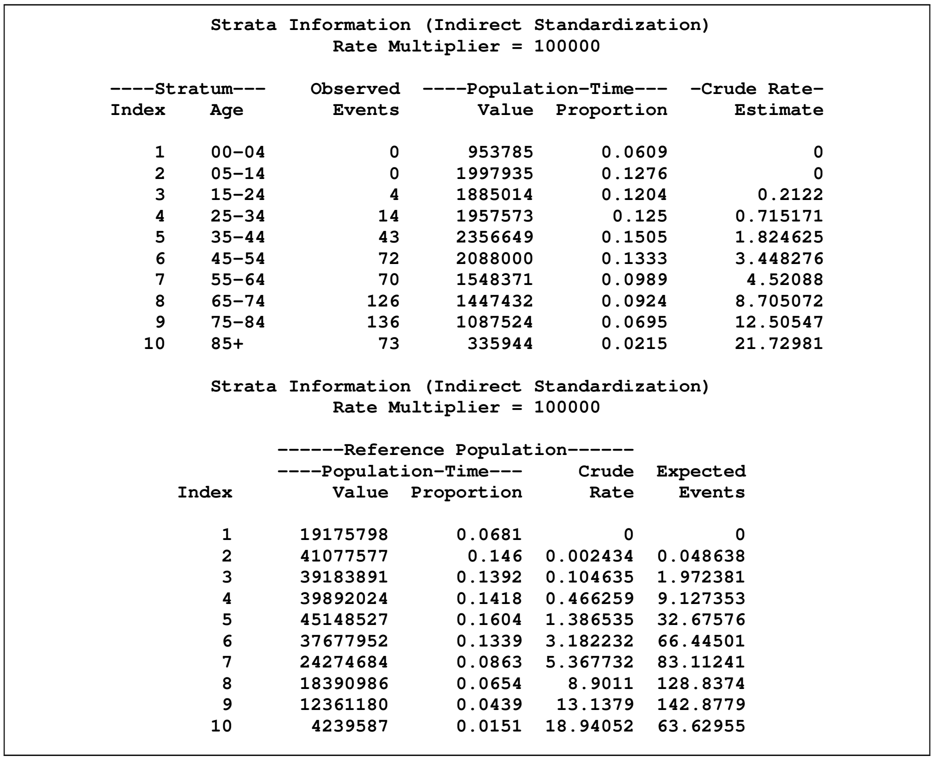
Figure 1 Standardization Information



```
                        The STDRATE Procedure

                      Standardization Information

              Data Set                        WORK.FLORIDA_C43
              Reference Data Set                    WORK.US_C43
              Method                    Indirect Standardization
              Statistic                                     Rate
              Number of Strata                                10
              Rate Multiplier                             100000
```

Figure 2 contains the strata information and the expected number of events at each stratum. Crude rates per 100,000 person-years are displayed. The “Expected Events” column displays the expected number of events when the stratum-specific rates in the reference data set are applied to the corresponding person-years in the study data set.

Figure 2 Strata Information



```
                Strata Information (Indirect Standardization)
                           Rate Multiplier = 100000

        ----Stratum---    Observed  ----Population-Time---  -Crude Rate-
        Index     Age       Events       Value  Proportion      Estimate

            1     00-04          0      953785      0.0609             0
            2     05-14          0     1997935      0.1276             0
            3     15-24          4     1885014      0.1204        0.2122
            4     25-34         14     1957573       0.125      0.715171
            5     35-44         43     2356649      0.1505      1.824625
            6     45-54         72     2088000      0.1333      3.448276
            7     55-64         70     1548371      0.0989       4.52088
            8     65-74        126     1447432      0.0924      8.705072
            9     75-84        136     1087524      0.0695      12.50547
           10     85+           73      335944      0.0215      21.72981

                Strata Information (Indirect Standardization)
                           Rate Multiplier = 100000

                      ------Reference Population------
                      ----Population-Time---     Crude   Expected
              Index         Value  Proportion      Rate     Events

                  1      19175798      0.0681         0          0
                  2      41077577       0.146  0.002434   0.048638
                  3      39183891      0.1392  0.104635   1.972381
                  4      39892024      0.1418  0.466259   9.127353
                  5      45148527      0.1604  1.386535   32.67576
                  6      37677952      0.1339  3.182232   66.44501
                  7      24274684      0.0863  5.367732   83.11241
                  8      18390986      0.0654    8.9011   128.8374
                  9      12361180      0.0439   13.1379   142.8779
                 10       4239587      0.0151  18.94052   63.62955
```

Figure 3 and Figure 4 display the strata distribution plot and the strata rate plot.

Figure 3 Strata Distribution Plot

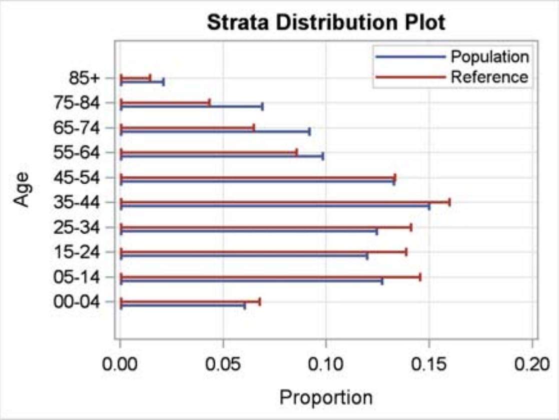
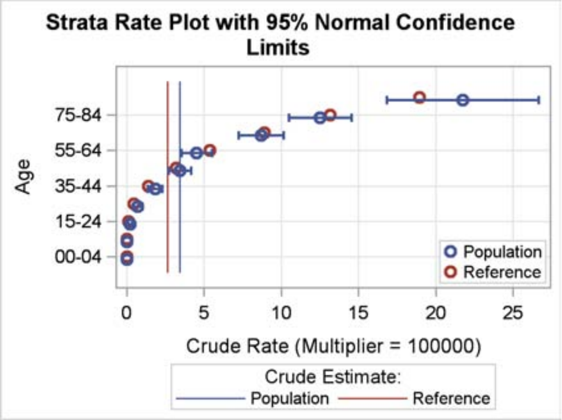


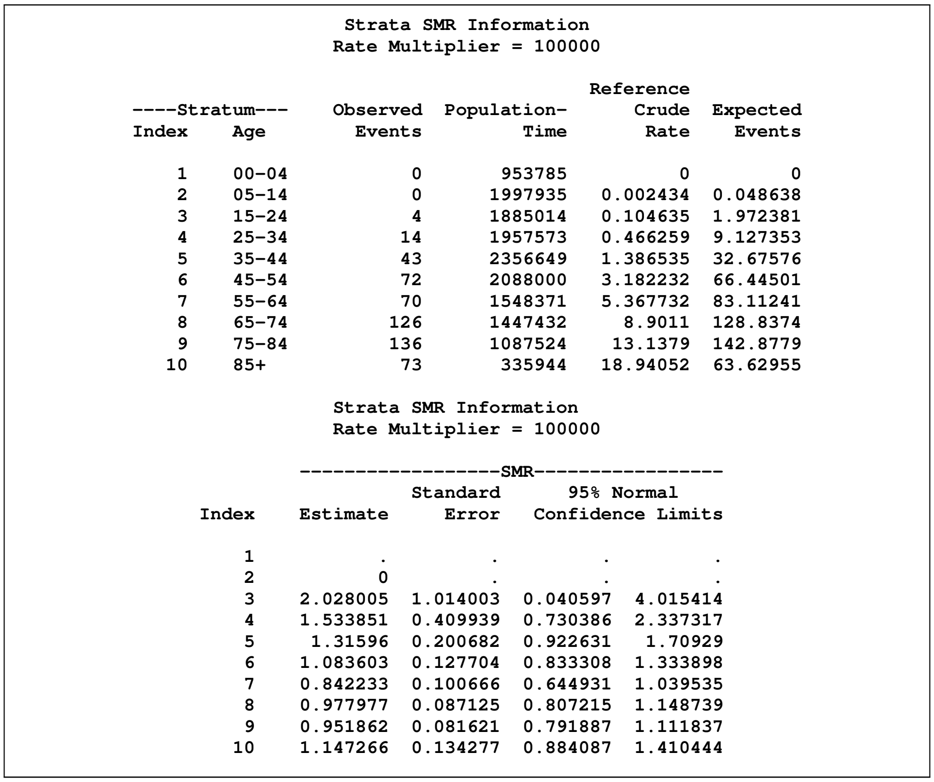
Figure 4 Strata Rate Plot



The distribution plot displays the strata proportions listed in Figure 2. It shows that the study population has higher proportions in older age groups and lower proportions in younger age groups than the reference population The strata rate plot displays stratum-specific rate estimates in the study and reference populations. It also displays the confidence limits for the rates in the study population and the overall crude rates for the two populations (the two vertical lines).

Figure 5 displays the SMR for each stratum. Since the MULT=100000 suboption was specified, the events per 100,000 person-years are displayed.

Figure 5 Strata SMR Information



```
                     Strata SMR Information
                    Rate Multiplier = 100000

                                                  Reference
----Stratum---     Observed  Population-          Crude    Expected
Index     Age        Events         Time           Rate      Events

    1     00-04           0       953785              0           0
    2     05-14           0      1997935       0.002434    0.048638
    3     15-24           4      1885014       0.104635    1.972381
    4     25-34          14      1957573       0.466259    9.127353
    5     35-44          43      2356649       1.386535    32.67576
    6     45-54          72      2088000       3.182232    66.44501
    7     55-64          70      1548371       5.367732    83.11241
    8     65-74         126      1447432         8.9011    128.8374
    9     75-84         136      1087524        13.1379    142.8779
   10     85+            73       335944       18.94052    63.62955

                    Strata SMR Information
                    Rate Multiplier = 100000

              ------------------SMR-----------------
                         Standard       95% Normal
      Index   Estimate      Error   Confidence Limits

          1          .          .          .          .
          2          0          .          .          .
          3   2.028005   1.014003   0.040597   4.015414
          4   1.533851   0.409939   0.730386   2.337317
          5    1.31596   0.200682   0.922631    1.70929
          6   1.083603   0.127704   0.833308   1.333898
          7   0.842233   0.100666   0.644931   1.039535
          8   0.977977   0.087125   0.807215   1.148739
          9   0.951862   0.081621   0.791887   1.111837
         10   1.147266   0.134277   0.884087   1.410444
```

Figure 6 displays these results graphically.

Figure 6 Strata SMR Plot

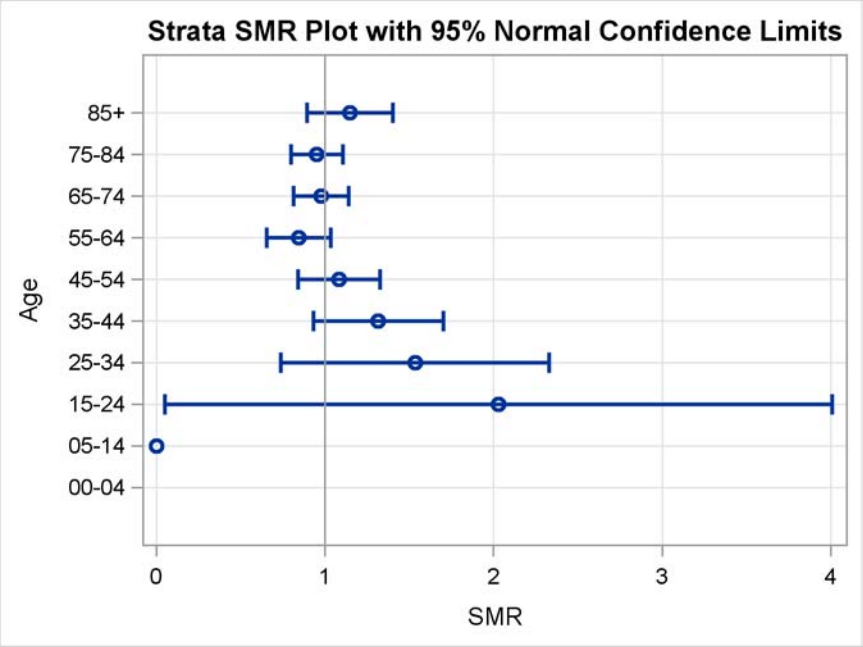


Figure 7 displays the overall SMR estimate, its confidence limits, and a test for the null hypothesis that the overall SMR equals 1.

Figure 7 Standardized Morbidity/Mortality Ratio

```
                 Standardized Morbidity/Mortality Ratio

                              --------------SMR-------------
      ------Events------                95% Normal
      Observed  Expected  Estimate    Confidence Limits

           538  528.7263   1.01754  0.931557    1.103522

                 Standardized Morbidity/Mortality Ratio

    --------------------Test of SMR=1---------------------
                               Standard
    Test        Estimate          Error           Z    Pr > |Z|

    SMR-1        0.01754       0.043869        0.40      0.6893
```

The 95% normal confidence limits contain 1, so the null hypothesis cannot be rejected.

Figure 8 contains the indirect standardized rate and related statistics.

Figure 8 Standardized Rate Estimates

```
         Standardized Rate Estimates (Indirect Standardization)
                        Rate Multiplier = 100000

                                      Reference
     Observed  Population-     Crude      Crude  Expected
       Events         Time      Rate       Rate    Events        SMR

          538     15658227  3.435893   2.636608  528.7263    1.01754

         Standardized Rate Estimates (Indirect Standardization)
                         Rate Multiplier = 100000

                 -----------Standardized Rate----------
                           Standard       95% Normal
                 Estimate     Error   Confidence Limits

                 2.682853  0.115666  2.456152  2.909554
```

The table shows that, although the crude rate in the state of Florida, 3.4359, is 30% higher than the crude rate in the US, 2.6366, the resulting standardized rate of 2.6829 is much closer to the crude rate in the US.

New QUANTSELECT Procedure

Ordinary least squares regression models the relationship between the conditional mean of a response variable with one or more covariates. Quantile regression extends that regression model to the relationship between the conditional quantiles of a response variable with one or more covariates. It is especially useful with data that are heterogeneous such that the tails and central location of the conditional distributions vary differently with the covariates. Quantile regression makes no distributional assumptions about the error term, and so it offers model robustness. It is a semi-parametric 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.

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, which provides model selection for univariate linear models. The experimental QUANTSELECT procedure includes:

- 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 players in the 1986 season; information is available for a number of measures, and 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 baseball variables are available for selection. The adaptive LASSO method is used for model selection, with AIC as the stopping criterion. Plots requested are the average check loss plot, the coefficient panel, and the criterion panel.

```
proc quantselect data=baseball plots=(acl crit 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 9 displays model information. The quantile type is single-level, the selection method is adaptive LASSO, AIC is both the select and stop criterion, and the choose criterion is SBC.

Figure 9 Model Information

```
                         The QUANTSELECT Procedure

                             Model Information

                   Data Set                      WORK.BASEBALL
                   Selection Method             Adaptive LASSO
                   Quantile Type                  Single Level
                   Select Criterion                        AIC
                   Stop Criterion                          AIC
                   Choose Criterion                        SBC
                   Test Type                Likelihood Ratio I
                   Dependent Variable                   salary
```

Figure 10 displays the selection summary information. You can see the values of AIC and AICC change as variables go into and come out of the model. The optimal value of AIC is 1057.6857 at the fifth step, which corresponds to a model with three variables: number of hits, career home runs, and division. These factors are the main factors in determining salary for the 90th percentile.

Figure 10 Selection Summary

```
                         The QUANTSELECT Procedure

           Selection stopped at a local minimum of the STOP criterion.

                             Selection Summary

           Parameter     Parameter           Number
     Step  Entered       Removed       Parameters In         AIC          AICC

        0                                          1   1219.3645     1219.3798
        1  division                                2   1199.2765     1199.3226
           East
        2  league                                  3   1200.9842     1201.0768
           National
        3  nHits                                   4   1150.8132     1150.9683
        4                league                    3   1153.0000     1153.0926
                         National
        5  crHome                                  4   1057.6857*    1057.8407*
        6  league                                  5   1059.5331     1059.7665
           National
        7                league                    4   1057.6857     1057.8407
                         National

                         * Optimal Value Of Criterion

                             Selection Summary

           Parameter     Parameter                       Model   Adjusted
     Step  Entered       Removed              SBC          R1          R1   p-Value

        0                               1222.9366      0.0000      0.0000        .
        1  division                     1206.4208      0.0806      0.0770    0.0043
           East
        2  league                       1211.7006      0.0816      0.0745    0.7335
           National
        3  nHits                        1165.1019      0.2468      0.2381    <.0001
        4                league         1163.7164      0.2347      0.2289    0.1775
                         National
        5  crHome                       1071.9743*     0.4714      0.4653*   <.0001
        6  league                       1077.3938      0.4717      0.4635    0.7519
           National
        7                league         1071.9743      0.4714      0.4653    0.7519
                         National

                         * Optimal Value Of Criterion
```

Figure 11 Selected Effects

```
Selected Effects: Intercept nHits crHome division East
```

Figure 12 displays the coefficient panel, which shows the progression of the standardized coefficients and the SBC throughout the selection process.

Figure 12 Coefficient Panel

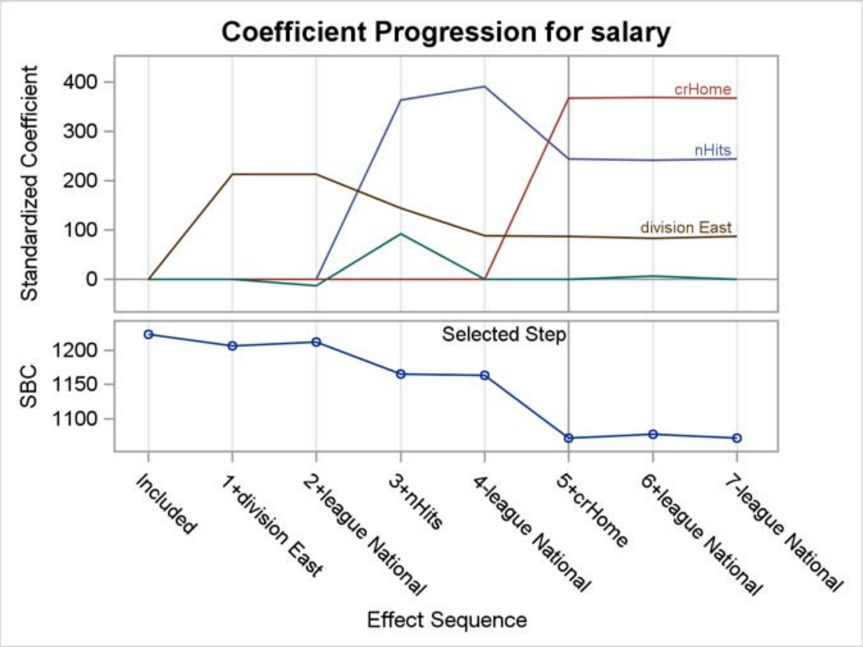


Figure 13 displays the progression of the average check loss for the selection process. It takes its lowest value at the fifth stage.

Figure 13 Average Check Loss Plot

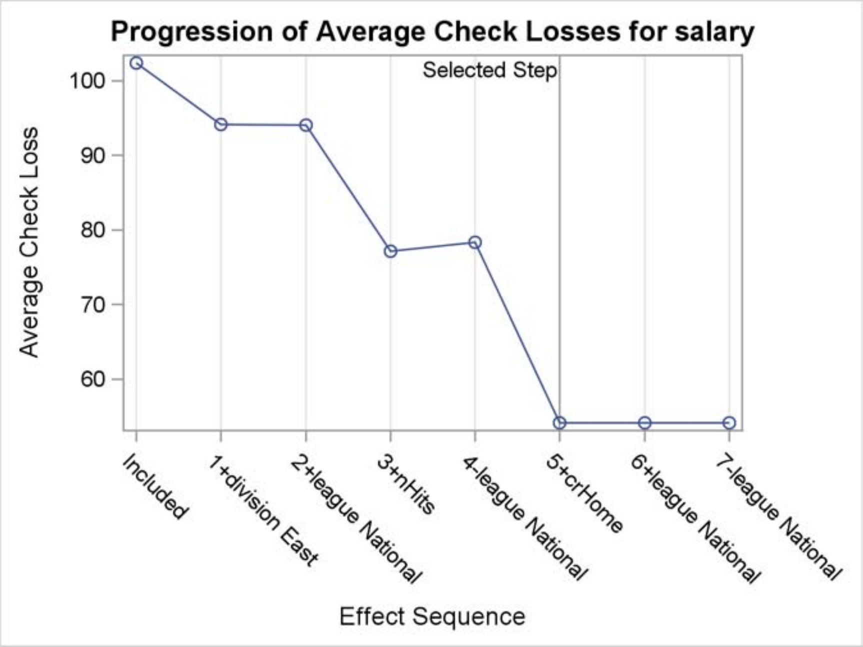


Figure 14 displays the fit criteria for the selection progression.

Figure 14 CriterionPanel

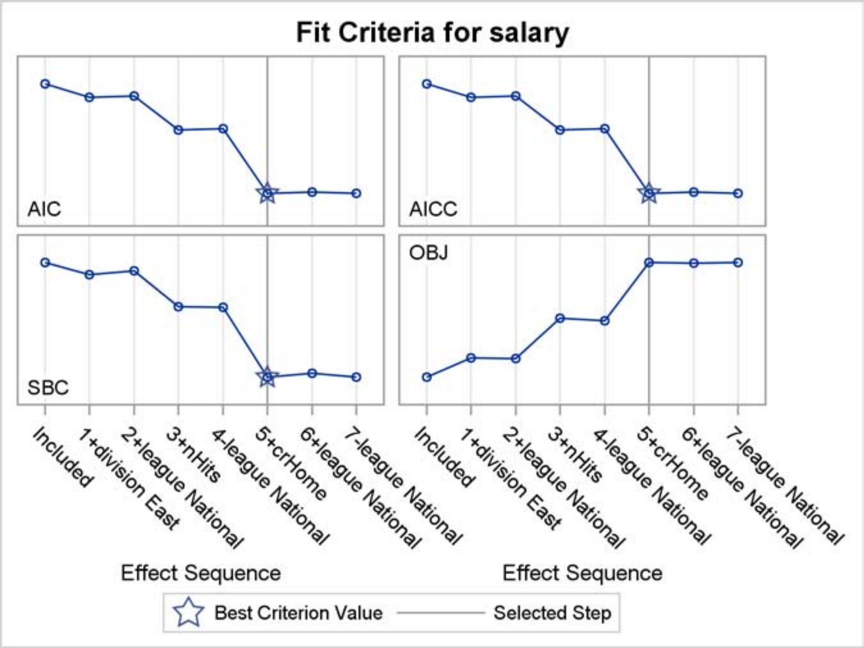
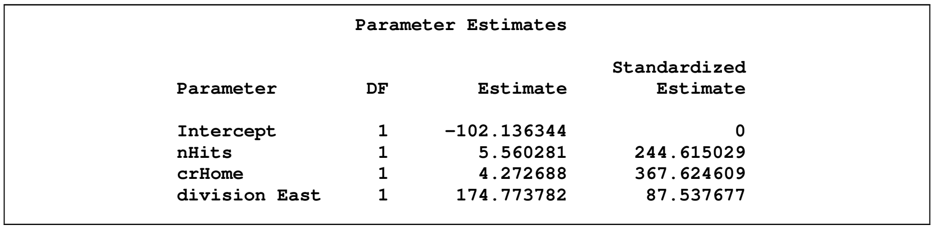


Figure 15 contains the parameter estimates and their standardized versions.

Figure 15 Parameter Estimates



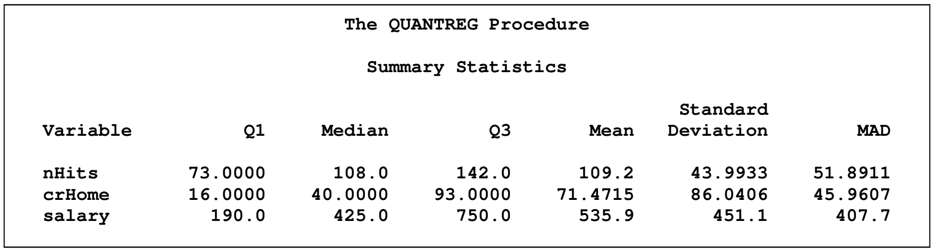
```
                    Parameter Estimates

                                                Standardized
   Parameter        DF        Estimate              Estimate

   Intercept         1     -102.136344                     0
   nHits             1        5.560281            244.615029
   crHome            1        4.272688            367.624609
   division East     1      174.773782             87.537677
```

You then perform a final analysis by using the QUANTREG procedure for the selected model:

```
proc quantreg data=baseball;
  class division;
  model Salary = nHits crHome division /
          quantile=.9;
run;
```

Figure 16 displays summary statistics for this analysis.

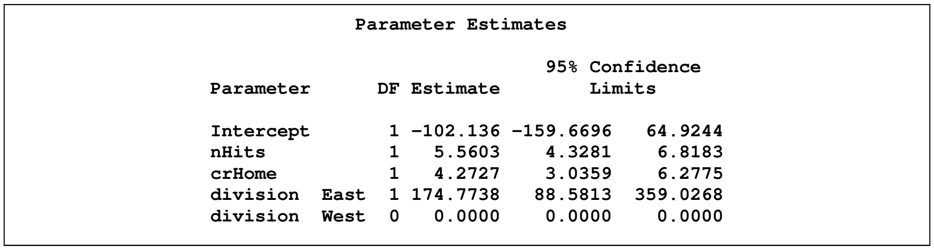
Figure 16 Summary Statistics



```
                          The QUANTREG Procedure

                            Summary Statistics

                                                           Standard
Variable          Q1      Median          Q3        Mean   Deviation          MAD

nHits        73.0000       108.0       142.0       109.2     43.9933      51.8911
crHome       16.0000     40.0000     93.0000     71.4715     86.0406      45.9607
salary         190.0       425.0       750.0       535.9       451.1        407.7
```

Figure 17 contains the parameter estimates and standard errors.

Figure 17 Parameter Estimates



```
                        Parameter Estimates

                                             95% Confidence
          Parameter       DF Estimate            Limits

          Intercept        1 -102.136 -159.6696   64.9244
          nHits            1   5.5603    4.3281    6.8183
          crHome           1   4.2727    3.0359    6.2775
          division   East  1 174.7738   88.5813  359.0268
          division   West  0   0.0000    0.0000    0.0000
```

New QUANTLIFE Procedure

Quantile regression also provides an alternative and flexible technique for the analysis of survival data. You can apply the method to right-censored responses, thus providing quantile-specific covariate effects and directly predicting lifetime. Two quantile regression approaches have been developed to account for right-censoring. Portnoy (2003) proposed a method to estimate conditional quantile functions from survival data based on the idea of the Kaplan-Meier estimator. For each quantile, this problem is framed as a weighted linear regression quantile problem that is solved for the conditional quantiles of a generalization of the Kaplan-Meier estimate. Peng and Huang (2008) developed a censored quantile regression approach based on the Nelson-Aalen estimator of the cumulative hazard function. This approach extends the martingale representation of that estimator to produce an estimating equation for conditional quantiles. Both methods can be solved with linear programming algorithms. When there are no censored observations, the Portnoy method produces the same estimates as are obtained from the QUANTREG procedure, and the Peng and Huang method produces approximately the same estimates.

The experimental QUANTLIFE procedure provides these two quantile regression methods for the analysis of survival data. PROC QUANTLIFE provides the following functionality:

- provides interior point algorithms for estimation
- enables parallel computing when multiple processors are available
- provides Wald tests for the regression parameter estimates
- produces survival plots, conditional quantile plots, and quantile process plots
- supports the EFFECT statement so it can fit regression quantile spline curves

Consider a study of primary biliary cirrhosis, a rare but fatal chronic liver disease discussed in Lin, Wei, and Ying (1993). Prognostic factors studied included age, edema, bilirubin, albumin, and prothrombin. Researchers at the Mayo Clinic 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:

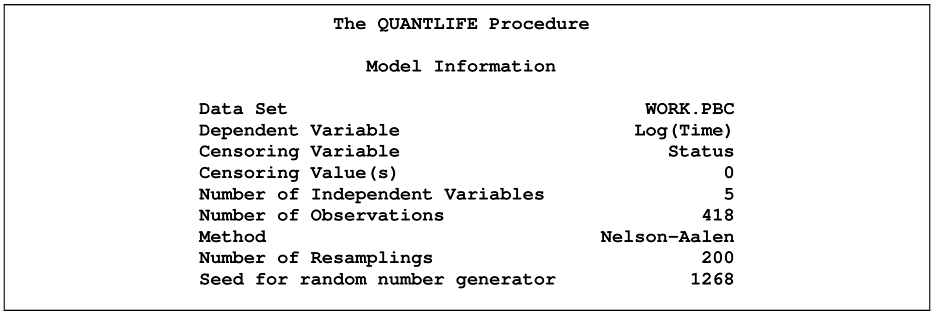
```
data pbc;
   input Time Status Age Albumin Bilirubin Edema Protime @@;
   label Time="Follow-up Time in Days";
   logAlbumin   =log(Albumin);
   logBilirubin =log(Bilirubin);
   logProtime   =log(Protime);
   datalines;
  400 1 58.7652 2.60 14.5 1.0 12.2 4500 0 56.4463 4.14  1.1 0.0 10.6
 1012 1 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. You indicate the censoring variable by crossing it with the response variable, and then you supply the censoring value in parentheses. 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 18 reports the model information. The Nelson-Aalen method is applied.

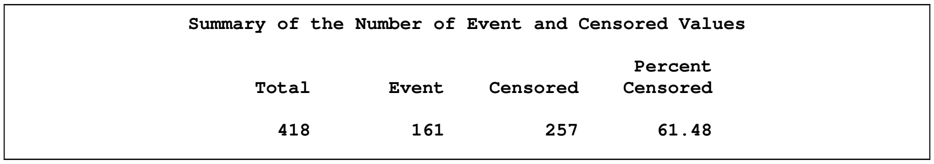
Figure 18 Model Information



```
                          The QUANTLIFE Procedure

                             Model Information

               Data Set                                   WORK.PBC
               Dependent Variable                        Log(Time)
               Censoring Variable                           Status
               Censoring Value(s)                                0
               Number of Independent Variables                   5
               Number of Observations                          418
               Method                                 Nelson-Aalen
               Number of Resamplings                           200
               Seed for random number generator               1268
```

Figure 19 reports the censoring statistics: 257 observations out of 418 observations have been censored.

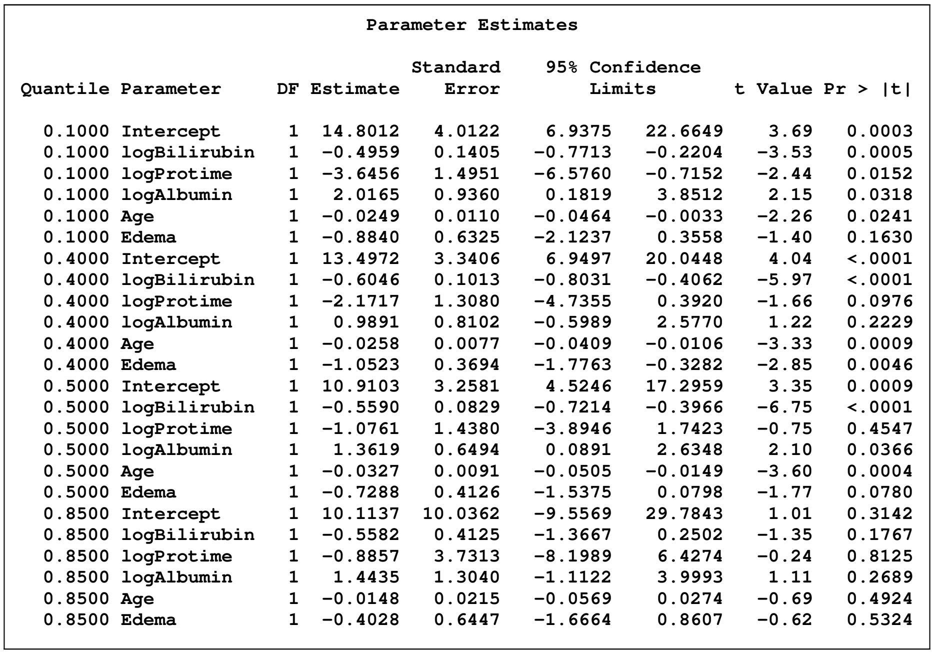
Figure 19 Censoring Summary



```
              Summary of the Number of Event and Censored Values

                                                       Percent
                    Total        Event     Censored    Censored

                      418          161          257       61.48
```

Figure 20 contains the parameter estimates. Each of the requested quantiles has its own set of parameter estimates. The confidence limits are computed by resampling methods.

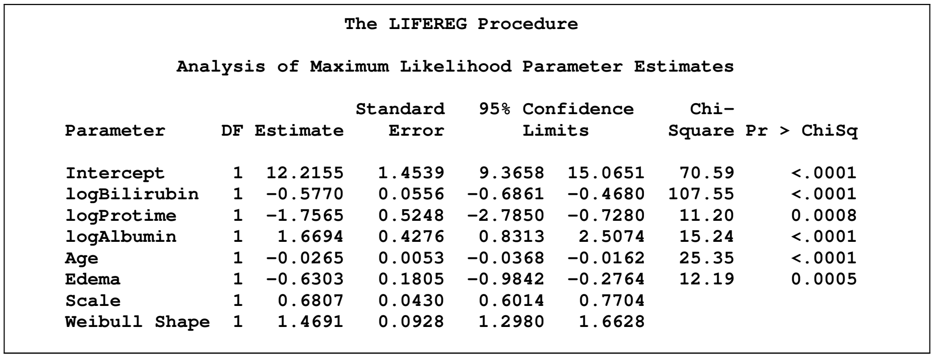
Figure 20 Parameter Estimates



```
                              Parameter Estimates

                                      Standard    95% Confidence
Quantile Parameter     DF Estimate     Error         Limits        t Value Pr > |t|

  0.1000 Intercept      1  14.8012    4.0122    6.9375   22.6649     3.69   0.0003
  0.1000 logBilirubin   1  -0.4959    0.1405   -0.7713   -0.2204    -3.53   0.0005
  0.1000 logProtime     1  -3.6456    1.4951   -6.5760   -0.7152    -2.44   0.0152
  0.1000 logAlbumin     1   2.0165    0.9360    0.1819    3.8512     2.15   0.0318
  0.1000 Age            1  -0.0249    0.0110   -0.0464   -0.0033    -2.26   0.0241
  0.1000 Edema          1  -0.8840    0.6325   -2.1237    0.3558    -1.40   0.1630
  0.4000 Intercept      1  13.4972    3.3406    6.9497   20.0448     4.04   <.0001
  0.4000 logBilirubin   1  -0.6046    0.1013   -0.8031   -0.4062    -5.97   <.0001
  0.4000 logProtime     1  -2.1717    1.3080   -4.7355    0.3920    -1.66   0.0976
  0.4000 logAlbumin     1   0.9891    0.8102   -0.5989    2.5770     1.22   0.2229
  0.4000 Age            1  -0.0258    0.0077   -0.0409   -0.0106    -3.33   0.0009
  0.4000 Edema          1  -1.0523    0.3694   -1.7763   -0.3282    -2.85   0.0046
  0.5000 Intercept      1  10.9103    3.2581    4.5246   17.2959     3.35   0.0009
  0.5000 logBilirubin   1  -0.5590    0.0829   -0.7214   -0.3966    -6.75   <.0001
  0.5000 logProtime     1  -1.0761    1.4380   -3.8946    1.7423    -0.75   0.4547
  0.5000 logAlbumin     1   1.3619    0.6494    0.0891    2.6348     2.10   0.0366
  0.5000 Age            1  -0.0327    0.0091   -0.0505   -0.0149    -3.60   0.0004
  0.5000 Edema          1  -0.7288    0.4126   -1.5375    0.0798    -1.77   0.0780
  0.8500 Intercept      1  10.1137   10.0362   -9.5569   29.7843     1.01   0.3142
  0.8500 logBilirubin   1  -0.5582    0.4125   -1.3667    0.2502    -1.35   0.1767
  0.8500 logProtime     1  -0.8857    3.7313   -8.1989    6.4274    -0.24   0.8125
  0.8500 logAlbumin     1   1.4435    1.3040   -1.1122    3.9993     1.11   0.2689
  0.8500 Age            1  -0.0148    0.0215   -0.0569    0.0274    -0.69   0.4924
  0.8500 Edema          1  -0.4028    0.6447   -1.6664    0.8607    -0.62   0.5324
```

For comparison purposes, consider the table of parameter estimates shown in Figure 21. These were produced by the LIFEREG procedure using the default Weibull distribution; PROC LIFEREG fits the accelerated failure time model, which assumes that the effect of independent variables is multiplicative on the event time. The variable LOGPROTIME has a very small p-value for this analysis. However, the same variable has much larger p-values for the quantile regression analysis; they are 0.4547 for the 0.5 quantile and 0.8125 for the 0.85 quantile. The p-values are much smaller for the lower quantiles. Apparently, the effect of this covariate depends on which side of the response distribution is being modeled.

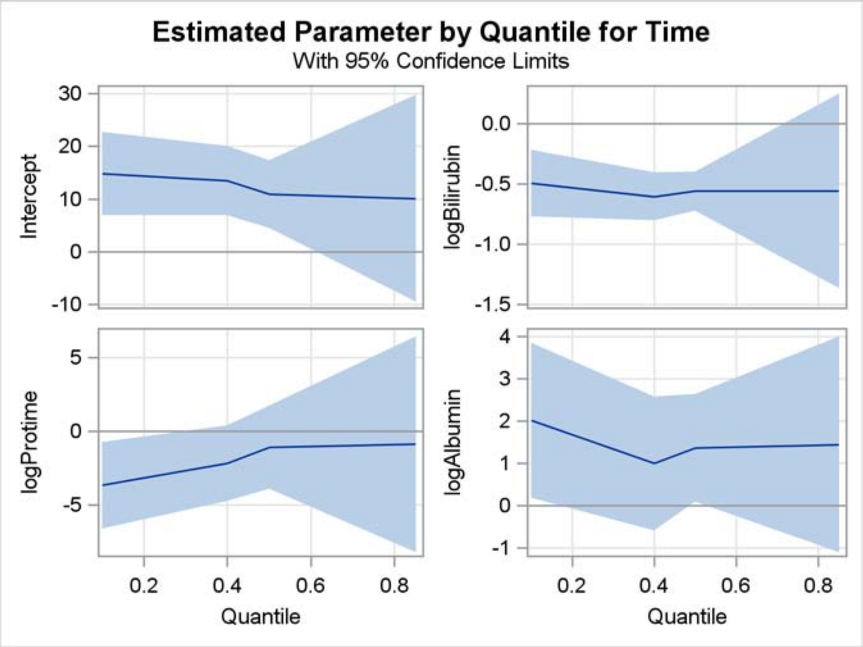
Figure 21 Parameter Estimates



```
                          The LIFEREG Procedure

             Analysis of Maximum Likelihood Parameter Estimates

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

Intercept       1  12.2155    1.4539   9.3658  15.0651   70.59     <.0001
logBilirubin    1  -0.5770    0.0556  -0.6861  -0.4680  107.55     <.0001
logProtime      1  -1.7565    0.5248  -2.7850  -0.7280   11.20     0.0008
logAlbumin      1   1.6694    0.4276   0.8313   2.5074   15.24     <.0001
Age             1  -0.0265    0.0053  -0.0368  -0.0162   25.35     <.0001
Edema           1  -0.6303    0.1805  -0.9842  -0.2764   12.19     0.0005
Scale           1   0.6807    0.0430   0.6014   0.7704
Weibull Shape   1   1.4691    0.0928   1.2980   1.6628
```

This behavior of the covariate coefficients is illustrated in the quantiles plot in Figure 22. 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 22 Estimated Parameter by Quantiles Plot



Finally, Figure 23 displays the survival probabilities for the range of survival times, and Figure 24 displays the predicted quantiles.

Figure 23 Survival Plot

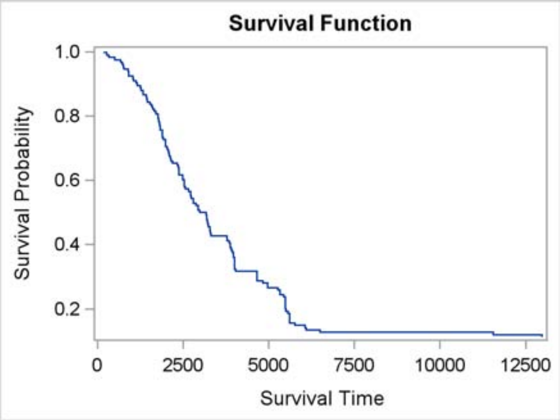
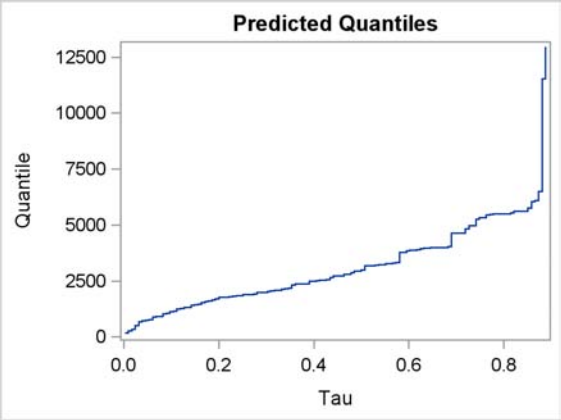


Figure 24 Quantile Plot



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 multivariate adaptive regression splines method is similar to recursive partitioning models (Breiman et al. 1984). PROC ADAPTIVEREG grows an overfitted model with the fast update algorithm (Friedman 1993) and prunes it back with the backward selection technique. During the forward selection process, bases are created from interactions between existing parent bases and nonparametric transformations of continuous or classification variables as candidate effects. After the model grows to a certain size, the backward selection process begins by deleting selected bases. The deletion continues until the null model is reached, and then an overall best model is chosen based on some goodness-of-fit criteria.

The ADAPTIVEREG procedure supports models with classification variables (Friedman 1991a), and it provides options for improving modeling speed. PROC ADAPTIVEREG extends the method to data with response distributions from the exponential family, such as binomial and Poisson (Buja et al. 1991). PROC ADAPTIVEREG is multithreaded so it takes advantage of multiple processors.

PROC ADAPTIVEREG

- 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

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.

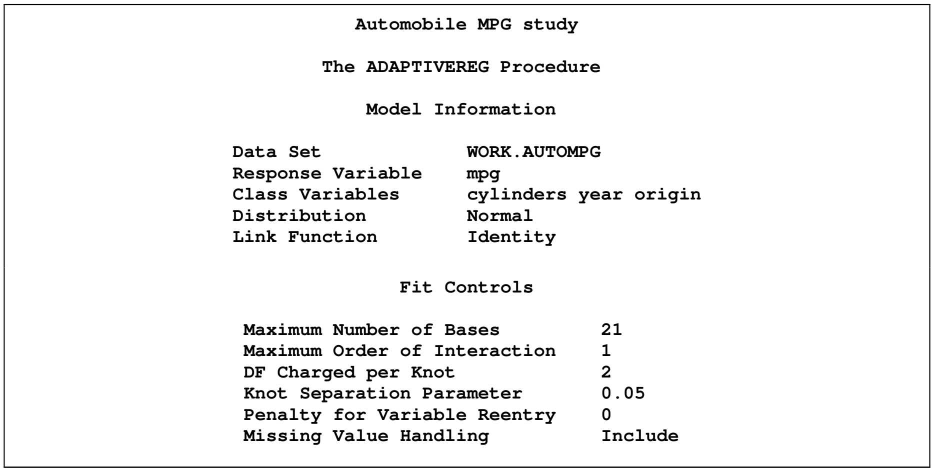
The dependency of vehicle fuel efficiency on these factors might be nonlinear. Dependency structures within the predictors might also mean that some of the predictors are redundant. For example, a model with more cylinders is likely to have more horsepower. The object of this analysis is to explore the 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;
```

PROC ADAPTIVEREG summarizes important information about the fitted model in Figure 25.

Figure 25 Model Information and Fit Controls



```
                         Automobile MPG study

                      The ADAPTIVEREG Procedure

                          Model Information

                 Data Set               WORK.AUTOMPG
                 Response Variable      mpg
                 Class Variables        cylinders year origin
                 Distribution           Normal
                 Link Function          Identity


                             Fit Controls

                  Maximum Number of Bases         21
                  Maximum Order of Interaction    1
                  DF Charged per Knot             2
                  Knot Separation Parameter       0.05
                  Penalty for Variable Reentry    0
                  Missing Value Handling          Include
```

In addition to listing the classification variables in the “Model Information” table, PROC ADAPTIVEREG displays class-level information about the classification variables specified in the CLASS statement. Figure 26 lists the levels of the classification variables CYLINDERS, YEAR, and ORIGIN. Although the values of CYLINDERS and YEAR are naturally ordered, they are treated as ordinary classification variables.

Figure 26 Class Level Information

```
                      Class Level Information

   Class        Levels    Values

   cylinders         5    3 4 5 6 8
   year             13    70 71 72 73 74 75 76 77 78 79 80 81 82
   origin            3    1 2 3
```

The “Fit Statistics” table in Figure 27 lists summary statistics for the fitted regression spline model. Because the final model is essentially a linear model, several naïve statistics are reported as if the model were fitted with predetermined basis functions. However, the determination of basis functions and the model selection process are highly nonlinear, so additional statistics that incorporate the extra sources of degrees of freedom are also displayed. These statistics include effective degrees of freedom, the GCV criterion, and the GCV R-Square.

Figure 27 Fit Statistics

```
                   Fit Statistics

    Naive R-Square                          0.853201
    Naive Adjusted R-Square                 0.850290
    Naive Mean Square Error                 9.185230
    Effective Degrees of Freedom           15.000000
    GCV                                     9.777318
    GCV R-Square                            0.841081
```

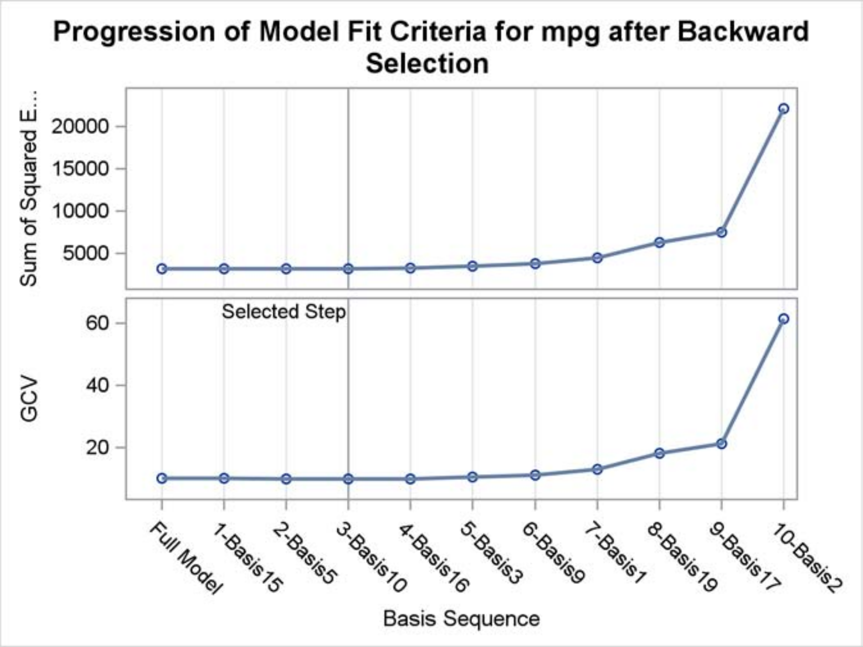
The “Parameter Estimates” table in Figure 28 displays parameter estimates for constructed basis functions in addition to each function’s construction component. For example, BASIS1 has an estimate of –0.003242. It is constructed from a parent basis function BASIS0 (intercept) and a linear spline function 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 28 Parameter Estimates

```
                Regression Spline Model after Backward Selection

Name       Coefficient   Parent   Variable               Knot   Levels

Basis0       17.862071            Intercept
Basis1       -0.003242   Basis0   weight          3139.000000
Basis2        0.010344   Basis0   weight          3139.000000
Basis3        2.045223   Basis0   year                          10 12 11 9 3 8 7
Basis9        2.539889   Basis0   acceleration      20.700000
Basis16      -0.241712   Basis0   displacement      85.000000
Basis17       4.767534   Basis0   year                          3 10 12 11 9
Basis19      -6.203451   Basis0   year                          3 9
```

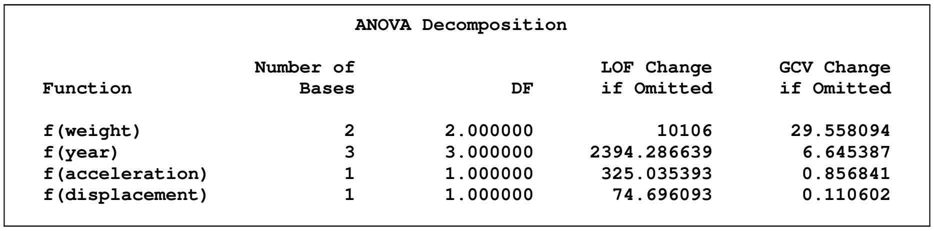
During the model construction and selection process, some basis function terms are removed. You can view the backward elimination process in the selection plot shown in Figure 29. The plot displays how the model sum of squared error and the corresponding GCV criterion change during the backward elimination process. The sum of squared error increases as more basis functions are removed from the full model. The GCV criterion decreases at first when three basis functions are dropped, and it increases afterwards. The vertical line indicates the selected model with the minimum GCV value. The model is formed by dropping BASIS15, BASIS5, and BASIS10 from the full model.

Figure 29 Selection Plot



The final model is an additive model. Basis functions of the same variables can be grouped together to form functional components. The "ANOVA Decomposition" table in Figure 30 shows four functional components and their contribution to the final model. The functional component of weight contributes the most, while the component of displacement contributes the least.

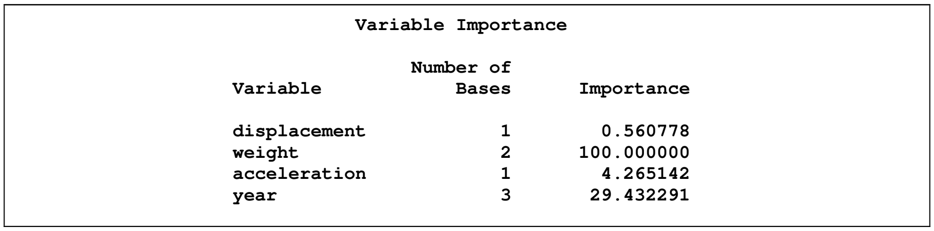
Figure 30 ANOVA Decomposition



```
                                ANOVA Decomposition

                     Number of                         LOF Change       GCV Change
Function                 Bases              DF        if Omitted       if Omitted

f(weight)                    2        2.000000             10106        29.558094
f(year)                      3        3.000000       2394.286639         6.645387
f(acceleration)              1        1.000000        325.035393         0.856841
f(displacement)              1        1.000000         74.696093         0.110602
```

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 31 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 31 Variable Importance



```
                     Variable Importance

                          Number of
        Variable              Bases      Importance

        displacement              1        0.560778
        weight                    2      100.000000
        acceleration              1        4.265142
        year                      3       29.432291
```

The component panel in Figure 32 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 32 Component Panel

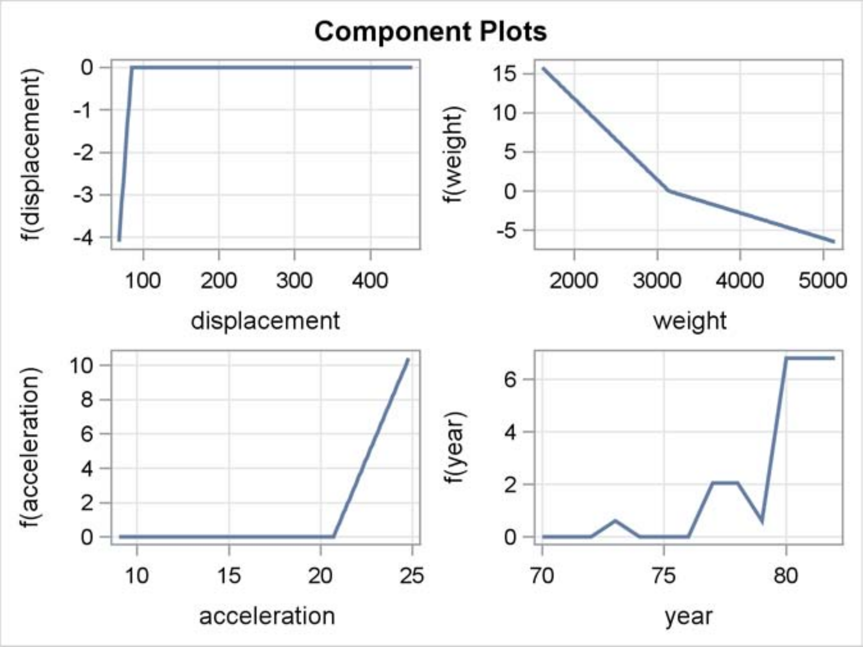
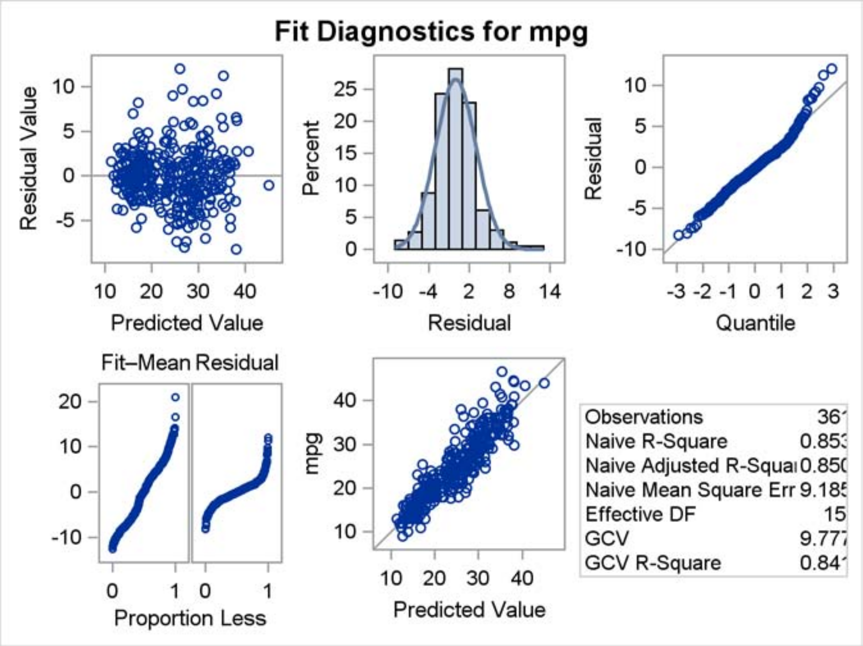


Figure 33 shows a panel of fit diagnostics for the selected model; all of these diagnostics indicate a reasonable fit.

Figure 33 Diagnostics Panel



Observations	361
Naive R-Square	0.853
Naive Adjusted R-Squa	0.850
Naive Mean Square Err	9.185
Effective DF	15
GCV	9.777
GCV R-Square	0.841

Finite Mixture Models

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 FMM procedure becomes production with SAS/STAT 12.1. In addition, it adds the truncated normal and truncated negative binomial distributions as well as support for output on both the probability and count scales.

Updated Frailty Models in Survival Analysis

When experimental units are clustered, the failure times of units within a cluster tend to be correlated. One approach is to account for within-cluster correlation by using a shared frailty model in which the cluster effects are incorporated into the model as 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.

SAS/STAT 12.1 also provides a Bayesian analysis of the shared frailty model.

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

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

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

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

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

The frailty can be distribution as gamma or lognormal:

Frailty	Distribution Details		
Gamma	$u_i \sim G\left(\frac{1}{\theta}, \frac{1}{\theta}\right)$	$E(u_i) = 1$	$V(u_i) = \theta$
LogNormal	$\gamma_i \sim N(0, \theta)$	$E(\gamma_i) = 0$	$V(\gamma_i) = \theta$

The following example illustrates the use of the Bayesian frailty model to assess whether laser treatment delays the occurrence of blindness in high risk diabetic patients. One eye of each patient is treated with laser photocoagulation, and the other eye is treated with standard remedies. Since juvenile and adult diabetes have very different courses, it is also desirable to examine how the age of onset of diabetes might affect the time of blindness. Since there are no biological differences between the left eye and the right eye, it is natural to assume a common baseline hazard function for the failure times of the left and right eyes. Each patient is a cluster that contributes two observations to the input data set, one for each eye.

The following DATA step creates the data set BLIND. Variables include those for ID, time to blindness, status for blindness, treatment, and type of diabetes.

```
proc format;
   value type 0='Juvenile' 1='Adult';
   value Rx  1='Laser' 0='Others';
run;
data Blind;
input ID Time Status dty trt @@;
Type= put(dty, type.);
Treat= put(trt, Rx.);
```

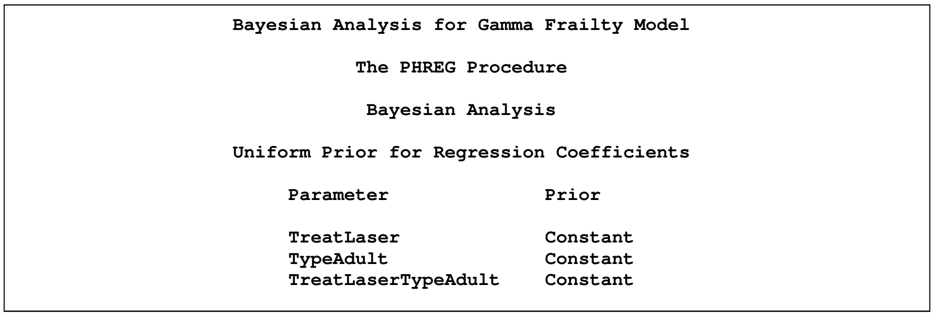
```
datalines;
  5 46.23 0 1 1    5 46.23 0 1 0   14 42.50 0 0 1   14 31.30 1 0 0
 16 42.27 0 0 1   16 42.27 0 0 0   25 20.60 0 0 1   25 20.60 0 0 0
 29 38.77 0 0 1   29  0.30 1 0 0   46 65.23 0 0 1   46 54.27 1 0 0
...
...
```

The following SAS statements request the Bayesian frailty model. Essentially, you add the BAYES statement. The DISPERSIONPRIOR=IGAMMA option specifies an inverse gamma distribution IG(3, 3) for the dispersion parameter for the frailty. No prior is specified for the regression coefficents, so the uniform prior is used by default. In the RANDOM statement, the option SOLUTION(2 4) requests that Bayesian summary statistics and diagnostics be computed for the second and fourth random effect parameters.

```
proc phreg data=Blind;
   class ID Treat Type;
   model Time*Status(0)=Treat|Type;
   random ID / dist=gamma solution (2 4);
   bayes seed=1 dispersionprior=igamma (shape=3, scale=3);
   title 'Bayesian Analysis for Gamma Frailty Model';
run;
```

Figure 34 displays the priors for the regression coefficients. By default, uniform priors are used.

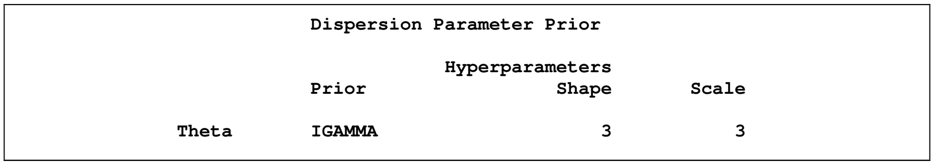
Figure 34 Coefficient Priors



```
                  Bayesian Analysis for Gamma Frailty Model

                            The PHREG Procedure

                             Bayesian Analysis

                  Uniform Prior for Regression Coefficients

                      Parameter               Prior

                      TreatLaser              Constant
                      TypeAdult               Constant
                      TreatLaserTypeAdult     Constant
```

Figure 35 displays the dispersion parameter prior, which was chosen to be inverse gamma (3, 3).

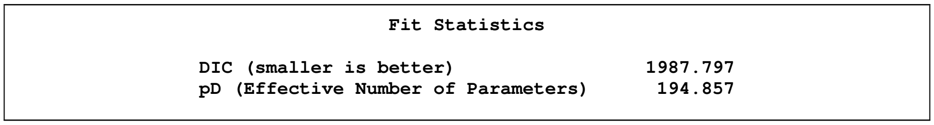
Figure 35 Dispersion Prior



```
                        Dispersion Parameter Prior

                                   Hyperparameters
                        Prior                Shape       Scale

             Theta      IGAMMA                   3           3
```

Figure 36 displays the fit statistics.

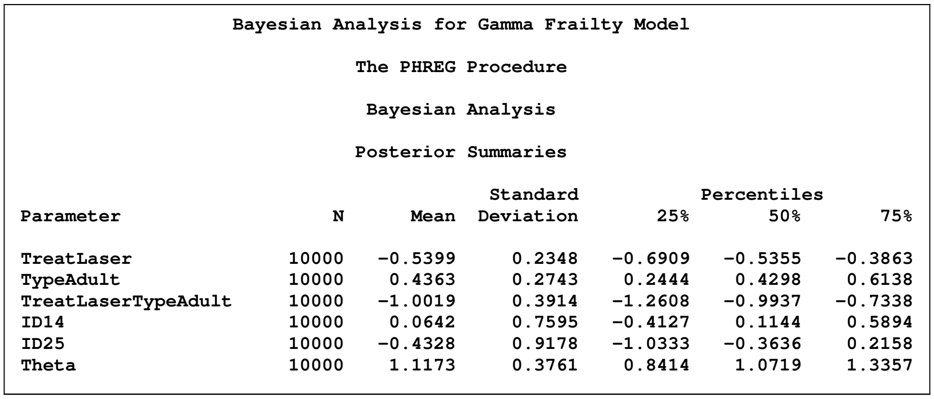
Figure 36 Fit Statistics



```
                              Fit Statistics

               DIC (smaller is better)                 1987.797
               pD (Effective Number of Parameters)      194.857
```

Figure 37 reports the posterior summaries. These values are similar to the parameter estimates obtained for the frequentist frailty analysis assuming the frailties are distributed as gamma.

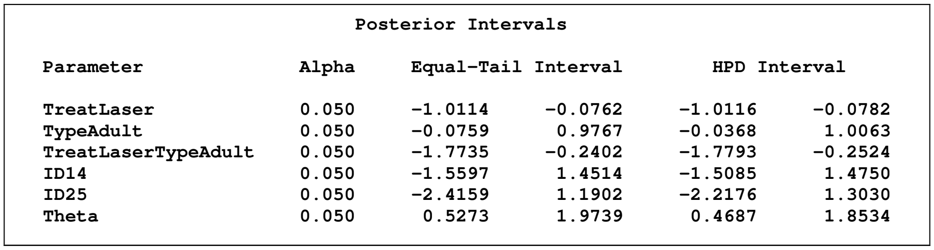
Figure 37 Posterior Summaries



```
                      Bayesian Analysis for Gamma Frailty Model

                                 The PHREG Procedure

                                  Bayesian Analysis

                                 Posterior Summaries

                                                 Standard               Percentiles
Parameter                    N        Mean      Deviation        25%          50%        75%

TreatLaser               10000     -0.5399       0.2348     -0.6909      -0.5355    -0.3863
TypeAdult                10000      0.4363       0.2743      0.2444       0.4298     0.6138
TreatLaserTypeAdult      10000     -1.0019       0.3914     -1.2608      -0.9937    -0.7338
ID14                     10000      0.0642       0.7595     -0.4127       0.1144     0.5894
ID25                     10000     -0.4328       0.9178     -1.0333      -0.3636     0.2158
Theta                    10000      1.1173       0.3761      0.8414       1.0719     1.3357
```

Figure 38 displays the credible intervals for the posterior parameters.

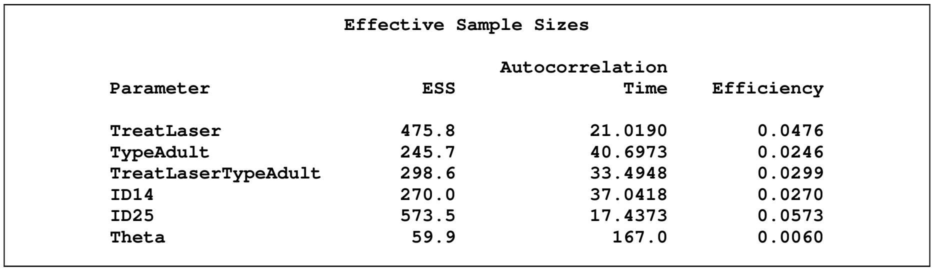
Figure 38 Posterior Intervals



```
                                 Posterior Intervals

  Parameter                Alpha      Equal-Tail Interval          HPD Interval

  TreatLaser               0.050      -1.0114     -0.0762     -1.0116     -0.0782
  TypeAdult                0.050      -0.0759      0.9767     -0.0368      1.0063
  TreatLaserTypeAdult      0.050      -1.7735     -0.2402     -1.7793     -0.2524
  ID14                     0.050      -1.5597      1.4514     -1.5085      1.4750
  ID25                     0.050      -2.4159      1.1902     -2.2176      1.3030
  Theta                    0.050       0.5273      1.9739      0.4687      1.8534
```

Figure 39 includes the effective sample sizes.

Figure 39 Effective Sample Sizes

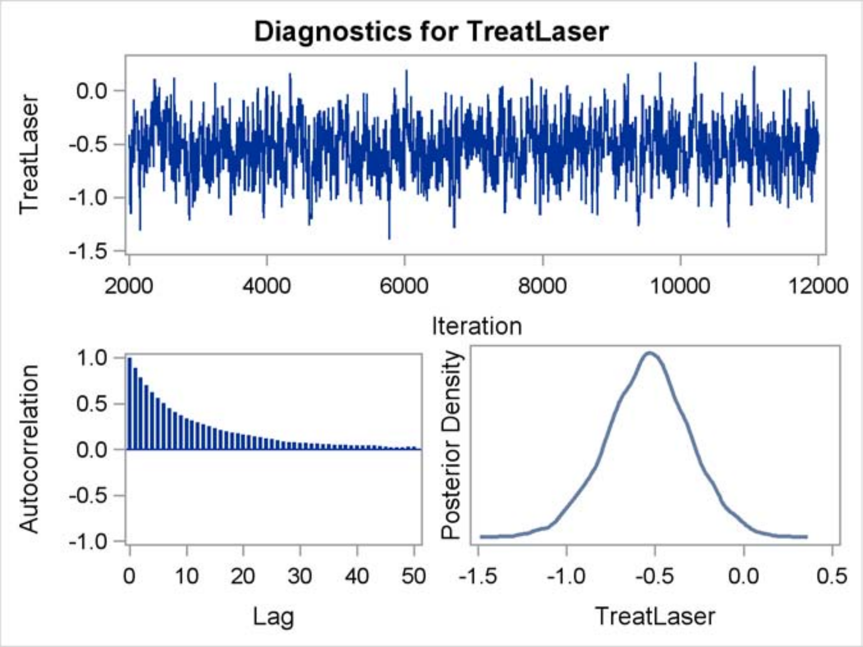


```
                                Effective Sample Sizes

                                             Autocorrelation
        Parameter                    ESS                Time     Efficiency

        TreatLaser                 475.8             21.0190         0.0476
        TypeAdult                  245.7             40.6973         0.0246
        TreatLaserTypeAdult        298.6             33.4948         0.0299
        ID14                       270.0             37.0418         0.0270
        ID25                       573.5             17.4373         0.0573
        Theta                       59.9               167.0         0.0060
```

These values are reasonable for this analysis.

The trace plot for TREATLASER is displayed in Figure 40. It shows reasonable mixing. The trace plots for the other parameters were acceptable.

Figure 40 Trace Plot



Updates to Bayesian Capabilities

Bayesian capabilities continue to grow in SAS/STAT software. These capabilities are available through two channels—as additional capabilities in existing procedures, with the BAYES statement, and through a general modeling paradigm with the MCMC procedure. Besides the availability of Bayesian frailty models in PROC PHREG, the Gamerman algorithm becomes the default sampling mechanism in the GENMOD procedure, except when you have conjugacy in the linear models.

In addition, the MCMC procedure has been enhanced in many different ways. The highlights are:

- The RANDOM statement supports arbitrary hierarchy.
- The MODEL statement supports missing value sampling.
- More conjugate sampling algorithms are available.
- Conjugate samplers now apply to random-effects parameters and missing value parameters, not just model parameters.
- A slice sampler is now available.

In addition, the MCMC procedure no longer uses optimization to find starting values when the sampling algorithms used are conjugate and/or direct, which can improve performance. You can now submit a combination of MODEL and RANDOM statements without needing a PARMS statement, and several postprocessing macros provide summary and diagnostic information. Additional distributions are available in the RANDOM statement, and the multivariate normal distribution with autocorrelation covariance structure is available for the PRIOR, RANDOM, and MODEL statements.

The MCMC procedure also includes facilities for managing missing data. Previously, missing responses for the dependent variable in the analysis resulted in those observations being deleted. Beginning with SAS/STAT 12.1, missing values for the responses are automatically sampled. In addition, the MCMC procedure can now accommodate missing values for the covariates. This new capability is illustrated with the following example, which uses the MCMC procedure to fit Bayesian logistic regression models to analyze air pollution data.

Researchers studied the effects of air pollution on respiratory disease in children. The response variable (Y) represented whether a child exhibited wheezing symptoms; it was recorded as 1 for symptoms exhibited and 0 for no symptoms exhibited. City of residence (X1) and maternal smoking status (X2) were the explanatory variables. The variable X1 was coded as 1 if the child lived in the more polluted city, Steel City, and 0 if the child lived in Green Hills. The variable X2 was the number of cigarettes the mother smoked per day. Both the covariates contain missing values: 17 for X1 and 30 for X2, respectively.

This example illustrates the treatment of missing at random (MAR) data by ignoring the missing mechanism. In other words, the missingness is assumed to depend only on the observed values, and not on the missing values, which implies that the modeling of the missingness can be ignored. You can model nonignorable missing data, also called MNAR (missing not at random), with the MCMC procedure. See Little and Rubin (2002) for further information about missing data analysis.

Suppose you want to fit a Bayesian logistic regression model for whether the subject develops wheezing symptoms with density as

$$\begin{aligned} Y_i &\sim \text{binary}(p_i) \\ \text{logit}(p_i) &= \beta_0 + \beta_1 \cdot X1_i + \beta_2 \cdot X2_i \end{aligned}$$

for the $i = 1, ..., 390$ subjects.

With this model, you can write the odds ratio for comparing Steel City to Green Hills as follows:

$$\text{OR}_{X1} = \exp(\beta_1)$$

The odds ratio is useful for interpreting how the odds of developing a wheeze change for a child living in the more polluted city. Similarly, the odds ratio for the maternal smoking effect is written as:

$$\text{OR}_{X2} = \exp(\beta_2)$$

The complete data likelihood function for each of the subjects is

$$p(Y_i|\beta_0, \beta_1, \beta_2, X1_{mis,i}, X2_{mis,i}, X1_{obs,i}, X2_{obs,i}) = \text{binary}(p_i)$$

where $p(\cdot|\cdot)$ denotes a conditional probability density. The binary density is evaluated at the specified value of Y_i and corresponding mean parameter p_i. The three parameters in the complete data likelihood are β_0, β_1, and β_2, which correspond to an intercept, adjustment for living in Steel City, and a slope for maternal smoking, respectively.

The covariates X1 and X2 are written in terms of whether they were missing ($X1_{mis}$ and $X2_{mis}$) or observed ($X1_{obs}$ and $X2_{obs}$). The goal is to make inferences from the observed data likelihood

$$p(Y_i|\beta_0, \beta_1, \beta_2, X1_{obs,i}, X2_{obs,i})$$

by multiplying the conditional distribution $p(X1_{mis,i}, X2_{mis,i}|X1_{obs,i}, X2_{obs,i})$ by the likelihood and integrating over the missing observations. To make inferences from the observed data likelihood, you need to specify a distribution for the missing covariates $p(X1_{mis,i}, X2_{mis,i}|X1_{obs,i}, X2_{obs,i}, \boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ represents the hyperparameters in the missing data distributions. Suppose you specify a joint distribution of X1 and X2 in terms of the product of a conditional and marginal distribution; that is,

$$p(X1_{mis}, X2_{mis}|\boldsymbol{\alpha}) = p(X1_{mis}|X2_{mis}, \alpha_{10}, \alpha_{11})p(X2_{mis}|\alpha_{20})$$

For this example, say $p(X1_{mis,i}|X2_{mis.i}, \alpha_{10}, \alpha_{11})$ is a logistic regression and $p(X2_{mis,i}|\alpha_{20})$ is a Poisson distribution. You treat the missing covariates as parameters, and you place prior distributions on them and their hyperparameters.

Suppose you place the following prior distributions on the three regression parameters, the missing covariates, and hyperparameters:

$$\begin{aligned} \pi(\beta_0), \pi(\beta_1), \pi(\beta_2) &= \text{normal}(0, \sigma^2 = 10) \\ p(X1_{mis,i}|X2_i, \alpha_{10}, \alpha_{11}) &= \text{binary}(p_{c,i}) \\ \text{logit}(p_{c,i}) &= \alpha_{10} + \alpha_{11} \cdot X2_i \\ \pi(\alpha_{10}), \pi(\alpha_{11}) &= \text{normal}(0, \sigma^2 = 10) \\ p(X2_{mis,i}|\alpha_{20}) &= \text{Poisson}(e^{\alpha_{20}}) \\ \pi(\alpha_{20}) &= \text{normal}(0, \sigma^2 = 2) \end{aligned}$$

where $\pi(\cdot)$ indicates a prior distribution.

The following SAS statements create the data set AIR:

```
data air;
   input y x1 x2;
   datalines;
0 0  0
0 0  0
0 1  0
0 0  0
0 0 11
0 1  7
0 0  8
0 1 10
0 1  9
0 0  0
....
....
;
```

The next set of SAS statements fit a Bayesian logistic regression with missing covariates. The SEED= option specifies a seed for the random number generator, which guarantees the reproducibility of the Markov chain. The NMC= option specifies the number of posterior simulation iterations. The MONITOR= option outputs analysis on selected variables of interest in the program. The STATS= option outputs posterior summary and interval statistics. The DIAG= option requests the effective sample sizes of parameters.

```
proc mcmc data=air seed=1181 nmc=10000 monitor=(_parms_ orx1 orx2)
   stats=(summary interval) diag=ess;
   parms beta0 -1 beta1 0.1 beta2 .01;
   parms alpha10 0 alpha11 0 alpha20 0;

   prior beta: alpha1: ~ normal(0,var=10);
   prior alpha20 ~ normal(0,var=2);

   beginnodata;
   pm = exp(alpha20);
   orx1 = exp(beta1);
   orx2 = exp(beta2);
   endnodata;
   model x2 ~ poisson(pm) monitor=(1 3 10);
   p1 = logistic(alpha10 + alpha11 * x2);
   model x1 ~ binary(p1) monitor=(4 10 16);
   p = logistic(beta0 + beta1*x1 + beta2*x2);
   model y ~ binary(p);
run;
```

The PARMS statements specify the parameters in the model and assign initial values to each of them. The PRIOR statements specify priors for all the model parameters. The notation BETA: and ALPHA: in the PRIOR statements are shorthand for all variables that start with 'BETA' and 'ALPHA,' respectively. The shorthand notation is not necessary, but it makes your code succinct.

The BEGINNODATA and ENDNODATA statements enclose three programming statements that calculate the Poisson mean PM, and the two odds ratios (ORX1 and ORX2). These enclosed statements are independent of any data set variables, and they are executed once per iteration to reduce unnecessary observation-level computations.

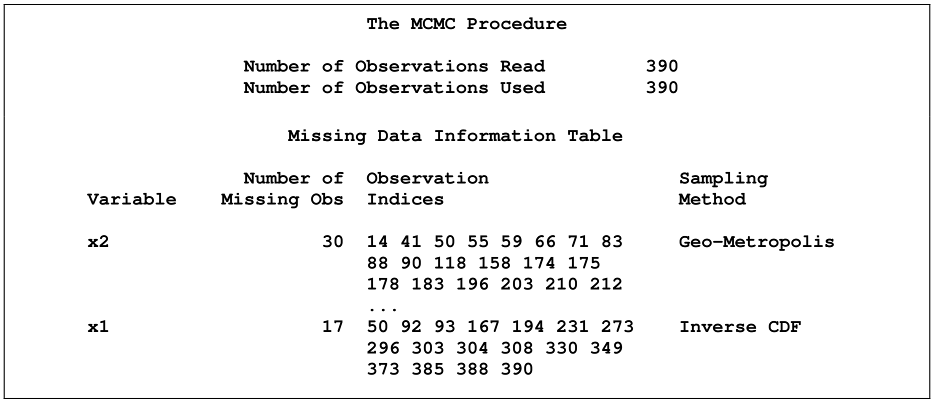
The first MODEL statement assigns a Poisson likelihood with mean PM to X2. The statement allows missing values in the variable, creates one variable for each of the missing values, and augments them automatically. In each iteration, PROC MCMC samples missing values from their posterior distributions and incorporates them as part of the simulation. By default, the procedure does not output analyses of the posterior samples of the missing values. You can use the MONITOR= option to choose the missing values that you want to monitor. In the example, the first, third, and tenth missing values are monitored.

The P1 assignment statement calculates $p_{c.i}$. The second MODEL statement assigns a binary likelihood with probability p1, and monitors the fourth, tenth, and sixteenth missing values in covariate X1.

The P1 assignment statement calculates p_i in the logistic model. The third MODEL statement specifies the complete data likelihood function for Y.

Figure 41 displays the "Number of Observations" and "Missing Data Information" tables. The "Number of Observations" table lists the number of observations read from the DATA= data set and the number of observations used in the analysis. No observations were omitted from the data set in the analysis. The “Missing Data Information” table lists the variables that contain missing values (X1 and X2), the number of missing observations in each variable, the observation indices of these missing values, and the sampling algorithms used. By default, the first 20 observation indices of each variable are listed.

Figure 41 Observation Information and Missing Data Information

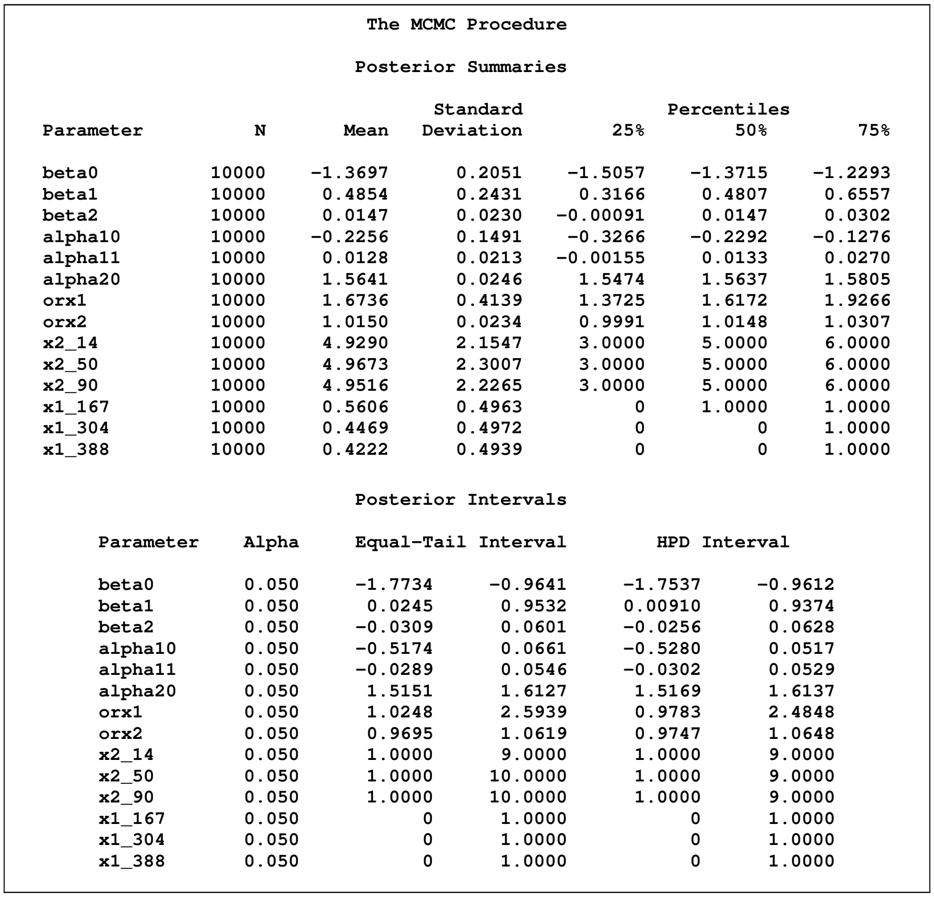


```
                            The MCMC Procedure

                 Number of Observations Read          390
                 Number of Observations Used          390

                     Missing Data Information Table

                 Number of  Observation                   Sampling
Variable       Missing Obs  Indices                       Method

x2                      30  14 41 50 55 59 66 71 83       Geo-Metropolis
                            88 90 118 158 174 175
                            178 183 196 203 210 212
                            ...
x1                      17  50 92 93 167 194 231 273      Inverse CDF
                            296 303 304 308 330 349
                            373 385 388 390
```

There are 30 missing values for the variable X2 and 17 missing values for variable X1. Internally, PROC MCMC creates 30 and 17 variables for the missing values in X2 and X1, respectively. The default naming convention of these missing values is determined by concatenating the response variable with the observation number. For example, the first missing value in X2 is the fourteenth observation, and the corresponding variable is X2_14.

Figure 42 displays summary and interval statistics for each parameters, the odds ratios, and the monitored missing values.

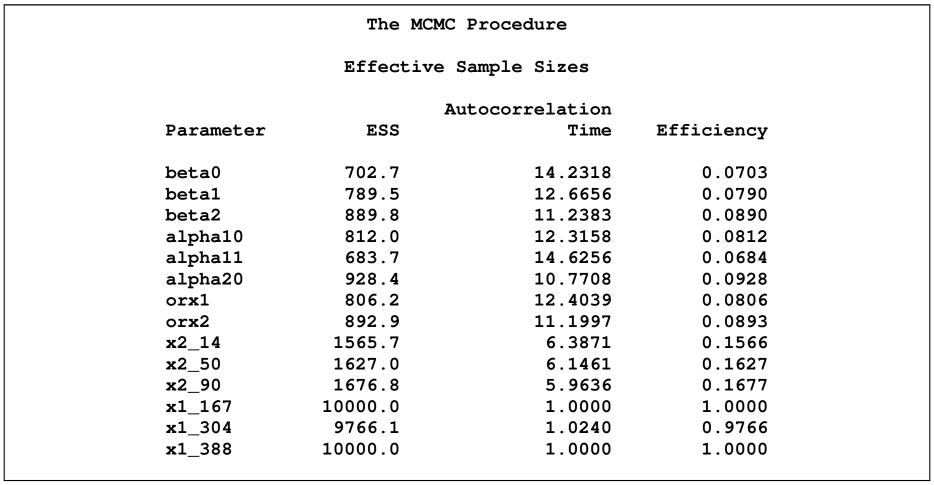
Figure 42 Posterior Summary and Interval Statistics



```
                               The MCMC Procedure

                               Posterior Summaries

                                      Standard                Percentiles
Parameter          N       Mean     Deviation        25%          50%         75%

beta0          10000    -1.3697        0.2051    -1.5057      -1.3715     -1.2293
beta1          10000     0.4854        0.2431     0.3166       0.4807      0.6557
beta2          10000     0.0147        0.0230   -0.00091       0.0147      0.0302
alpha10        10000    -0.2256        0.1491    -0.3266      -0.2292     -0.1276
alpha11        10000     0.0128        0.0213   -0.00155       0.0133      0.0270
alpha20        10000     1.5641        0.0246     1.5474       1.5637      1.5805
orx1           10000     1.6736        0.4139     1.3725       1.6172      1.9266
orx2           10000     1.0150        0.0234     0.9991       1.0148      1.0307
x2_14          10000     4.9290        2.1547     3.0000       5.0000      6.0000
x2_50          10000     4.9673        2.3007     3.0000       5.0000      6.0000
x2_90          10000     4.9516        2.2265     3.0000       5.0000      6.0000
x1_167         10000     0.5606        0.4963          0       1.0000      1.0000
x1_304         10000     0.4469        0.4972          0            0      1.0000
x1_388         10000     0.4222        0.4939          0            0      1.0000

                               Posterior Intervals

          Parameter     Alpha      Equal-Tail Interval        HPD Interval

          beta0         0.050     -1.7734     -0.9641     -1.7537     -0.9612
          beta1         0.050      0.0245      0.9532     0.00910      0.9374
          beta2         0.050     -0.0309      0.0601     -0.0256      0.0628
          alpha10       0.050     -0.5174      0.0661     -0.5280      0.0517
          alpha11       0.050     -0.0289      0.0546     -0.0302      0.0529
          alpha20       0.050      1.5151      1.6127      1.5169      1.6137
          orx1          0.050      1.0248      2.5939      0.9783      2.4848
          orx2          0.050      0.9695      1.0619      0.9747      1.0648
          x2_14         0.050      1.0000      9.0000      1.0000      9.0000
          x2_50         0.050      1.0000     10.0000      1.0000      9.0000
          x2_90         0.050      1.0000     10.0000      1.0000      9.0000
          x1_167        0.050           0      1.0000           0      1.0000
          x1_304        0.050           0      1.0000           0      1.0000
          x1_388        0.050           0      1.0000           0      1.0000
```

Lastly, Figure 43 displays the effective sample sizes (ESS) of monitored variables. The ESSs indicate reasonable mixing for all of these variables.

Figure 43 Effective Sample Sizes



```
                          The MCMC Procedure

                        Effective Sample Sizes

                                     Autocorrelation
          Parameter          ESS                Time    Efficiency

          beta0            702.7             14.2318        0.0703
          beta1            789.5             12.6656        0.0790
          beta2            889.8             11.2383        0.0890
          alpha10          812.0             12.3158        0.0812
          alpha11          683.7             14.6256        0.0684
          alpha20          928.4             10.7708        0.0928
          orx1             806.2             12.4039        0.0806
          orx2             892.9             11.1997        0.0893
          x2_14           1565.7              6.3871        0.1566
          x2_50           1627.0              6.1461        0.1627
          x2_90           1676.8              5.9636        0.1677
          x1_167         10000.0              1.0000        1.0000
          x1_304          9766.1              1.0240        0.9766
          x1_388         10000.0              1.0000        1.0000
```

The odds ratio for X1 is the multiplicative change in the odds of a child wheezing in Steel City compared to the odds of the child wheezing in Green Hills. The estimated odds ratio (ORX1) value is 1.6736 with a corresponding 95% equal-tail credible interval of (1.0248, 2.5939). City of residence is a significant factor in a child's wheezing status. The estimated odds ratio for X2 is the multiplicative change in the odds of developing a wheeze for each additional reported cigarette smoked per day. The odds ratio of ORX2 indicates that the odds of a child developing a wheeze is 1.0150 times higher for each reported cigarette a mother smokes. The corresponding 95% equal-tail credible interval is (0.9695,1.0619). Since this interval contains the value 1, maternal smoking is not considered to be an influential effect.

See Chen (2009) and Chen (2011) for more information about the MCMC procedure.

Additional Postprocessing

The LIFEREG and PROBIT procedures have been updated to include additional postprocessing statements. They now provide the TEST, LSMEANS, LSMESTIMATE, ESTIMATE, SLICE, and EFFECTPLOT statements, and so does the LOGISTIC procedure for stratified analyses.

Statistical Graphics

Each release of SAS/STAT software includes additional graphs. As seen in the examples in this paper, new procedures come equipped with the appropriate graphs. The STDRATE procedure provides the strata SMR plot, the QUANTLIFE procedure produces quantile plots, and the QUANTSELECT procedure displays a graph of the progression of the average check loss. Existing procedures are also actively updating their existing graphs and adding useful new ones. For example, the FREQ procedure adds a mosaic plot in this release, and it also displays the common odds ratio in the odds ratios plot.

Other Highlights

A number of existing procedures have also had important updates; many of these are the result of user requests. A few of these enhancements are listed:

- WEIGHT statement in PROC LIFETEST
- case-level (observation-level) residual diagnostics with latent variables in PROC CALIS
- partial R-square for relative importance of parameters in PROC LOGISTIC
- Miettinen-Nurminen confidence limits for the difference of proportions in PROC FREQ
- Poisson sampling in PROC SURVEYSELECT

- group sequential design with nonbinding acceptance boundary in the SEQDESIGN and SEQTEST procedures
- post-stratification estimation in the SURVEYMEANS procedure
- REF= option added to the CLASS statement for GLM, MIXED, GLIMMIX, and ORTHOREG procedures

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 when it becomes available. 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.

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.

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.

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

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.

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.

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 Guixian Lin, Ying So, and Yonggong Yao for their contributions to the manuscript and to Allen Mcdowell and Robert Rodriguez for their reviews. They are also grateful to Tim Arnold for his documentation programming and to Anne Baxter for editing.

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