# SAS/STAT® 12.3 User's Guide
# The LIFEREG Procedure (Chapter)

# Chapter 51
# The LIFEREG Procedure

## Contents

# Overview: LIFEREG Procedure

The LIFEREG procedure fits parametric models to failure time data that can be uncensored, right censored, left censored, or interval censored. The models for the response variable consist of a linear effect composed of the covariates and a random disturbance term. The distribution of the random disturbance can be taken from a class of distributions that includes the extreme value, normal, logistic, and, by using a log transformation, the exponential, Weibull, lognormal, log-logistic, and three-parameter gamma distributions.

The model assumed for the response $\mathbf{y}$ is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}$$

where $\mathbf{y}$ is a vector of response values, often the log of the failure times, $\mathbf{X}$ is a matrix of covariates or independent variables (usually including an intercept term), $\boldsymbol{\beta}$ is a vector of unknown regression parameters, $\sigma$ is an unknown scale parameter, and $\boldsymbol{\epsilon}$ is a vector of errors assumed to come from a known distribution (such as the standard normal distribution). If an offset variable O is specified, the form of the model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{O} + \sigma\boldsymbol{\epsilon}$, where $\mathbf{O}$ is a vector of values of the offset variable O. The distribution might also depend on additional shape parameters. These models are equivalent to accelerated failure time models when the log of the response is the quantity being modeled. The effect of the covariates in an accelerated failure time model is to change the scale, and not the location, of a baseline distribution of failure times.

The LIFEREG procedure estimates the parameters by maximum likelihood with a Newton-Raphson algorithm. PROC LIFEREG estimates the standard errors of the parameter estimates from the inverse of the observed information matrix.

The accelerated failure time model assumes that the effect of independent variables on an event time distribution is multiplicative on the event time. Usually, the scale function is $\exp(\mathbf{x}_c'\boldsymbol{\beta}_c)$, where $\mathbf{x}_c$ is the vector of covariate values (not including the intercept term) and $\boldsymbol{\beta}_c$ is a vector of unknown parameters. Thus, if $T_0$ is an event time sampled from the baseline distribution corresponding to values of zero for the covariates,

then the accelerated failure time model specifies that, if the vector of covariates is $\mathbf{x}_c$, the event time is $T = \exp(\mathbf{x}_c'\boldsymbol{\beta}_c)T_0$. If $y = \log(T)$ and $y_0 = \log(T_0)$, then

$$y = \mathbf{x}_c'\boldsymbol{\beta}_c + y_0$$

This is a linear model with $y_0$ as the error term.

In terms of survival or exceedance probabilities, this model is

$$\Pr(T > t \mid \mathbf{x}_c) = \Pr(T_0 > \exp(-\mathbf{x}_c'\boldsymbol{\beta}_c)t)$$

The probability on the left-hand side of the equal sign is evaluated given the value $\mathbf{x}_c$ for the covariates, and the right-hand side is computed using the baseline probability distribution but at a scaled value of the argument. The right-hand side of the equation represents the value of the baseline survival function evaluated at $\exp(-\mathbf{x}_c'\boldsymbol{\beta}_c)t$.

Models usually have an intercept parameter and a scale parameter. In terms of the original untransformed event times, the effects of the intercept term and the scale term are to scale the event time and to raise the event time to a power, respectively. That is, if

$$\log(T_0) = \mu + \sigma \log(T_\epsilon)$$

then

$$T_0 = \exp(\mu)T_\epsilon^\sigma$$

Although it is possible to fit these models to the original response variable by using the NOLOG option, it is more common to model the log of the response variable. Because of this log transformation, zero values for the observed failure times are not allowed unless the NOLOG option is specified. Similarly, small values for the observed failure times lead to large negative values for the transformed response. The NOLOG option should be used only if you want to fit a distribution appropriate for the untransformed response, such as the extreme value instead of the Weibull. If you specify the normal or logistic distributions, the responses are not log transformed; that is, the NOLOG option is implicitly assumed.

Parameter estimates for the normal distribution are sensitive to large negative values, and care must be taken that the fitted model is not unduly influenced by them. Large negative values for the normal distribution can occur when fitting the lognormal distribution by log transforming the response, and some response values are near zero. Likewise, values that are extremely large after the log transformation have a strong influence in fitting the Weibull distribution (that is, the extreme value distribution for log responses). You should examine the residuals and check the effects of removing observations with large residuals or extreme values of covariates on the model parameters. The logistic distribution gives robust parameter estimates in the sense that the estimates have a bounded influence function.

The standard errors of the parameter estimates are computed from large sample normal approximations by using the observed information matrix. In small samples, these approximations might be poor. See Lawless (2003) for additional discussion and references. You can sometimes construct better confidence intervals by transforming the parameters. For example, large sample theory is often more accurate for $\log(\sigma)$ than $\sigma$. Therefore, it might be more accurate to construct confidence intervals for $\log(\sigma)$ and transform these into confidence intervals for $\sigma$. The parameter estimates and their estimated covariance matrix are available in an output SAS data set and can be used to construct additional tests or confidence intervals for the parameters. Alternatively, tests of parameters can be based on log-likelihood ratios. See Cox and Oakes (1984) for a

discussion of the merits of some possible test methods including score, Wald, and likelihood ratio tests. Likelihood ratio tests are generally more reliable for small samples than tests based on the information matrix.

The log-likelihood function is computed using the log of the failure time as a response. This log likelihood differs from the log likelihood obtained using the failure time as the response by an additive term of $\sum \log(t_i)$, where the sum is over the uncensored failure times. This term does not depend on the unknown parameters and does not affect parameter or standard error estimates. However, many published values of log likelihoods use the failure time as the basic response variable and, hence, differ by the additive term from the value computed by the LIFEREG procedure.

The classic Tobit model also fits into this class of models but with data usually censored on the left. The data considered by Tobin (1958) in his original paper came from a survey of consumers where the response variable is the ratio of expenditures on durable goods to the total disposable income. The two explanatory variables are the age of the head of household and the ratio of liquid assets to total disposable income. Because many observations in this data set have a value of zero for the response variable, the model fit by Tobin is

$$y = \max(\mathbf{x}'\boldsymbol{\beta} + \epsilon, 0)$$

which is a regression model with left censoring, where $\mathbf{x}' = (1, \mathbf{x}'_c)$.

Bayesian analysis of parametric survival models can be requested by using the BAYES statement in the LIFEREG procedure. In Bayesian analysis, the model parameters are treated as random variables, and inference about parameters is based on the posterior distribution of the parameters, given the data. The posterior distribution is obtained using Bayes' theorem as the likelihood function of the data weighted with a prior distribution. The prior distribution enables you to incorporate knowledge or experience of the likely range of values of the parameters of interest into the analysis. If you have no prior knowledge of the parameter values, you can use a noninformative prior distribution, and the results of the Bayesian analysis will be very similar to a classical analysis based on maximum likelihood. A closed form of the posterior distribution is often not feasible, and a Markov chain Monte Carlo method by Gibbs sampling is used to simulate samples from the posterior distribution. See Chapter 7, "Introduction to Bayesian Analysis Procedures," for an introduction to the basic concepts of Bayesian statistics. Also see the section "Bayesian Analysis: Advantages and Disadvantages" on page 132 for a discussion of the advantages and disadvantages of Bayesian analysis. See Ibrahim, Chen, and Sinha (2001) and Gilks, Richardson, and Spiegelhalter (1996) for more information about Bayesian analysis, including guidance in choosing prior distributions.

For Bayesian analysis, PROC LIFEREG generates a Gibbs chain for the posterior distribution of the model parameters. Summary statistics (mean, standard deviation, quartiles, HPD and credible intervals, correlation matrix) and convergence diagnostics (autocorrelations; Gelman-Rubin, Geweke, Raftery-Lewis, and Heidelberger and Welch tests; and the effective sample size) are computed for each parameter, as well as the correlation matrix of the posterior sample. Trace plots, posterior density plots, and autocorrelation function plots that are created using ODS Graphics are also provided for each parameter.

The LIFEREG procedure uses ODS Graphics to create graphs as part of its output. For general information about ODS Graphics, see Chapter 21, "Statistical Graphics Using ODS."

# Getting Started: LIFEREG Procedure

The following examples demonstrate how you can use the LIFEREG procedure to fit a parametric model to failure time data.

Suppose you have a response variable y that represents failure time; a binary variable, censor, with censor=0 indicating censored values; and two linearly independent variables, x1 and x2. The following statements perform a typical accelerated failure time model analysis. Higher-order effects such as interactions and nested effects are allowed in the independent variables list, but they are not shown in this example.

```
proc lifereg;
   model y*censor(0) = x1 x2;
run;
```

PROC LIFEREG can fit models to interval-censored data. The syntax for specifying interval-censored data is as follows:

```
proc lifereg;
   model (begin, end) = x1 x2;
run;
```

You can also model binomial data by using the *events/trials* syntax for the response, as illustrated in the following statements:

```
proc lifereg;
   model r/n=x1 x2;
run;
```

The variable n represents the number of trials, and the variable r represents the number of events.

## Modeling Right-Censored Failure Time Data

The following example demonstrates how you can use the LIFEREG procedure to fit a model to right-censored failure time data.

Suppose you conduct a study of two headache pain relievers. You divide patients into two groups, with each group receiving a different type of pain reliever. You record the time taken (in minutes) for each patient to report headache relief. Because some of the patients never report relief for the entire study, some of the observations are censored.

The following DATA step creates the SAS data set headache:

```
data Headache;
   input Minutes Group Censor @@;
   datalines;
11  1  0    12  1  0    19  1  0    19  1  0
19  1  0    19  1  0    21  1  0    20  1  0
21  1  0    21  1  0    20  1  0    21  1  0
20  1  0    21  1  0    25  1  0    27  1  0
30  1  0    21  1  1    24  1  1    14  2  0
16  2  0    16  2  0    21  2  0    21  2  0
23  2  0    23  2  0    23  2  0    23  2  0
```

```
25   2   1     23   2   0     24   2   0     24   2   0
26   2   1     32   2   1     30   2   1     30   2   0
32   2   1     20   2   1
;
```

The data set Headache contains the variable Minutes, which represents the reported time to headache relief; the variable Group, the group to which the patient is assigned; and the variable Censor, a binary variable indicating whether the observation is censored. Valid values of the variable Censor are 0 (no) and 1 (yes). Figure 51.1 shows the first five records of the data set Headache.

**Figure 51.1** Headache Data

| Obs | Minutes | Group | Censor |
|-----|---------|-------|--------|
| 1 | 11 | 1 | 0 |
| 2 | 12 | 1 | 0 |
| 3 | 19 | 1 | 0 |
| 4 | 19 | 1 | 0 |
| 5 | 19 | 1 | 0 |

The following statements invoke the LIFEREG procedure:

```
proc lifereg data=Headache;
   class Group;
   model Minutes*Censor(1)=Group;
   output out=New cdf=Prob;
run;
```

The CLASS statement specifies the variable Group as the classification variable. The MODEL statement syntax indicates that the response variable Minutes is right censored when the variable Censor takes the value 1. The MODEL statement specifies the variable Group as the single explanatory variable. Because the MODEL statement does not specify the DISTRIBUTION= option, the LIFEREG procedure fits the default type 1 extreme-value distribution by using $\log$(Minutes) as the response. This is equivalent to fitting the Weibull distribution.

The OUTPUT statement creates the output data set New. In addition to containing the variables in the original data set Headache, the SAS data set New also contains the variable Prob. This new variable is created by the CDF= option to contain the estimates of the cumulative distribution function evaluated at the observed response.

The results of this analysis are displayed in the following figures.

**Figure 51.2** Model Fitting Information from the LIFEREG Procedure

```
                      The LIFEREG Procedure

                        Model Information

           Data Set                   WORK.HEADACHE
           Dependent Variable          Log(Minutes)
           Censoring Variable                Censor
           Censoring Value(s)                     1
           Number of Observations               38
           Noncensored Values                   30
           Right Censored Values                 8
           Left Censored Values                  0
           Interval Censored Values              0
           Number of Parameters                  3
           Name of Distribution            Weibull
           Log Likelihood             -9.37930239


                   Class Level Information

              Name        Levels    Values

              Group          2       1 2
```

Figure 51.2 displays the class level information and model fitting information. There are 30 uncensored observations and 8 right-censored observations. The log likelihood for the Weibull distribution is –9.3793. The log-likelihood value can be used to compare the goodness of fit for nested models with different covariates, but with the same distribution.

**Figure 51.3** Model Fit Statistics from the LIFEREG Procedure

```
                        Fit Statistics

           -2 Log Likelihood                   18.759
           AIC (smaller is better)             24.759
           AICC (smaller is better)            25.464
           BIC (smaller is better)             29.671


                Fit Statistics (Unlogged Response)

           -2 Log Likelihood                  199.747
           Weibull AIC (smaller is better)    205.747
           Weibull AICC (smaller is better)   206.453
           Weibull BIC (smaller is better)    210.660
```

Figure 51.3 displays fit statistics for the model. The "Fit Statistics" table displays statistics based on the maximum extreme-value log likelihood fit by using $\log(\mathsf{Minutes})$ as the response. These statistics are useful

in comparing the fit of a different model when the fit criteria from the model that you compare is also based on the log likelihood using log(Minutes) as the response. The "Fit Statistics (Unlogged Response)" table is based on the maximum Weibull log likelihood using Minutes as the response. The AIC, BIC, and AICC statistics in this table can be used to compare models with different covariates, in addition to models with different distributions, as long as the fit statistics for the models that you compare use Minutes as the response.

**Figure 51.4** Model Parameter Estimates from the LIFEREG Procedure

```
            Analysis of Maximum Likelihood Parameter Estimates

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

    Intercept           1   3.3091   0.0589   3.1938   3.4245 3161.70   <.0001
    Group           1   1  -0.1933   0.0786  -0.3473  -0.0393   6.05    0.0139
    Group           2   0   0.0000     .        .        .        .       .
    Scale               1   0.2122   0.0304   0.1603   0.2809
    Weibull Shape       1   4.7128   0.6742   3.5604   6.2381
```

The table of parameter estimates is displayed in Figure 51.4. Both the intercept and the slope parameter for the variable group are significantly different from 0 at the 0.05 level. Because the variable group has only one degree of freedom, parameter estimates are given for only one level of the variable group (group=1). However, the estimate for the intercept parameter provides a baseline for group=2.

The resulting model is as follows:

$$
\log(\text{minutes}) = \begin{cases} 3.30911843 - 0.1933025 & \text{for } \text{group} = 1 \\ 3.30911843 & \text{for } \text{group} = 2 \end{cases}
$$

Note that the Weibull shape parameter for this model is the reciprocal of the extreme-value scale parameter estimate shown in Figure 51.4 ($1/0.21219 = 4.7128$).

The following statements produce a graph of the cumulative distribution values versus the variable Minutes.

```
proc sgplot data=New;
   scatter x=Minutes y=Prob / group=Group;
   discretelegend;
run;
```

Figure 51.5 displays the estimated cumulative distribution function values contained in the output data set New for each group.

**Figure 51.5** Plot of the Estimated Cumulative Distribution Function
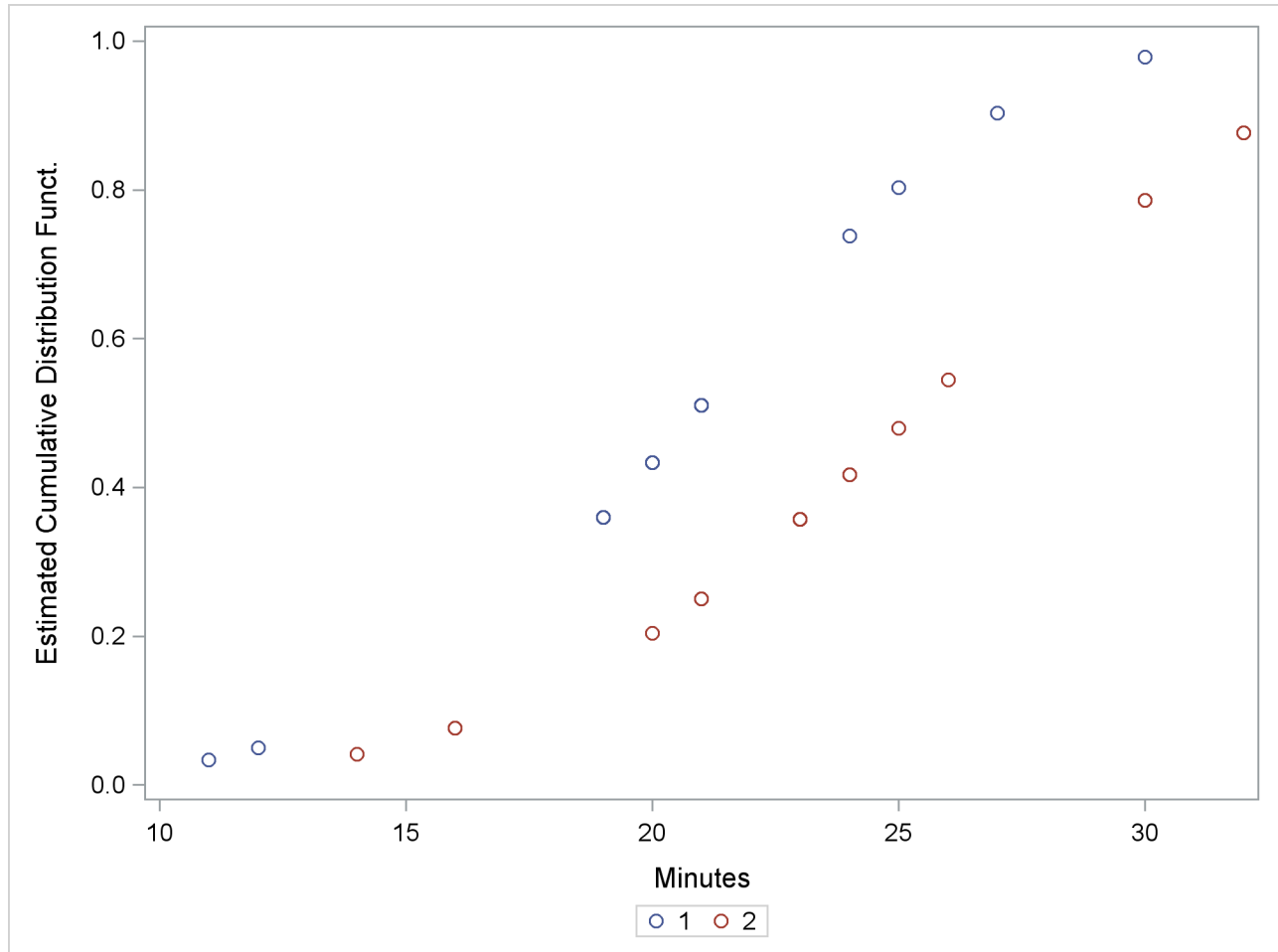


## Bayesian Analysis of Right-Censored Data

Nelson (1982) describes a study of the lifetimes of locomotive engine fans. This example shows how to use PROC LIFEREG to carry out a Bayesian analysis of the engine fan data. In this example, a lognormal distribution is used to model the engine lifetimes, but other survival time distributions, such as the Weibull, can also be used.

The following SAS statements create the SAS data set Fan. This data set contains a censoring indicator variable and right-censored survival times for the 70 locomotive engine fans in the study.

```
data Fan;
   input Lifetime Censor@@;
   datalines;
 450 0    460 1   1150 0   1150 0   1560 1
1600 0   1660 1   1850 1   1850 1   1850 1
1850 1   1850 1   2030 1   2030 1   2030 1
2070 0   2070 0   2080 0   2200 1   3000 1
3000 1   3000 1   3000 1   3100 0   3200 1
3450 0   3750 1   3750 1   4150 1   4150 1
4150 1   4150 1   4300 1   4300 1   4300 1
4300 1   4600 0   4850 1   4850 1   4850 1
4850 1   5000 1   5000 1   5000 1   6100 1
6100 0   6100 1   6100 1   6300 1   6450 1
6450 1   6700 1   7450 1   7800 1   7800 1
8100 1   8100 1   8200 1   8500 1   8500 1
8500 1   8750 1   8750 0   8750 1   9400 1
9900 1  10100 1  10100 1  10100 1  11500 1
;
```

Some of the fans had not failed at the time the data were collected, and the unfailed units have right-censored lifetimes. The variable Lifetime represents either a failure time or a censoring time. The variable Censor is equal to 0 if the value of Lifetime is a failure time, and it is equal to 1 if the value is a censoring time.

The following SAS statements specify a Bayesian analysis that uses a lognormal model for the engine lifetimes. There are no covariates, so the model is an intercept-only model. The OUTPOST= option saves the samples from the posterior distribution in the SAS data set Post for further processing.

```
ods graphics on;
proc lifereg data=Fan;
   model Lifetime*Censor( 1 )= / dist=lognormal;
   bayes seed=1  outpost=Post;
run;
ods graphics off;
```

The SEED= option is specified to maintain reproducibility; no other options are specified in the BAYES statement. By default, a uniform prior distribution is assumed for the intercept coefficient. The uniform prior is a flat prior on the real line with a distribution that reflects ignorance of the location of the parameter, placing equal probability on all possible values the regression coefficient can take. Using the uniform prior in the following example, you would expect the Bayesian estimates to resemble the classical results of maximizing the likelihood. If you can elicit an informative prior on the regression coefficients, you should use the COEFFPRIOR= option to specify it. A default noninformative gamma prior is used for the lognormal scale parameter $\sigma$.

You should make sure that the posterior distribution samples have achieved convergence before using them for Bayesian inference. If you do not specify additional options, PROC LIFEREG produces by default three convergence diagnostics: autocorrelations of the posterior sample, effective sample size, and the Geweke statistic. See the section "Assessing Markov Chain Convergence" on page 139 for information about assessing the convergence of the chain of posterior samples. Trace plots, posterior density plots, and autocorrelation function plots that are created using ODS Graphics are also provided for each parameter. See the section "Visual Analysis via Trace Plots" on page 139 for help in interpreting these plots.

The "Analysis of Maximum Likelihood Parameter Estimates" table in Figure 51.6 summarizes maximum likelihood estimates of the lognormal intercept and scale parameters.

**Figure 51.6** Maximum Likelihood Estimates from the LIFEREG Procedure

```
                        The LIFEREG Procedure

                          Bayesian Analysis

             Analysis of Maximum Likelihood Parameter Estimates

                                       Standard    95% Confidence
                Parameter     DF Estimate   Error        Limits

                Intercept     1  10.1432   0.5211    9.1219   11.1646
                Scale         1   1.6796   0.3893    1.0664    2.6453
```

Since no prior distribution for the intercept was specified, the default uniform improper distribution shown in the "Uniform Prior for Regression Coefficients" table in Figure 51.7 is used.

Noninformative prior distributions are appropriate if you have no prior knowledge of the likely range of values of the parameters, and if you want to make probability statements about the parameters or functions of the parameters. Refer, for example, to Ibrahim, Chen, and Sinha (2001) for more information about choosing prior distributions.

The default noninformative gamma prior distribution for the lognormal scale parameter is shown in the "Independent Prior Distributions for Model Parameters" table in Figure 51.7.

**Figure 51.7** Noninformative Prior Distributions

```
                        The LIFEREG Procedure

                          Bayesian Analysis

                Uniform Prior for Regression Coefficients

                      Parameter     Prior

                      Intercept     Constant


            Independent Prior Distributions for Model Parameters

                   Prior
      Parameter    Distribution              Hyperparameters

      Scale        Gamma         Shape       0.001    Inverse Scale     0.001
```

By default, posterior mode estimates of the model parameters are used as the starting value for the simulation. These are listed in the "Initial Values of the Chain" table in Figure 51.8.

**Figure 51.8** Markov Chain Initial Values

```
                      Initial Values of the Chain

         Chain             Seed      Intercept        Scale

           1                 1        10.0501       1.59544
```

Summary statistics for the posterior sample are displayed in the "Fit Statistics," "Descriptive Statistics for the Posterior Sample," "Interval Statistics for the Posterior Sample," and "Posterior Correlation Matrix" tables in Figure 51.9. Since noninformative prior distributions were used, these results are consistent with the maximum likelihood estimates shown in Figure 51.6.

**Figure 51.9** Posterior Sample Summary Statistics

```
                            Fit Statistics

         DIC (smaller is better)                    87.245
         pD (effective number of parameters)         1.823


                         The LIFEREG Procedure

                          Bayesian Analysis

                          Posterior Summaries

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

   Intercept     10000   10.4196      0.6172    9.9670    10.3259    10.7959
   Scale         10000    1.9196      0.4809    1.5675     1.8476     2.1931


                          Posterior Intervals

       Parameter    Alpha     Equal-Tail Interval        HPD Interval

       Intercept    0.050       9.4477    11.8994      9.3216    11.6752
       Scale        0.050       1.1906     3.0570      1.1104     2.8834


                       Posterior Correlation Matrix

              Parameter          Intercept        Scale

              Intercept           1.0000         0.8297
              Scale               0.8297         1.0000
```

By default, PROC LIFEREG computes three convergence diagnostics: the lag1, lag5, lag10, and lag50 autocorrelations; the Geweke diagnostic; and the effective sample size. These are displayed in Figure 51.10. There is no indication that the Markov chain has not converged. See the section "Assessing Markov Chain Convergence" on page 139 for more information about convergence diagnostics and their interpretation.

**Figure 51.10** Posterior Sample Summary Statistics

```
                    The LIFEREG Procedure

                      Bayesian Analysis

                   Posterior Autocorrelations

       Parameter      Lag 1      Lag 5      Lag 10      Lag 50

       Intercept     0.6973     0.1765     0.0190     -0.0017
       Scale         0.6955     0.1713     0.0172     -0.0002


                       Geweke Diagnostics

               Parameter            z     Pr > |z|

               Intercept       -0.9183      0.3585
               Scale           -0.9233      0.3559


                     Effective Sample Sizes

                              Autocorrelation
       Parameter       ESS            Time      Efficiency

       Intercept      1772.8         5.6408       0.1773
       Scale          1805.0         5.5400       0.1805
```

Summary statistics of the posterior distribution samples are produced by default. However, these statistics might not be sufficient for carrying out your Bayesian inference. The samples from the posterior distribution saved in the SAS data set Post created with the OUTPOST= option can be used for further analysis.

Trace, autocorrelation, and density plots for the three model parameters shown in Figure 51.11 and Figure 51.12 are useful in diagnosing whether the Markov chain of posterior samples has converged. These plots show no evidence that the chain has not converged. See the section "Visual Analysis via Trace Plots" on page 139 for more information about interpreting these types of diagnostic plots.

**Figure 51.11** Diagnostic Plots
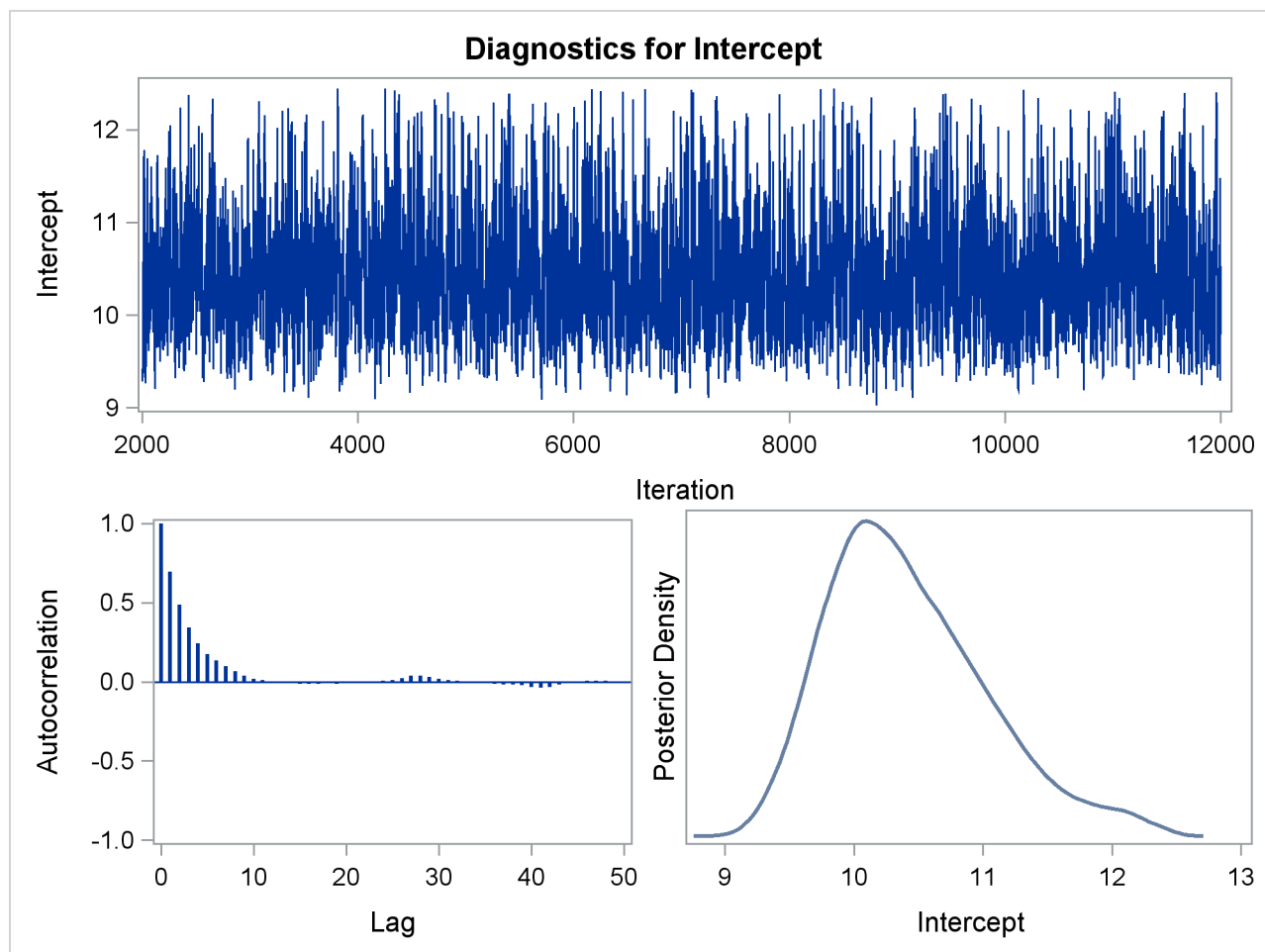
**Figure 51.12** Diagnostic Plots
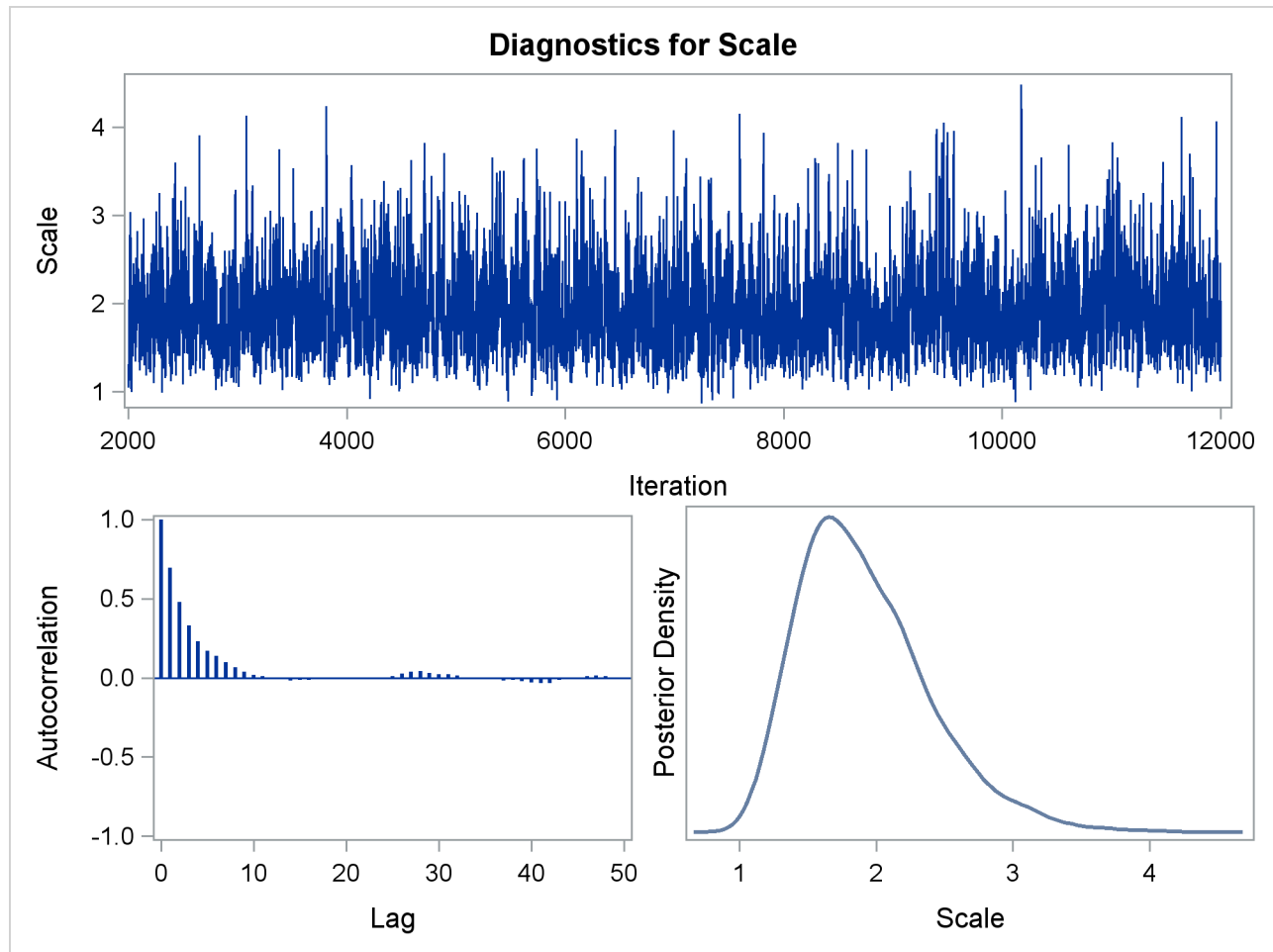


The fraction failing in the first 8000 hours of operation might be a quantity of interest. This kind of information could be useful, for example, in determining whether to improve the reliability of the engine components due to warranty considerations. The following SAS statements compute the mean and percentiles of the distribution of the fraction failing in the first 8000 hours from the posterior sample data set Post:

```
data Prob;
   set Post;
   Frac = ProbNorm(( log(8000) - Intercept ) / Scale );
   label Frac= 'Fraction Failing in 8000 Hours';
run;

proc means data = Prob(keep=Frac) n mean p10 p25 p50 p75 p90;
run;
```

The mean fraction of failures in the first 8000 hours, shown in Figure 51.13, is about 0.24, which could be used in further analysis of warranty costs. The 10th percentile is about 0.16 and the 90th percentile is about 0.32, which gives an assessment of the probable range of the fraction failing in the first 8000 hours.

**Figure 51.13** Fraction Failing in 8000 Hours

```
                        The MEANS Procedure

            Analysis Variable : Frac Fraction Failing in 8000 Hours

      N           Mean      10th Pctl      25th Pctl      50th Pctl      75th Pctl
  -----------------------------------------------------------------------------------
  10000      0.2381467      0.1628591      0.1953691      0.2336756      0.2766051
  -----------------------------------------------------------------------------------


            Analysis Variable : Frac Fraction Failing in 8000 Hours

                                   90th Pctl
                                 ------------
                                   0.3190883
                                 ------------
```

# Syntax: LIFEREG Procedure

The following statements are available in the LIFEREG procedure:

**PROC LIFEREG** *< options >* ;
    **BAYES** *< options >* ;
    **BY** *variables* ;
    **CLASS** *variables* ;
    **ESTIMATE** *< 'label' > estimate-specification < (divisor=n) >*
                *< , ... < 'label' > estimate-specification < (divisor=n) > > </ options >* ;
    **EFFECTPLOT** *< plot-type < (plot-definition-options) > > </ options >* ;
    **INSET** *< keyword-list > </ options >* ;
    **LSMEANS** *< model-effects > </ options >* ;
    **LSMESTIMATE** *model-effect < 'label' > values < (divisor=n) >*
                *< , ... < 'label' > values < (divisor=n) > > </ options >* ;
    **MODEL** *response = < effects > </ options >* ;
    **OUTPUT** *< **OUT=**SAS-data-set > < keyword=name ... keyword=name > < options >* ;
    **PROBPLOT** *</ options >* ;
    **SLICE** *model-effect </ options >* ;
    **STORE** *< **OUT=** >item-store-name </ **LABEL=***'label'> ;
    **TEST** *< model-effects > </ options >* ;
    **WEIGHT** *variable* ;

The MODEL statement is required; it specifies both the variables that are used in the regression part of the model and the distribution that is used for the error (random) component of the model. Each invocation of the LIFEREG procedure can use only one MODEL statement. If multiple MODEL statements are present, only

the last is used. You can specify main effects and interaction terms in the MODEL statement, as in the GLM procedure. You can specify initial values in the MODEL statement or in an INEST= data set. If no initial values are specified, the starting estimates are obtained by ordinary least squares. The CLASS statement determines which explanatory variables are treated as categorical. The WEIGHT statement identifies a *variable* with values that are used to weight the observations. Observations with zero or negative weights are not used to fit the model, although predicted values can be computed for them. The OUTPUT statement creates an output data set that contains predicted values and residuals.

The ESTIMATE, EFFECTPLOT, LSMEANS, LSMESTIMATE, SLICE, STORE, and TEST statements are common to many procedures. Summary descriptions of functionality and syntax for these statements are also given after the PROC LIFEREG statement in alphabetical order, and full documentation about them is available in Chapter 19, "Shared Concepts and Topics."

## PROC LIFEREG Statement

**PROC LIFEREG** < *options* > ;

The PROC LIFEREG statement invokes the LIFEREG procedure. Table 51.1 summarizes the options available in the PROC LIFEREG statement.

**Table 51.1** PROC LIFEREG Statement Options

| Option | Description |
|---|---|
| COVOUT | Writes the estimated covariance matrix to the OUTEST= data set |
| DATA= | Specifies the input SAS data set |
| GOUT= | Specifies a graphics catalog |
| INEST= | Specifies an input SAS data set that contains initial estimates |
| NAMELEN= | Specifies the length of effect names |
| NOPRINT | Suppresses the display of the output |
| ORDER= | Specifies the sort order for the levels of the classification variables |
| OUTEST= | Specifies an output SAS data set |
| PLOTS= | Controls graphics created by ODS Graphics |
| XDATA= | Specifies a SAS input data containing values for the independent variables |

You can specify the following *options* in the PROC LIFEREG statement.

**COVOUT**
> writes the estimated covariance matrix to the OUTEST= data set if convergence is attained.

**DATA=**SAS-data-set
> specifies the input SAS data set used by PROC LIFEREG. By default, the most recently created SAS data set is used.

**GOUT=**graphics-catalog
> specifies a graphics catalog in which to save graphics output.

**INEST=***SAS-data-set*

> specifies an input SAS data set that contains initial estimates for all the parameters in the model. See the section "INEST= Data Set" on page 3945 for a detailed description of the contents of the INEST= data set.

**NAMELEN=***n*

> specifies the length of effect names in tables and output data sets to be *n* characters, where *n* is a value between 20 and 200. The default length is 20 characters.

**NOPRINT**

> suppresses the display of the output. Note that this option temporarily disables the Output Delivery System (ODS). For more information, see Chapter 20, "Using the Output Delivery System."

**ORDER=DATA | FORMATTED | FREQ | INTERNAL**

> specifies the sort order for the levels of the classification variables (which are specified in the CLASS statement). This option applies to the levels for all classification variables, except when you use the (default) ORDER=FORMATTED option with numeric classification variables that have no explicit format. With this option, the levels of such variables are ordered by their internal value.
>
> The ORDER= option can take the following values:

| Value of ORDER= | Levels Sorted By |
|---|---|
| DATA | Order of appearance in the input data set |
| FORMATTED | External formatted value, except for numeric variables with no explicit format, which are sorted by their unformatted (internal) value |
| FREQ | Descending frequency count; levels with the most observations come first in the order |
| INTERNAL | Unformatted value |

> By default, ORDER=FORMATTED. For ORDER=FORMATTED and ORDER=INTERNAL, the sort order is machine-dependent. For more information about sort order, see the chapter on the SORT procedure in the *Base SAS Procedures Guide* and the discussion of BY-group processing in *SAS Language Reference: Concepts*.

**OUTEST=***SAS-data-set*

> specifies an output SAS data set containing the parameter estimates, the maximized log likelihood, and, if the COVOUT option is specified, the estimated covariance matrix. See the section "OUTEST= Data Set" on page 3946 for a detailed description of the contents of the OUTEST= data set.

**PLOTS=NONE | PROBPLOT**

> specifies options that control graphics created by ODS Graphics.

ODS Graphics must be enabled before plots can be requested. For example:

```
ods graphics on;

proc lifereg plots=probplot;
   model y = x;
run;

ods graphics off;
```

For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 600 in Chapter 21, "Statistical Graphics Using ODS."

The following plot requests are available.

NONE    suppresses any plots created by ODS Graphics specified in other LIFEREG statements, such as the BAYES or PROBPLOT statement.

PROBPLOT    creates a default probability plot based on information in the MODEL statement. If a PROBPLOT option is also specified, the probability plot specified in the PROBPLOT statement is created, and this option is ignored.

**XDATA=***SAS-data-set*

specifies an input SAS data set that contains values for all the independent variables in the MODEL statement and variables in the CLASS statement for probability plotting. If there are covariates specified in a MODEL statement and a probability plot is requested with a PROBPLOT statement, you specify fixed values for the effects in the MODEL statement with the XDATA= data set. See the section "XDATA= Data Set" on page 3947 for a detailed description of the contents of the XDATA= data set.

## BAYES Statement

> **BAYES** < *options* > ;

The BAYES statement requests a Bayesian analysis of the regression model by using Gibbs sampling. The Bayesian posterior samples (also known as the chain) for the regression parameters are not tabulated. The Bayesian posterior samples (also known as the chain) for the model parameters can be output to a SAS data set.

Table 51.2 summarizes the options available in the BAYES statement.

**Table 51.2**   BAYES Statement Options

| Option | Description |
|---|---|
| **Monte Carlo Options** | |
| INITIAL= | Specifies initial values of the chain |
| INITIALMLE | Specifies that maximum likelihood estimates be used as initial values of the chain |

**Table 51.2** (*continued*)

| Option | Description |
|---|---|
| METROPOLIS= | Specifies the use of a Metropolis step |
| NBI= | Specifies the number of burn-in iterations |
| NMC= | Specifies the number of iterations after burn-in |
| SEED= | Specifies the random number generator seed |
| THINNING= | Controls the thinning of the Markov chain |
| **Model and Prior Options** | |
| COEFFPRIOR= | Specifies the prior of the regression coefficients |
| EXPONENTIALSCALEPRIOR= | Specifies the prior of the exponential scale parameter |
| GAMMASHAPEPRIOR= | Specifies the prior of the three-parameter gamma shape parameter |
| SCALEPRIOR= | Specifies the prior of the scale parameter |
| WEIBULLSCALEPRIOR= | Specifies the prior of the Weibull scale parameter |
| WEIBULLSHAPEPRIOR= | Specifies the prior of the Weibull shape parameter |
| **Summary Statistics and Convergence Diagnostics** | |
| DIAGNOSTICS= | Displays convergence diagnostics |
| PLOTS= | Displays diagnostic plots |
| STATISTICS= | Displays summary statistics of the posterior samples |
| **Posterior Samples** | |
| OUTPOST= | Names a SAS data set for the posterior samples |

The following list describes these options and their suboptions.

**COEFFPRIOR=UNIFORM | NORMAL <(***normal-options***)>**

**CPRIOR=UNIFORM | NORMAL <(***option***)>**

**COEFF=UNIFORM | NORMAL <(***option***)>**

> specifies the prior distribution for the regression coefficients. The default is COEFFPRIOR=UNIFORM. The available prior distributions are as follows:

> **NORMAL<(***normal-option***)>**
>
> > specifies a normal distribution. The *normal-options* include the following:
> >
> > **CONDITIONAL**
> >
> > > specifies that the normal prior, conditional on the current Markov chain value of the location-scale model precision parameter $\tau = \frac{1}{\sigma^2}$, is $N(\boldsymbol{\mu}, \tau^{-1}\boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and covariance of the normal prior specified by other normal options.
> >
> > **INPUT=** *SAS-data-set*
> >
> > > specifies a SAS data set that contains the mean and covariance information of the normal prior. The data set must have a _TYPE_ variable to represent the type of each observation and a variable for each regression coefficient. If the data set also contains a _NAME_ variable, the values of this variable are used to identify the covariances for the _TYPE_='COV' observations; otherwise, the _TYPE_='COV' observations are assumed to be in the same order as the explanatory variables in the MODEL statement. PROC LIFEREG reads the

mean vector from the observation with _TYPE_='MEAN' and reads the covariance matrix from observations with _TYPE_='COV'. For an independent normal prior, the variances can be specified with _TYPE_='VAR'; alternatively, the precisions (inverse of the variances) can be specified with _TYPE_='PRECISION'.

**RELVAR< =c >**

specifies the normal prior $N(\mathbf{0}, c\mathbf{J})$, where $\mathbf{J}$ is a diagonal matrix with diagonal elements equal to the variances of the corresponding ML estimator. By default, $c = 10^6$.

**VAR< =c >**

specifies the normal prior $N(\mathbf{0}, c\mathbf{I})$, where $\mathbf{I}$ is the identity matrix.

If you do not specify an option, the normal prior $N(\mathbf{0}, 10^6\mathbf{I})$, where $\mathbf{I}$ is the identity matrix, is used. See the section "Normal Prior" on page 3949 for more details.

**UNIFORM**

specifies a flat prior—that is, the prior that is proportional to a constant ($p(\beta_1, \ldots, \beta_k) \propto 1$ for all $-\infty < \beta_i < \infty$).

**DIAGNOSTICS=ALL | NONE |** *(keyword-list)*

**DIAG=ALL | NONE |** *(keyword-list)*

controls the number of diagnostics produced. You can request all the following diagnostics by specifying DIAGNOSTICS=ALL. If you do not want any of these diagnostics, specify DIAGNOSTICS=NONE. If you want some but not all of the diagnostics, or if you want to change certain settings of these diagnostics, specify a subset of the following keywords. The default is DIAGNOSTICS=(AUTOCORR ESS GEWEKE).

**AUTOCORR < (LAGS=** *numeric-list***) >**

computes the autocorrelations of lags given by LAGS= list for each parameter. Elements in the list are truncated to integers and repeated values are removed. If the LAGS= option is not specified, autocorrelations of lags 1, 5, 10, and 50 are computed for each variable. See the section "Autocorrelations" on page 152 for details.

**ESS**

computes Carlin's estimate of the effective sample size, the correlation time, and the efficiency of the chain for each parameter. See the section "Effective Sample Size" on page 152 for details.

**GELMAN <** *(gelman-options)* **>**

computes the Gelman and Rubin convergence diagnostics. You can specify one or more of the following *gelman-options*:

**NCHAIN=***number*

**N=***number*

specifies the number of parallel chains used to compute the diagnostic, and must be 2 or larger. The default is NCHAIN=3. If an INITIAL= data set is used, NCHAIN defaults to the number of rows in the INITIAL= data set. If any number other than this is specified with the NCHAIN= option, the NCHAIN= value is ignored.

**ALPHA=**_value_

specifies the significance level for the upper bound. The default is ALPHA=0.05, resulting in a 97.5% bound.

See the section "Gelman and Rubin Diagnostics" on page 145 for details.

**GEWEKE** < (_geweke-options_) >

computes the Geweke spectral density diagnostics, which are essentially a two-sample $t$ test between the first $f_1$ portion and the last $f_2$ portion of the chain. The default is $f_1 = 0.1$ and $f_2 = 0.5$, but you can choose other fractions by using the following _geweke-options_:

**FRAC1=**_value_

specifies the fraction $f_1$ for the first window.

**FRAC2=**_value_

specifies the fraction $f_2$ for the second window.

See the section "Geweke Diagnostics" on page 146 for details.

**HEIDELBERGER** < (_heidel-options_) >

computes the Heidelberger and Welch diagnostic for each variable, which consists of a stationarity test of the null hypothesis that the sample values form a stationary process. If the stationarity test is not rejected, a halfwidth test is then carried out. Optionally, you can specify one or more of the following _heidel-options_:

**SALPHA=**_value_

specifies the $\alpha$ level $(0 < \alpha < 1)$ for the stationarity test.

**HALPHA=**_value_

specifies the $\alpha$ level $(0 < \alpha < 1)$ for the halfwidth test.

**EPS=**_value_

specifies a positive number $\epsilon$ such that if the halfwidth is less than $\epsilon$ times the sample mean of the retained iterates, the halfwidth test is passed.

See the section "Heidelberger and Welch Diagnostics" on page 148 for details.

**MCSE**

**MCERROR**

computes the Monte Carlo standard error for each parameter. The Monte Caro standard error, which measures the simulation accuracy, is the standard error of the posterior mean estimate and is calculated as the posterior standard deviation divided by the square root of the effective sample size. See the section "Standard Error of the Mean Estimate" on page 153 for details.

**RAFTERY**< (_raftery-options_) >

computes the Raftery and Lewis diagnostics that evaluate the accuracy of the estimated quantile ($\hat{\theta}_Q$ for a given $Q \in (0, 1)$) of a chain. $\hat{\theta}_Q$ can achieve any degree of accuracy when the chain is allowed to run for a long time. A stopping criterion is when the estimated probability $\hat{P}_Q = \Pr(\theta \le \hat{\theta}_Q)$ reaches within $\pm R$ of the value $Q$ with probability $S$; that is, $\Pr(Q - R \le \hat{P}_Q \le Q + R) = S$. The following _raftery-options_ enable you to specify $Q, R, S$, and a precision level $\epsilon$ for the test:

**QUANTILE | Q=***value*

specifies the order (a value between 0 and 1) of the quantile of interest. The default is 0.025.

**ACCURACY | R=***value*

specifies a small positive number as the margin of error for measuring the accuracy of estimation of the quantile. The default is 0.005.

**PROBABILITY | S=***value*

specifies the probability of attaining the accuracy of the estimation of the quantile. The default is 0.95.

**EPSILON | EPS=***value*

specifies the tolerance level (a small positive number) for the stationary test. The default is 0.001.

See the section "Raftery and Lewis Diagnostics" on page 149 for details.

**EXPSCALEPRIOR=GAMMA<(***options***)>| IMPROPER**

**ESCALEPRIOR=GAMMA<(***options***)>| IMPROPER**

**ESCPRIOR=GAMMA<(***options***)>| IMPROPER**

specifies that Gibbs sampling be performed on the exponential distribution scale parameter and the prior distribution for the scale parameter. This prior distribution applies only when the exponential distribution and no covariates are specified.

A gamma prior $G(a, b)$ with density $f(t) = \frac{b(bt)^{a-1}e^{-bt}}{\Gamma(a)}$ is specified by EXPSCALEPRIOR=GAMMA, which can be followed by one of the following *gamma-options* enclosed in parentheses. The hyperparameters $a$ and $b$ are the shape and inverse-scale parameters of the gamma distribution, respectively. See the section "Gamma Prior" on page 3949 for more details. The default is $G(10^{-4}, 10^{-4})$.

**RELSHAPE<=***c***>**

specifies independent $G(c\hat{\alpha}, c)$ distribution, where $\hat{\alpha}$ is the MLE of the exponential scale parameter. With this choice of hyperparameters, the mean of the prior distribution is $\hat{\alpha}$ and the variance is $\frac{\hat{\alpha}}{c^2}$. By default, $c=10^{-4}$.

**SHAPE=***a*

**ISCALE=***b*

when both specified, results in a $G(a, b)$ prior.

**SHAPE=***c*

when specified alone, results in a $G(c, c)$ prior.

**ISCALE=***c*

when specified alone, results in a $G(c, c)$ prior.

An improper prior with density $f(t)$ proportional to $t^{-1}$ is specified with EXPSCALEPRIOR=IMPROPER.

**GAMMASHAPEPRIOR=NORMAL<(***options***)>**

**GAMASHAPEPRIOR=NORMAL<(***options***)>**

**SHAPE1PRIOR=NORMAL<(***options***)>**

specifies the prior distribution for the gamma distribution shape parameter. If you do not specify any options in a gamma model, the $N(0, 10^6)$ prior for the shape is used. You can specify MEAN= and VAR= or RELVAR= options, either alone or together, to specify the mean and variance of the normal prior for the gamma shape parameter.

**MEAN=***a*
> specifies a normal prior $N(a, 10^6)$. By default, *a*=0.

**RELVAR<=***b***>**
> specifies the normal prior $N(0, bJ)$, where $J$ is the variance of the MLE of the shape parameter. By default, *b*=$10^6$.

**VAR=***c*
> specifies the normal prior $N(0, c)$. By default, *c*=$10^6$.

**INITIAL=***SAS-data-set*
> specifies the SAS data set that contains the initial values of the Markov chains. The INITIAL= data set must contain all the variables of the model. You can specify multiple rows as the initial values of the parallel chains for the Gelman-Rubin statistics, but posterior summaries, diagnostics, and plots are computed only for the first chain. If the data set also contains the variable _SEED_, the value of the _SEED_ variable is used as the seed of the random number generator for the corresponding chain.

**INITIALMLE**
> specifies that maximum likelihood estimates of the model parameters be used as initial values of the Markov chain. If this option is not specified, estimates of the mode of the posterior distribution obtained by optimization are used as initial values.

**METROPOLIS=YES**
**METROPOLIS=NO**
> specifies the use of a Metropolis step to generate Gibbs samples for posterior distributions that are not log concave. The default value is METROPOLIS=YES.

**NBI=***number*
> specifies the number of burn-in iterations before the chains are saved. The default is 2000.

**NMC=***number*
> specifies the number of iterations after the burn-in. The default is 10000.

**OUTPOST=***SAS-data-set*
**OUT=***SAS-data-set*
> names the SAS data set that contains the posterior samples. See the section "OUTPOST= Output Data Set" on page 3951 for more information. Alternatively, you can create the output data set by specifying an ODS OUTPUT statement as follows:

> **ODS OUTPUT POSTERIORSAMPLE=***SAS-data-set*

**PLOTS<(***global-plot-options***)>=** *plot-request*
**PLOTS<(***global-plot-options***)>=** *(plot-request< ...plot-request>)*
> controls the display of diagnostic plots. Three types of plots can be requested: trace plots, autocorrelation function plots, and kernel density plots. By default, the plots are displayed in panels unless the global plot option UNPACK is specified. Also, when specifying more than one type of plots, the plots are displayed by parameters unless the global plot option GROUPBY is specified. When you specify only one plot request, you can omit the parentheses around the plot request. For example:

```
plots=none
plots(unpack)=trace
plots=(trace autocorr)
```

ODS Graphics must be enabled before plots can be requested. For example:

```
ods graphics on;
proc lifereg;
   model y=x;
   bayes plots=trace;
run;
ods graphics off;
```

For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 600 in Chapter 21, "Statistical Graphics Using ODS."

The global plot options are as follows:

**FRINGE**
>   creates a fringe plot on the X axis of the density plot.

**GROUPBY=PARAMETER**
**GROUPBY=TYPE**
>   specifies how the plots are grouped when there is more than one type of plot.

>   **GROUPBY=TYPE**
>   >   specifies that the plots be grouped by type.

>   **GROUPBY=PARAMETER**
>   >   specifies that the plots be grouped by parameter.

>   GROUPBY=PARAMETER is the default.

**LAGS=***n*
>   specifies that autocorrelations be plotted up to lag *n*. If this option is not specified, autocorrelations are plotted up to lag 50.

**SMOOTH**
>   displays a fitted penalized B-spline curve for each trace plot.

**UNPACKPANEL**
**UNPACK**
>   specifies that all paneled plots be unpacked, meaning that each plot in a panel is displayed separately.

The plot requests include the following:

**ALL**
> specifies all types of plots. PLOTS=ALL is equivalent to specifying PLOTS=(TRACE AUTO-CORR DENSITY).

**AUTOCORR**
> displays the autocorrelation function plots for the parameters.

**DENSITY**
> displays the kernel density plots for the parameters.

**NONE**
> suppresses all diagnostic plots.

**TRACE**
> displays the trace plots for the parameters. See the section "Visual Analysis via Trace Plots" on page 139 for details.

**SCALEPRIOR=GAMMA< (*options*) >**
specifies that Gibbs sampling be performed on the location-scale model scale parameter and the prior distribution for the scale parameter.

A gamma prior $G(a, b)$ with density $f(t) = \frac{b(bt)^{a-1}e^{-bt}}{\Gamma(a)}$ is specified by SCALEPRIOR=GAMMA, which can be followed by one of the following *gamma-options* enclosed in parentheses. The hyper-parameters $a$ and $b$ are the shape and inverse-scale parameters of the gamma distribution, respectively. See the section "Gamma Prior" on page 3949 for details. The default is $G(10^{-4}, 10^{-4})$.

**RELSHAPE< =*c* >**
> specifies independent $G(c\hat{\sigma}, c)$ distribution, where $\hat{\sigma}$ is the MLE of the scale parameter. With this choice of hyperparameters, the mean of the prior distribution is $\hat{\sigma}$ and the variance is $\frac{\hat{\sigma}}{c}$. By default, $c=10^{-4}$.

**SHAPE=*a***
**ISCALE=*b***
> when both specified, results in a $G(a, b)$ prior.

**SHAPE=*c***
> when specified alone, results in a $G(c, c)$ prior.

**ISCALE=*c***
> when specified alone, results in a $G(c, c)$ prior.

**SEED=*number***
specifies an integer seed in the range 1 to $2^{31} - 1$ for the random number generator in the simulation. Specifying a seed enables you to reproduce identical Markov chains for the same specification. If the SEED= option is not specified, or if you specify a nonpositive seed, a random seed is derived from the time of day.

**STATISTICS** < (*global-options*) > = **ALL** | **NONE** | *keyword* | *(keyword-list)*

**STATS** < (*global-statoptions*) > = **ALL** | **NONE** | *keyword* | *(keyword-list)*

controls the number of posterior statistics produced. Specifying STATISTICS=ALL is equivalent to specifying STATISTICS= (SUMMARY INTERVAL COV CORR). If you do not want any posterior statistics, you specify STATISTICS=NONE. The default is STATISTICS=(SUMMARY INTERVAL). See the section "Summary Statistics" on page 153 for details. The *global-options* include the following:

**ALPHA=***numeric-list*

controls the probabilities of the credible intervals. The ALPHA= values must be between 0 and 1. Each ALPHA= value produces a pair of $100(1–ALPHA)\%$ equal-tail and HPD intervals for each parameters. The default is ALPHA=0.05, which yields the 95% credible intervals for each parameter.

**PERCENT=***numeric-list*

requests the percentile points of the posterior samples. The PERCENT= values must be between 0 and 100. The default is PERCENT=25, 50, 75, which yields the 25th, 50th, and 75th percentile points, respectively, for each parameter.

The list of *keywords* includes the following:

**CORR**

produces the posterior correlation matrix.

**COV**

produces the posterior covariance matrix.

**SUMMARY**

produces the means, standard deviations, and percentile points for the posterior samples. The default is to produce the 25th, 50th, and 75th percentile points, but you can use the global PERCENT= option to request specific percentile points.

**INTERVAL**

produces equal-tail credible intervals and HPD intervals. The default is to produce the 95% equal-tail credible intervals and 95% HPD intervals, but you can use the global ALPHA= option to request intervals of any probabilities.

**NONE**

suppresses printing all summary statistics.

**THINNING=***number*

**THIN=***number*

controls the thinning of the Markov chain. Only one in every $k$ samples is used when THINNING=$k$, and if NBI=$n_0$ and NMC=$n$, the number of samples kept is

$$\left[\frac{n_0 + n}{k}\right] - \left[\frac{n_0}{k}\right]$$

where $[a]$ represents the integer part of the number $a$. The default is THINNING=1.

**WEIBULLSCALEPRIOR=GAMMA< (*options*) >**

**WSCALEPRIOR=GAMMA< (*options*) >**

**WSCPRIOR=GAMMA< (*options*) >**

>    specifies that Gibbs sampling be performed on the Weibull model scale parameter and the prior distribution for the scale parameter. This option applies only when a Weibull distribution and no covariates are specified. When this option is specified, PROC LIFEREG performs Gibbs sampling on the Weibull scale parameter, which is defined as $\exp(\mu)$, where $\mu$ is the intercept term.

>    A gamma prior $G(a, b)$ is specified by WEIBULLSCALEPRIOR=GAMMA, which can be followed by one of the following *gamma-options* enclosed in parentheses. The gamma probability density is given by $g(t) = \frac{b(bt)^{a-1}e^{-bt}}{\Gamma(a)}$. The hyperparameters $a$ and $b$ are the shape and inverse-scale parameters of the gamma distribution, respectively. See the section "Gamma Prior" on page 3949 for details about the gamma prior. The default is $G(10^{-4}, 10^{-4})$.

>    **RELSHAPE< =*c* >**
>    >    specifies independent $G(c\hat{\alpha}, c)$ distribution, where $\hat{\alpha}$ is the MLE of the Weibull scale parameter. With this choice of hyperparameters, the mean of the prior distribution is $\hat{\alpha}$ and the variance is $\frac{\hat{\alpha}}{c}$. By default, $c=10^{-4}$.

>    **SHAPE=*a***
>    **ISCALE=*b***
>    >    when both specified, results in a $G(a, b)$ prior.

>    **SHAPE=*c***
>    >    when specified alone, results in a $G(c, c)$ prior.

>    **ISCALE=*c***
>    >    when specified alone, results in a $G(c, c)$ prior.

**WEIBULLSHAPEPRIOR=GAMMA< (*options*) >**

**WSHAPEPRIOR=GAMMA< (*options*) >**

**WSHPRIOR=GAMMA< (*options*) >**

>    specifies that Gibbs sampling be performed on the Weibull model shape parameter and the prior distribution for the shape parameter. When this option is specified, PROC LIFEREG performs Gibbs sampling on the Weibull shape parameter, which is defined as $\sigma^{-1}$, where $\sigma$ is the location-scale model scale parameter.

>    A gamma prior $G(a, b)$ with density $f(t) = \frac{b(bt)^{a-1}e^{-bt}}{\Gamma(a)}$ is specified by WEIBULL-SHAPEPRIOR=GAMMA, which can be followed by one of the following *gamma-options* enclosed in parentheses. The hyperparameters $a$ and $b$ are the shape and inverse-scale parameters of the gamma distribution, respectively. See the section "Gamma Prior" on page 3949 for details about the gamma prior. The default is $G(10^{-4}, 10^{-4})$.

>    **RELSHAPE< =*c* >**
>    >    specifies independent $G(c\hat{\beta}, c)$ distribution, where $\hat{\beta}$ is the MLE of the Weibull shape parameter. With this choice of hyperparameters, the mean of the prior distribution is $\hat{\beta}$ and the variance is $\frac{\hat{\beta}}{c}$. By default, $c=10^{-4}$.

> **SHAPE**< =*a*>
> **ISCALE=***b*
>> when both specified, results in a $G(a, b)$ prior.
>
> **SHAPE=***c*
>> when specified alone, results in a $G(c, c)$ prior.
>
> **ISCALE=***c*
>> when specified alone, results in a $G(c, c)$ prior.

# BY Statement

> **BY** *variables* ;

You can specify a BY statement with PROC LIFEREG to obtain separate analyses of observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data by using the SORT procedure with a similar BY statement.

- Specify the NOTSORTED or DESCENDING option in the BY statement for the LIFEREG procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.

- Create an index on the BY variables by using the DATASETS procedure (in Base SAS software).

For more information about BY-group processing, see the discussion in *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in the *Base SAS Procedures Guide*.

# CLASS Statement

> **CLASS** *variables* < / **TRUNCATE** > ;

The CLASS statement names the classification variables to be used in the model. Typical classification variables are Treatment, Sex, Race, Group, and Replication. If you use the CLASS statement, it must appear before the MODEL statement.

Classification variables can be either character or numeric. By default, class levels are determined from the entire set of formatted values of the CLASS variables.

**NOTE:** Prior to SAS 9, class levels were determined by using no more than the first 16 characters of the formatted values. To revert to this previous behavior, you can use the TRUNCATE option in the CLASS statement.

In any case, you can use formats to group values into levels. See the discussion of the FORMAT procedure in the *Base SAS Procedures Guide* and the discussions of the FORMAT statement and SAS formats in *SAS*

*Formats and Informats: Reference*. You can adjust the order of CLASS variable levels with the ORDER= option in the PROC LIFEREG statement. You can specify the following option in the CLASS statement after a slash (/):

**TRUNCATE**

    specifies that class levels should be determined by using only up to the first 16 characters of the formatted values of CLASS variables. When formatted values are longer than 16 characters, you can use this option to revert to the levels as determined in releases prior to SAS 9.

## EFFECTPLOT Statement

    **EFFECTPLOT** < *plot-type* < (*plot-definition-options*) > > < / *options* > ;

The EFFECTPLOT statement produces a display of the fitted model and provides options for changing and enhancing the displays. Table 51.3 describes the available *plot-types* and their *plot-definition-options*.

**Table 51.3**  *Plot-Types* and *Plot-Definition-Options*

| Plot-Type and Description | Plot-Definition-Options |
|---|---|
| **BOX** <br> Displays a box plot of continuous response data at each level of a CLASS effect, with predicted values superimposed and connected by a line. This is an alternative to the INTERACTION *plot-type*. | PLOTBY= variable or CLASS effect <br> X= CLASS variable or effect |
| **CONTOUR** <br> Displays a contour plot of predicted values against two continuous covariates. | PLOTBY= variable or CLASS effect <br> X= continuous variable <br> Y= continuous variable |
| **FIT** <br> Displays a curve of predicted values versus a continuous variable. | PLOTBY= variable or CLASS effect <br> X= continuous variable |
| **INTERACTION** <br> Displays a plot of predicted values (possibly with error bars) versus the levels of a CLASS effect. The predicted values are connected with lines and can be grouped by the levels of another CLASS effect. | PLOTBY= variable or CLASS effect <br> SLICEBY= variable or CLASS effect <br> X= CLASS variable or effect |
| **SLICEFIT** <br> Displays a curve of predicted values versus a continuous variable grouped by the levels of a CLASS effect. | PLOTBY= variable or CLASS effect <br> SLICEBY= variable or CLASS effect <br> X= continuous variable |

For full details about the syntax and options of the EFFECTPLOT statement, see the section "EFFECTPLOT Statement" on page 411 in Chapter 19, "Shared Concepts and Topics."

# ESTIMATE Statement

> **ESTIMATE** < *'label'* > *estimate-specification* < **(divisor=n)** >
>     < **,** ... < *'label'* > *estimate-specification* < **(divisor=n)** > >
>     < **/** *options* > **;**

The ESTIMATE statement provides a mechanism for obtaining custom hypothesis tests. Estimates are formed as linear estimable functions of the form $\mathbf{L}\boldsymbol{\beta}$. You can perform hypothesis tests for the estimable functions, construct confidence limits, and obtain specific nonlinear transformations.

Table 51.4 summarizes the options available in the ESTIMATE statement.

**Table 51.4**   ESTIMATE Statement Options

| Option | Description |
| --- | --- |
| **Construction and Computation of Estimable Functions** | |
| DIVISOR= | Specifies a list of values to divide the coefficients |
| NOFILL | Suppresses the automatic fill-in of coefficients for higher-order effects |
| SINGULAR= | Tunes the estimability checking difference |
| **Degrees of Freedom and *p*-values** | |
| ADJUST= | Determines the method for multiple comparison adjustment of estimates |
| ALPHA=$\alpha$ | Determines the confidence level $(1 - \alpha)$ |
| LOWER | Performs one-sided, lower-tailed inference |
| STEPDOWN | Adjusts multiplicity-corrected *p*-values further in a step-down fashion |
| TESTVALUE= | Specifies values under the null hypothesis for tests |
| UPPER | Performs one-sided, upper-tailed inference |
| **Statistical Output** | |
| CL | Constructs confidence limits |
| CORR | Displays the correlation matrix of estimates |
| COV | Displays the covariance matrix of estimates |
| E | Prints the $\mathbf{L}$ matrix |
| JOINT | Produces a joint *F* or chi-square test for the estimable functions |
| PLOTS= | Requests ODS statistical graphics if the analysis is sampling-based |
| SEED= | Specifies the seed for computations that depend on random numbers |

For details about the syntax of the ESTIMATE statement, see the section "ESTIMATE Statement" on page 437 in Chapter 19, "Shared Concepts and Topics."

# INSET Statement

**INSET** *< keyword-list >* *< / options >* **;**

The box or table of summary information produced on plots made with the PROBPLOT statement is called an *inset*. You can use the INSET statement to customize the information that is displayed in the inset box as well as to customize the appearance of the inset box. To supply the information that is displayed in the inset box, you specify *keywords* corresponding to the information that you want shown. For example, the following statements produce a probability plot with the number of observations, the number of right-censored observations, the name of the distribution, and the estimated Weibull shape parameter in the inset:

```
proc lifereg data=epidemic;
   model life = dose / dist = Weibull;
   probplot;
   inset nobs right dist shape;
run;
```

By default, inset entries are identified with appropriate labels. However, you can provide a customized label by specifying the *keyword* for that entry followed by the equal sign (=) and the label in quotes. For example, the following INSET statement produces an inset containing the number of observations and the name of the distribution, labeled "Sample Size" and "Distribution" in the inset:

```
inset nobs='Sample Size' dist='Distribution';
```

If you specify a keyword that does not apply to the plot you are creating, then the keyword is ignored.

If you specify more than one INSET statement, only the first one is used.

Table 51.5 lists keywords available in the INSET statement to display summary statistics, distribution parameters, and distribution fitting information.

**Table 51.5**   INSET Statement Keywords

| Keyword | Description |
| --- | --- |
| CONFIDENCE | Confidence coefficient for all confidence intervals |
| DIST | Name of the distribution |
| INTERVAL | Number of interval-censored observations |
| LEFT | Number of left-censored observations |
| NOBS | Number of observations |
| NMISS | Number of observations with missing values |
| RIGHT | Number of right-censored observations |
| SCALE | Value of the scale parameter |
| SHAPE | Value of the shape parameter |
| UNCENSORED | Number of uncensored observations |

The following *options* control the appearance of the box when you use traditional graphics. These options are not available if ODS Graphics is enabled. Table 51.6 summarizes the options available in the INSET statement.

**Table 51.6** INSET Statement Options

| Option | Description |
| --- | --- |
| CFILL= | Specifies the color for the filling box |
| CFILLH= | Specifies the color for the filling box header |
| CFRAME= | Specifies the color for the frame |
| CHEADER= | Specifies the color for text in the header |
| CTEXT= | Specifies the color for the text |
| FONT= | Specifies the software font for the text |
| HEIGHT= | Specifies the height of the text |
| HEADER= | Specifies the text for the header or box title |
| NOFRAME | Omits the frame around the box |
| POS= | Determines the position of the inset |
| REFPOINT= | Specifies the reference point for an inset |

All *options* are specified after the slash (/) in the INSET statement.

**CFILL=***color*
  specifies the color for the filling box.

**CFILLH=***color*
  specifies the color for the filling box header.

**CFRAME=***color*
  specifies the color for the frame.

**CHEADER=***color*
  specifies the color for text in the header.

**CTEXT=***color*
  specifies the color for the text.

**FONT=***font*
  specifies the software font for the text.

**HEIGHT=***value*
  specifies the height of the text.

**HEADER=***'quoted string'*
  specifies the text for the header or box title.

**NOFRAME**
  omits the frame around the box.

**POS=***value* **< DATA | PERCENT >**
  determines the position of the inset. The *value* can be a compass point (N, NE, E, SE, S, SW, W, NW) or a pair of coordinates (x, y) enclosed in parentheses. The coordinates can be specified in screen percentage units or axis data units. The default is screen percentage units.

**REFPOINT=**name

 specifies the reference point for an inset that is positioned by a pair of coordinates with the POS= option. You use the REFPOINT= option in conjunction with the POS= coordinates. The REFPOINT= option specifies which corner of the inset frame you have specified with coordinates (x, y), and it can take the value of BR (bottom right), BL (bottom left), TR (top right), or TL (top left). The default is REFPOINT=BL. If the inset position is specified as a compass point, then the REFPOINT= option is ignored.

## LSMEANS Statement

 **LSMEANS** < *model-effects* > < / *options* > **;**

The LSMEANS statement computes and compares least squares means (LS-means) of fixed effects. LS-means are *predicted population margins*—that is, they estimate the marginal means over a balanced population. In a sense, LS-means are to unbalanced designs as class and subclass arithmetic means are to balanced designs.

Table 51.7 summarizes the options available in the LSMEANS statement.

**Table 51.7** LSMEANS Statement Options

| Option | Description |
|---|---|
| **Construction and Computation of LS-Means** | |
| AT | Modifies the covariate value in computing LS-means |
| BYLEVEL | Computes separate margins |
| DIFF | Requests differences of LS-means |
| OM= | Specifies the weighting scheme for LS-means computation as determined by the input data set |
| SINGULAR= | Tunes estimability checking |
| | |
| **Degrees of Freedom and *p*-values** | |
| ADJUST= | Determines the method for multiple-comparison adjustment of LS-means differences |
| ALPHA=$\alpha$ | Determines the confidence level $(1 - \alpha)$ |
| STEPDOWN | Adjusts multiple-comparison *p*-values further in a step-down fashion |
| | |
| **Statistical Output** | |
| CL | Constructs confidence limits for means and mean differences |
| CORR | Displays the correlation matrix of LS-means |
| COV | Displays the covariance matrix of LS-means |
| E | Prints the **L** matrix |
| LINES | Produces a "Lines" display for pairwise LS-means differences |
| MEANS | Prints the LS-means |
| PLOTS= | Requests graphs of means and mean comparisons |
| SEED= | Specifies the seed for computations that depend on random numbers |

For details about the syntax of the LSMEANS statement, see the section "LSMEANS Statement" on page 453 in Chapter 19, "Shared Concepts and Topics."

## LSMESTIMATE Statement

> **LSMESTIMATE** *model-effect < 'label' > values < divisor=n >*
> *<, ...< 'label' > values < divisor=n > >*
> *< / options > ;*

The LSMESTIMATE statement provides a mechanism for obtaining custom hypothesis tests among least squares means.

Table 51.8 summarizes the options available in the LSMESTIMATE statement.

**Table 51.8** LSMESTIMATE Statement Options

| Option | Description |
|---|---|
| **Construction and Computation of LS-Means** | |
| AT | Modifies covariate values in computing LS-means |
| BYLEVEL | Computes separate margins |
| DIVISOR= | Specifies a list of values to divide the coefficients |
| OM= | Specifies the weighting scheme for LS-means computation as determined by a data set |
| SINGULAR= | Tunes estimability checking |
| **Degrees of Freedom and *p*-values** | |
| ADJUST= | Determines the method for multiple-comparison adjustment of LS-means differences |
| ALPHA=$\alpha$ | Determines the confidence level $(1 - \alpha)$ |
| LOWER | Performs one-sided, lower-tailed inference |
| STEPDOWN | Adjusts multiple-comparison *p*-values further in a step-down fashion |
| TESTVALUE= | Specifies values under the null hypothesis for tests |
| UPPER | Performs one-sided, upper-tailed inference |
| **Statistical Output** | |
| CL | Constructs confidence limits for means and mean differences |
| CORR | Displays the correlation matrix of LS-means |
| COV | Displays the covariance matrix of LS-means |
| E | Prints the **L** matrix |
| ELSM | Prints the **K** matrix |
| JOINT | Produces a joint *F* or chi-square test for the LS-means and LS-means differences |
| PLOTS= | Requests graphs of means and mean comparisons |
| SEED= | Specifies the seed for computations that depend on random numbers |

For details about the syntax of the LSMESTIMATE statement, see the section "LSMESTIMATE Statement" on page 470 in Chapter 19, "Shared Concepts and Topics."

## MODEL Statement

> < *label*: > **MODEL** *response*< * *censor*(*list*) > = *effects* < / *options* > ;

> < *label*: > **MODEL** (*lower*,*upper*)= *effects* < / *options* > ;

> < *label*: > **MODEL** *events/trials* = *effects* < / *options* > ;

Only a single MODEL statement can be used with one invocation of the LIFEREG procedure. If multiple MODEL statements are present, only the last is used. The optional *label* is used to label the model estimates in the output SAS data set and OUTEST= data set.

The first MODEL syntax is appropriate for right censoring. The variable *response* is possibly right censored. If the *response* variable can be right censored, then a second variable, denoted *censor*, must appear after the *response* variable with a list of parenthesized values, separated by commas or blanks, to indicate censoring. That is, if the *censor* variable takes on a value given in the list, the *response* is a right-censored value; otherwise, it is an observed value.

The second MODEL syntax specifies two variables, *lower* and *upper*, that contain values of the endpoints of the censoring interval. If the two values are the same (and not missing), it is assumed that there is no censoring and the actual response value is observed. If the lower value is missing, then the upper value is used as a left-censored value. If the upper value is missing, then the lower value is taken as a right-censored value. If both values are present and the lower value is less than the upper value, it is assumed that the values specify a censoring interval. If the lower value is greater than the upper value or both values are missing, then the observation is not used in the analysis, although predicted values can still be obtained if none of the covariates are missing.

The following table summarizes the ways of specifying censoring.

| *lower* | *upper* | **Comparison** | **Interpretation** |
|---|---|---|---|
| Not missing | Not missing | Equal | No censoring |
| Not missing | Not missing | Lower < upper | Censoring interval |
| Missing | Not missing | | Upper used as left-censoring value |
| Not missing | Missing | | Lower used as right-censoring value |
| Not missing | Not missing | Lower > upper | Observation not used |
| Missing | Missing | | Observation not used |

The third MODEL syntax specifies two variables that contain count data for a binary response. The value of the first variable, *events*, is the number of successes. The value of the second variable, *trials*, is the number of tries. The values of both *events* and (*trials-events*) must be nonnegative, and *trials* must be positive for the response to be valid. The values of the two variables do not need to be integers and are not modified to be integers.

The *effects* following the equal sign are the covariates in the model. Higher-order effects, such as interactions and nested terms, are allowed in the list, similar to the GLM procedure. Variable names and combinations of variable names representing higher-order terms are allowed to appear in this list. Classification, or CLASS, variables can be used as effects, and indicator variables are generated for the class levels. If you do not specify any covariates following the equal sign, an intercept-only model is fit.

Examples of three valid MODEL statements follow:

```
a: model time*flag(1,3)=temp;

b: model (start, finish)=;

c: model r/n=dose;
```

MODEL statement a indicates that the response is contained in a variable named time and that, if the variable flag takes on the values 1 or 3, the observation is right censored. The explanatory variable is temp, which could be a CLASS variable. MODEL statement b indicates that the response is known to be in the interval between the values of the variables start and finish and that there are no covariates except for a default intercept term. MODEL statement c indicates a binary response, with the variable r containing the number of responses and the variable n containing the number of trials.

Table 51.9 summarizes the options available in the MODEL statement.

**Table 51.9** MODEL Statement Options

| Option | Description |
|--------|-------------|
| **Model specification** | |
| ALPHA= | Sets the significance level |
| DISTRIBUTION= | Specifies the distribution type for failure time |
| NOLOG | Requests no log transformation of response |
| INTERCEPT= | Specifies initial estimate for intercept term |
| NOINT | Holds the intercept term fixed |
| INITIAL= | Specifies initial estimates for regression parameters |
| OFFSET= | Specifies an offset variable |
| SCALE= | Initializes the scale parameter |
| NOSCALE | Holds the scale parameter fixed |
| SHAPE1= | Initializes the first shape parameter |
| NOSHAPE1 | Holds the first shape parameter fixed |
| **Model fitting** | |
| CONVERGE= | Sets the convergence criterion |
| MAXITER= | Sets the maximum number of iterations |
| SINGULAR= | Sets the tolerance for testing singularity |
| **Output** | |
| CORRB | Displays the estimated correlation matrix |
| COVB | Displays the estimated covariance matrix |
| ITPRINT | Displays the iteration history, final gradient, and second derivative matrix |

The following options can appear in the MODEL statement.

**ALPHA=***value*

sets the significance level for the confidence intervals for regression parameters and estimated survival probabilities. The value must be between 0 and 1. By default, ALPHA=0.05.

**CONVERGE=***value*

sets the convergence criterion. Convergence is declared when the maximum change in the parameter estimates between Newton-Raphson steps is less than the value specified. The change is a relative change if the parameter is greater than 0.01 in absolute value; otherwise, it is an absolute change. By default, CONVERGE=1E–8.

**CONVG=***value*

sets the relative Hessian convergence criterion; *value* must be between 0 and 1. After convergence is determined with the change in parameter criterion specified with the CONVERGE= option, the quantity $tc = \frac{\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}}{|f|}$ is computed and compared to *value*, where $\mathbf{g}$ is the gradient vector, $\mathbf{H}$ is the Hessian matrix for the model parameters, and $f$ is the log-likelihood function. If $tc$ is greater than *value*, a warning that the relative Hessian convergence criterion has been exceeded is displayed. This criterion detects the occasional case where the change in parameter convergence criterion is satisfied, but a maximum in the log-likelihood function has not been attained. By default, CONVG=1E–4.

**CORRB**

produces the estimated correlation matrix of the parameter estimates.

**COVB**

produces the estimated covariance matrix of the parameter estimates.

**DISTRIBUTION=***distribution-type*

**DIST=***distribution-type*

**D=***distribution-type*

specifies the distribution type assumed for the failure time. By default, PROC LIFEREG fits a type 1 extreme-value distribution to the log of the response. This is equivalent to fitting the Weibull distribution, since the scale parameter for the extreme-value distribution is related to a Weibull shape parameter and the intercept is related to the Weibull scale parameter in this case. When the NOLOG option is specified, PROC LIFEREG models the untransformed response with a type 1 extreme-value distribution as the default. See the section "Supported Distributions" on page 3933 for descriptions of the distributions. The following are valid values for *distribution-type*:

| | |
|---|---|
| **EXPONENTIAL** | the exponential distribution, which is treated as a restricted Weibull distribution |
| **GAMMA** | a generalized gamma distribution (Lawless 2003, p. 240). The standard two-parameter gamma distribution is not available in PROC LIFEREG. |
| **LLOGISTIC** | a log-logistic distribution |
| **LNORMAL** | a lognormal distribution |
| **LOGISTIC** | a logistic distribution (equivalent to LLOGISTIC when the NOLOG option is specified) |
| **NORMAL** | a normal distribution (equivalent to LNORMAL when the NOLOG option is specified) |

**WEIBULL**    a Weibull distribution. If NOLOG is specified, it fits a type 1 extreme-value distribution to the raw, untransformed data.

By default, PROC LIFEREG transforms the response with the natural logarithm before fitting the specified model when you specify the GAMMA, LLOGISTIC, LNORMAL, or WEIBULL option. You can suppress the log transformation with the NOLOG option. The following table summarizes the resulting distributions when the preceding distribution options are used in combination with the NOLOG option.

| DISTRIBUTION= | NOLOG Specified? | Resulting Distribution |
|---|---|---|
| EXPONENTIAL | No | Exponential |
| EXPONENTIAL | Yes | One-parameter extreme value |
| GAMMA | No | Generalized log-gamma using the log of the response. (This is the same as fitting the generalized gamma using the untransformed response.) |
| GAMMA | Yes | Generalized log-gamma with untransformed responses |
| LOGISTIC | No | Logistic |
| LOGISTIC | Yes | Logistic (NOLOG has no effect) |
| LLOGISTIC | No | Log-logistic |
| LLOGISTIC | Yes | Logistic |
| LNORMAL | No | Lognormal |
| LNORMAL | Yes | Normal |
| NORMAL | No | Normal |
| NORMAL | Yes | Normal (NOLOG has no effect) |
| WEIBULL | No | Weibull |
| WEIBULL | Yes | Extreme value |

**INITIAL=***values*

sets initial values for the regression parameters. This option can be helpful in the case of convergence difficulty. Specified values are used to initialize the regression coefficients for the covariates specified in the MODEL statement. The intercept parameter is initialized with the INTERCEPT= option and is not included here. The values are assigned to the variables in the MODEL statement in the same order in which they are listed in the MODEL statement. Note that a CLASS variable requires $k - 1$ values when the CLASS variable takes on $k$ different levels. The order of the CLASS levels is determined by the ORDER= option. If there is no intercept term, the first CLASS variable requires $k$ initial values. If a BY statement is used, all CLASS variables must take on the same number of levels in each BY group or no meaningful initial values can be specified. The INITIAL= option can be specified as follows.

| Type of List | Specification |
|---|---|
| List separated by blanks | `initial=3 4 5` |
| List separated by commas | `initial=3,4,5` |
| x to y | `initial=3 to 5` |
| x to y by z | `initial=3 to 5 by 1` |
| Combination of methods | `initial=1,3 to 5,9` |

By default, PROC LIFEREG computes initial estimates with ordinary least squares. See the section "Computational Method" on page 3931 for details.

**NOTE:** The INITIAL= option is overwritten by the INEST= option. See the section "INEST= Data Set" on page 3945 for details.

**INTERCEPT=**_value_

initializes the intercept term to _value_. By default, the intercept is initialized by an ordinary least squares estimate.

**ITPRINT**

displays the iteration history for computing maximum likelihood estimates, the final evaluation of the gradient, and the final evaluation of the negative of the second derivative matrix—that is, the negative of the Hessian. If you perform a Bayesian analysis by specifying the BAYES statement, the iteration history for computing the mode of the posterior distribution is also displayed.

**MAXITER=**_n_

sets the maximum allowable number of iterations during the model estimation. By default, MAX-ITER=50.

**NOINT**

holds the intercept term fixed. Because of the usual log transformation of the response, the intercept parameter is usually a scale parameter for the untransformed response, or a location parameter for a transformed response.

**NOLOG**

requests that no log transformation of the response variable be performed. By default, PROC LIF-EREG models the log of the response variable for the GAMMA, LLOGISTIC, LOGNORMAL, and WEIBULL distribution options. NOLOG is implicitly assumed for the NORMAL and LOGISTIC distribution options.

**NOSCALE**

holds the scale parameter fixed. Note that if the log transformation has been applied to the response, the effect of the scale parameter is a power transformation of the original response. If no SCALE= value is specified, the scale parameter is fixed at the value 1.

**NOSHAPE1**

holds the first shape parameter, SHAPE1, fixed. If no SHAPE1= value is specified, SHAPE1 is fixed at a value that depends on the DISTRIBUTION type.

**OFFSET=**_variable_

specifies a variable in the input data set to be used as an offset variable. This variable cannot be a CLASS variable, and it cannot be the response variable or one of the explanatory variables.

**SCALE=**_value_

initializes the scale parameter to _value_. If the Weibull distribution is specified, this scale parameter is the scale parameter of the type 1 extreme-value distribution, not the Weibull scale parameter. Note that, with a log transformation, the exponential model is the same as a Weibull model with the scale parameter fixed at the value 1.

**SHAPE1=**_value_

initializes the first shape parameter to _value_. If the specified distribution does not depend on this parameter, then this option has no effect. The only distribution that depends on this shape parameter is the generalized gamma distribution. See the section "Supported Distributions" on page 3933 for descriptions of the parameterizations of the distributions.

**SINGULAR=**_value_

sets the tolerance for testing singularity of the information matrix and the crossproducts matrix for the initial least squares estimates. Roughly, the test requires that a pivot be at least this value times the original diagonal value. By default, SINGULAR=1E–12.

## OUTPUT Statement

> **OUTPUT** < **OUT=**_SAS-data-set_ > < _keyword=name_ > ... < _keyword=name_ > **;**

The OUTPUT statement creates a new SAS data set containing statistics calculated after fitting the model. At least one specification of the form _keyword=name_ is required.

All variables in the original data set are included in the new data set, along with the variables created as options for the OUTPUT statement. These new variables contain fitted values and estimated quantiles. If you want to create a SAS data set in a permanent library, you must specify a two-level name. For more information about permanent libraries and SAS data sets, see _SAS Language Reference: Concepts_. Each OUTPUT statement applies to the preceding MODEL statement. See Example 51.1 for illustrations of the OUTPUT statement.

The following specifications can appear in the OUTPUT statement:

**OUT=**_SAS-data-set_

specifies the new data set. By default, the procedure uses the DATA*n* convention to name the new data set.

_keyword=name_

specifies the statistics to include in the output data set and gives names to the new variables. Specify a keyword for each desired statistic (see the following list of keywords), an equal sign, and the variable to contain the statistic.

The keywords allowed and the statistics they represent are as follows:

**CENSORED=**_variable_

specifies a _variable_ to signal whether an observation is censored, and the type of censoring. The variable takes on values according to Table 51.10.

**Table 51.10** Censoring Variable Values

| Type of Response | CENSORED Variable Value |
|---|:---:|
| Uncensored | 0 |
| Right-censored | 1 |
| Left-censored | 2 |
| Interval-censored | 3 |

**CDF=**_variable_

specifies a *variable* to contain the estimates of the cumulative distribution function evaluated at the observed response. If the data are interval censored, then the cumulative distribution function is evaluated at the response lower interval endpoint. See the section "Predicted Values" on page 3937 for more information.

**CONTROL=**_variable_

specifies a *variable* in the input data set to control the estimation of quantiles. See Example 51.1 for an illustration. If the specified variable has the value 1, estimates for all the values listed in the QUANTILE= list are computed for that observation in the input data set; otherwise, no estimates are computed. If no CONTROL= variable is specified, all quantiles are estimated for all observations. If the response variable in the MODEL statement is binomial, then this option has no effect.

**CRESIDUAL | CRES=**_variable_

specifies a *variable* to contain the Cox-Snell residuals

$$-\log(S(u_i))$$

where $S$ is the standard survival function and

$$u_i = \frac{y_i - \mathbf{x}_i'\mathbf{b}}{\sigma}$$

If the data are interval censored, residuals are computed for $y_i$ values corresponding to lower interval endpoints. If the response variable in the corresponding model statement is binomial, then the residuals are not computed, and this variable contains missing values.

**SRESIDUAL | SRES=**_variable_

specifies a *variable* to contain the standardized residuals

$$\frac{y_i - \mathbf{x}_i'\mathbf{b}}{\sigma}$$

If the data are interval censored, residuals are computed for $y_i$ values corresponding to lower interval endpoints. If the response variable in the corresponding model statement is binomial, then the residuals are not computed, and this variable contains missing values.

**PREDICTED | P=**_variable_

specifies a *variable* to contain the quantile estimates. If the response variable in the corresponding model statement is binomial, then this variable contains the estimated probabilities, $1 - F(-\mathbf{x}'\mathbf{b})$.

**QUANTILES | QUANTILE | Q=**_value-list_

gives a list of *values* for which quantiles are calculated. The values must be between 0 and 1, noninclusive. For each value, a corresponding quantile is estimated. This option is not used if the response variable in the corresponding MODEL statement is binomial.

By default, QUANTILES=0.5. When the response is not binomial, a numeric variable, _PROB_, is added to the OUTPUT data set whenever the QUANTILES= option is specified. The variable _PROB_ gives the probability value for the quantile estimates. These are the values taken from the QUANTILES= list and are given as values between 0 and 1, not as values between 0 and 100. The list of QUANTILES values can be specified as in Table 51.11.

**Table 51.11** Types of Value Lists

| Type of List | Specification |
|---|---|
| List separated by blanks | `.2 .4 .6 .8` |
| List separated by commas | `.2,.4,.6,.8` |
| x to y | `.2 to .8` |
| x to y by z | `.2 to .8 by .1` |
| Combination of methods | `.1,.2 to .8 by .2` |

**STD_ERR | STD=**_variable_

specifies a _variable_ to contain the estimates of the standard errors of the estimated quantiles or $x'b$. If the response used in the MODEL statement is a binomial response, then these are the standard errors of $x'b$. Otherwise, they are the standard errors of the quantile estimates. These estimates can be used to compute confidence intervals for the quantiles. However, if the model is fit to the log of the event time, better confidence intervals can usually be computed by transforming the confidence intervals for the log response. See Example 51.1 for such a transformation.

**XBETA=**_variable_

specifies a _variable_ to contain the computed value of $x'b$, where $x$ is the covariate vector and $b$ is the vector of parameter estimates.

# PROBPLOT Statement

**PROBPLOT | PPLOT** < /*options* > **;**

You can use the PROBPLOT statement to create a probability plot from lifetime data. The data can be uncensored, right censored, or arbitrarily censored. You can specify any number of PROBPLOT statements after a MODEL statement. The syntax used for the response in the MODEL statement determines the type of censoring assumed in creating the probability plot. The model fit with the MODEL statement is plotted along with the data. If there are covariates in the model, they are set to constant values specified in the XDATA= data set when creating the probability plot. If no XDATA= data set is specified, continuous variables are set to their overall mean values and categorical variables specified in the CLASS statement are set to their highest levels.

Table 51.12 summarizes the options available in the PROBPLOT statement.

**Table 51.12** PROBPLOT Statement Options

| Option | Description |
|---|---|
| **Traditional Graphics** | |
| ANNOTATE= | Specifies an Annotate data set |
| CAXIS= | Specifies the color used for the axes and tick marks |
| CCENSOR= | Specifies the color for filling the censor plot area |
| CENBIN | Plots censored data as frequency counts |
| CENCOLOR= | Specifies the color for the censor symbol |
| CENSYMBOL= | Specifies symbols for censored values |

**Table 51.12** *continued*

| Option | Description |
|---|---|
| CFIT= | Specifies the color for the fitted probability line and confidence curves |
| CFRAME= | Specifies the color for the area enclosed by the axes and frame |
| CGRID= | Specifies the color for grid lines |
| CHREF= | Specifies the color for lines requested by the HREF= option |
| CTEXT= | Specifies the color for tick mark values and axis labels |
| CVREF= | Specifies the color for lines requested by the VREF= option |
| DESCRIPTION= | Specifies a description that appears in the PROC GREPLAY master menu |
| FONT= | Specifies a software font for reference line and axis labels |
| HCL | Computes and draws confidence limits |
| HEIGHT= | Specifies the height of text used outside framed areas |
| HLOWER= | Specifies the lower limit on the lifetime axis scale |
| HOFFSET= | Specifies the offset for the horizontal axis |
| HUPPER= | Specifies *value* as the upper lifetime axis tick mark |
| HREF | Draws reference lines perpendicular to the horizontal axis |
| HREFLABELS= | Specifies labels for the lines requested by the HREF= option |
| HREFLABPOS= | Specifies the vertical position of labels for HREF= lines |
| INBORDER | Requests a border around probability plots |
| INTERTILE= | Specifies the distance between tiles |
| ITPRINTEM | Displays the iteration history for the Turnbull algorithm |
| JITTER= | Specifies the amount to jitter overlaying plot symbols, in units of symbol width |
| LFIT= | Specifies a line style for fitted curves and confidence limits |
| LGRID= | Specifies a line style for all grid lines |
| LHREF= | Specifies the line type for lines requested by the HREF= option |
| LVREF= | Specifies the line type for lines requested by the VREF= option |
| MAXITEM= | Specifies the maximum number of iterations for the Turnbull algorithm |
| NAME= | Specifies a name for the plot |
| NOCENPLOT | Suppresses the plotting of censored data points |
| NOCONF | Suppresses the default confidence bands |
| NODATA | Suppresses plotting of the estimated empirical probability plot |
| NOFIT | Suppresses the fitted probability (percentile) line and confidence bands |
| NOFRAME | Suppresses the frame around plotting areas |
| NOGRID | Suppresses grid lines |
| NOHLABEL | Suppresses horizontal label |
| NOHTICK | Suppresses horizontal tick marks |
| NOPOLISH | Suppresses setting small interval probabilities to zero |
| NOVLABEL | Suppresses vertical labels |
| NOVTICK | Suppresses vertical tick marks |
| NPINTERVALS= | Displays one of the two kinds of confidence limit |
| PCTLIST= | Specifies the list of percentages for which to compute percentile estimates |
| PLOWER= | Specifies the lower limit on the probability axis scale |
| PRINTPROBS | Displays intervals and associated probabilities for the Turnbull algorithm |
| PUPPER= | Specifies the upper limit on the probability axis scale |
| PPOS= | Specifies the plotting position type |
| PPOUT | Displays a table of the cumulative probabilities |

**Table 51.12** *continued*

| Option | Description |
|---|---|
| PROBLIST= | Specifies the list of initial values for the Turnbull algorithm |
| ROTATE | Requests probability plots with probability scale on the horizontal axis |
| SQUARE | Makes the layout of the probability plots square |
| TOLLIKE= | Specifies the criterion for convergence in the Turnbull algorithm |
| TOLPROB= | Specifies the criterion for setting the interval probability to zero in the Turnbull algorithm |
| VAXISLABEL= | Specifies a label for the vertical axis |
| VREF | Draws reference lines perpendicular to the vertical axis |
| VREFLABELS= | Specifies labels for the lines requested by the VREF= option |
| VREFLABPOS= | Specifies the horizontal position of labels for VREF= lines |
| WAXIS= | Specifies line thickness for axes and frame |
| WFIT= | Specifies line thickness for fitted curves |
| WGRID= | Specifies line thickness for grids |
| WREFL= | Specifies line thickness for reference lines |

**ODS Graphics**

| Option | Description |
|---|---|
| HCL | Computes and draws confidence limits for the predicted probabilities |
| HLOWER= | Specifies *value* as the lower lifetime axis tick mark |
| HUPPER= | Specifies *value* as the upper lifetime axis tick mark |
| HREF | Draws reference lines perpendicular to the horizontal axis |
| HREFLABELS= | Specifies labels for the lines requested by the HREF= option |
| ITPRINTEM | Displays the iteration history for the Turnbull algorithm |
| MAXITEM= | Specifies the maximum number of iterations for the Turnbull algorithm |
| NOCENPLOT | Suppresses the plotting of censored data points |
| NOCONF | Suppresses the default confidence bands |
| NODATA | Suppresses plotting of the estimated empirical probability plot |
| NOFIT | Suppresses the fitted probability (percentile) line and confidence bands |
| NOFRAME | Suppresses the frame around plotting areas |
| NOGRID | Suppresses grid lines |
| NOPOLISH | Suppresses setting small interval probabilities to zero in the Turnbull algorithm |
| NPINTERVALS= | Displays one of the two kinds of confidence limits |
| PCTLIST= | Specifies the list of percentages for which to compute percentile estimates |
| PLOWER= | Specifies the lower limit on the probability axis scale |
| PRINTPROBS | Displays intervals and associated probabilities for the Turnbull algorithm |
| PUPPER= | Specifies the upper limit on the probability axis scale |
| PPOS= | Specifies the plotting position type |
| PPOUT | Displays a table of the cumulative probabilities |
| PROBLIST= | Specifies the list of initial values for the Turnbull algorithm |
| ROTATE | Requests probability plots with probability scale on the horizontal axis |
| SQUARE | Makes the layout of the probability plots square |
| TOLLIKE= | Specifies the criterion for convergence in the Turnbull algorithm |
| TOLPROB= | Specifies the criterion for setting the interval probability to zero in the Turnbull algorithm |
| VREF | Draws reference lines perpendicular to the vertical axis |

You can specify the following options to control the content, layout, and appearance of a probability plot.

## Traditional Graphics

The following options are available if you use traditional graphics—that is, if ODS Graphics is not enabled.

**ANNOTATE=**_SAS-data-set_

**ANNO=**_SAS-data-set_

> specifies an Annotate data set, as described in *SAS/GRAPH: Reference*, that enables you to add features to the probability plot. The data set you specify with the ANNOTATE= option in the PROB-PLOT statement provides the Annotate data set for all plots created by the statement.

**CAXIS=**_color_

**CAXES=**_color_

> specifies the color used for the axes and tick marks. This option overrides any COLOR= specifications in an AXIS statement. The default is the first color in the device color list.

**CCENSOR=**_color_

> specifies the color for filling the censor plot area. The default is the first color in the device color list.

**CENBIN**

> plots censored data as frequency counts (rounding for noninteger frequency) rather than as individual points.

**CENCOLOR=**_color_

> specifies the color for the censor symbol. The default is the first color in the device color list.

**CENSYMBOL=**_symbol_ | (_symbol list_)

> specifies symbols for censored values. The *symbol* is one of the symbol names (plus, star, square, diamond, triangle, hash, paw, point, dot, and circle) or a letter (A–Z). If you do not specify the CENSYMBOL= option, the symbol used for censored values is the same as for failures.

**CFIT=**_color_

> specifies the color for the fitted probability line and confidence curves. The default is the first color in the device color list.

**CFRAME=**_color_

**CFR=**_color_

> specifies the color for the area enclosed by the axes and frame. This area is not shaded by default.

**CGRID=**_color_

> specifies the color for grid lines. The default is the first color in the device color list.

**CHREF=**_color_

**CH=**_color_

> specifies the color for lines requested by the HREF= option. The default is the first color in the device color list.

**CTEXT=**_color_

> specifies the color for tick mark values and axis labels. The default is the color specified for the CTEXT= option in the most recent GOPTIONS statement.

**CVREF=***color*

**CV=***color*

 specifies the color for lines requested by the VREF= option. The default is the first color in the device color list.

**DESCRIPTION='***string***'**

**DES='***string***'**

 specifies a description, up to 40 characters, that appears in the PROC GREPLAY master menu. The default is the variable name.

**FONT=***font*

 specifies a software font for reference line and axis labels. You can also specify fonts for axis labels in an AXIS statement. The FONT= font takes precedence over the FTEXT= font specified in the most recent GOPTIONS statement. Hardware characters are used by default.

**HCL**

 computes and draws confidence limits for the predicted probabilities based on distribution percentiles instead of the default CDF limits. See the section "Confidence Limits for Percentiles" on page 3945 for details of the computation.

**HEIGHT=***value*

 specifies the height of text used outside framed areas. The default value is 3.846 (in percentage).

**HLOWER=***value*

 specifies the lower limit on the lifetime axis scale. The HLOWER= option specifies *value* as the lower lifetime axis tick mark. The tick mark interval and the upper axis limit are determined automatically.

**HOFFSET=***value*

 specifies the offset for the horizontal axis. The default value is 1.

**HUPPER=***value*

 specifies *value* as the upper lifetime axis tick mark. The tick mark interval and the lower axis limit are determined automatically.

**HREF < (INTERSECT) > =***value-list*

 requests reference lines perpendicular to the horizontal axis be drawn at horizontal axis values in the *value-list*. If (INTERSECT) is specified, a second reference line perpendicular to the vertical axis is drawn that intersects the fit line at the same point as the horizontal axis reference line. If a horizontal axis reference line label is specified with the HREFLABELS= option, the intersecting vertical axis reference line is labeled with the vertical axis value. See also the CHREF=, HREFLABELS=, and LHREF= options.

**HREFLABELS=***'label1'* **...** *'labeln'*

**HREFLABEL=***'label1'* **...** *'labeln'*

**HREFLAB=***'label1'* **...** *'labeln'*

 specifies labels for the lines requested by the HREF= option. The number of labels must equal the number of lines. Enclose each label in quotes. Labels can be up to 16 characters.

**HREFLABPOS=**$n$

    specifies the vertical position of labels for HREF= lines. The following table shows the valid values for $n$ and the corresponding label placements.

| $n$ | Label Placement |
|---|---|
| 1 | Top |
| 2 | Staggered from top |
| 3 | Bottom |
| 4 | Staggered from bottom |
| 5 | Alternating from top |
| 6 | Alternating from bottom |

**INBORDER**

    requests a border around probability plots.

**INTERTILE=***value*

    specifies the distance between tiles.

**ITPRINTEM**

    displays the iteration history for the Turnbull algorithm.

**JITTER=***value*

    specifies the amount to jitter overlaying plot symbols, in units of symbol width.

**LFIT=***linetype*

    specifies a line style for fitted curves and confidence limits. By default, fitted curves are drawn by connecting solid lines (*linetype = 1*), and confidence limits are drawn by connecting dashed lines (*linetype = 3*).

**LGRID=***linetype*

    specifies a line style for all grid lines; *linetype* is between 1 and 46. The default is 35.

**LHREF=***linetype*

**LH=***linetype*

    specifies the line type for lines requested by the HREF= option. The default is 2, which produces a dashed line.

**LVREF=***linetype*

**LV=***linetype*

    specifies the line type for lines requested by the VREF= option. The default is 2, which produces a dashed line.

**MAXITEM=***n1* < *,n2* >

    specifies the maximum number of iterations for the Turnbull algorithm. Iteration history will be displayed in increments of *n2* if requested with the ITPRINTEM option. See the section "Arbitrarily Censored Data" on page 3942 for details.

**NAME=***'string'*

    specifies a name for the plot, up to eight characters, that appears in the PROC GREPLAY master menu. The default is 'LIFEREG'.

**NOCENPLOT**

suppresses the plotting of censored data points.

**NOCONF**

suppresses the default confidence bands on the probability plot.

**NODATA**

suppresses plotting of the estimated empirical probability plot.

**NOFIT**

suppresses the fitted probability (percentile) line and confidence bands.

**NOFRAME**

suppresses the frame around plotting areas.

**NOGRID**

suppresses grid lines.

**NOHLABEL**

suppresses horizontal labels.

**NOHTICK**

suppresses horizontal tick marks.

**NOPOLISH**

suppresses setting small interval probabilities to zero in the Turnbull algorithm.

**NOVLABEL**

suppresses vertical labels.

**NOVTICK**

suppresses vertical tick marks.

**NPINTERVALS=***interval type*

specifies one of the two kinds of confidence limits for the estimated cumulative probabilities, pointwise (NPINTERVALS=POINT) or simultaneous (NPINTERVALS=SIMUL), requested by the PPOUT option to be displayed in the tabular output.

**PCTLIST=***value-list*

specifies the list of percentages for which to compute percentile estimates; *value-list* must be a list of values separated by blanks or commas. Each value in the list must be between 0 and 100.

**PLOWER=***value*

specifies the lower limit on the probability axis scale. The PLOWER= option specifies *value* as the lower probability axis tick mark. The tick mark interval and the upper axis limit are determined automatically.

**PRINTPROBS**

displays intervals and associated probabilities for the Turnbull algorithm.

**PUPPER=***value*

    specifies the upper limit on the probability axis scale. The PUPPER= option specifies *value* as the upper probability axis tick mark. The tick mark interval and the lower axis limit are determined automatically.

**PPOS=***character-list*

    specifies the plotting position type. See the section "Probability Plotting" on page 3940 for details.

| PPOS= | Method |
|---|---|
| **EXPRANK** | Expected ranks |
| **MEDRANK** | Median ranks |
| **MEDRANK1** | Median ranks (exact formula) |
| **KM** | Kaplan-Meier |
| **MKM** | Modified Kaplan-Meier (default) |

**PPOUT**

    specifies that a table of the cumulative probabilities plotted on the probability plot be displayed. Kaplan-Meier estimates of the cumulative probabilities are also displayed, along with standard errors and confidence limits. The confidence limits can be pointwise or simultaneous, as specified by the NPINTERVALS= option.

**PROBLIST=***value-list*

    specifies the list of initial values for the Turnbull algorithm.

**ROTATE**

    requests probability plots with probability scale on the horizontal axis.

**SQUARE**

    makes the layout of the probability plots square.

**TOLLIKE=***value*

    specifies the criterion for convergence in the Turnbull algorithm.

**TOLPROB=***value*

    specifies the criterion for setting the interval probability to zero in the Turnbull algorithm.

**VAXISLABEL=‘***string***’**

    specifies a label for the vertical axis.

**VREF< (INTERSECT) >=***value-list*

    requests reference lines perpendicular to the vertical axis be drawn at vertical axis values in the *value-list*. If (INTERSECT) is specified, a second reference line perpendicular to the horizontal axis is drawn that intersects the fit line at the same point as the vertical axis reference line. If a vertical axis reference line label is specified with the VREFLABELS= option, the intersecting horizontal axis reference line is labeled with the horizontal axis value. See also the CVREF=, LVREF=, and VREFLABELS= options.

**VREFLABELS=**'*label1*' ... '*labeln*'

**VREFLABEL=**'*label1*' ... '*labeln*'

**VREFLAB=**'*label1*' ... '*labeln*'

> specifies labels for the lines requested by the VREF= option. The number of labels must equal the number of lines. Enclose each label in quotes. Labels can be up to 16 characters.

**VREFLABPOS=**n

> specifies the horizontal position of labels for VREF= lines. The valid values for *n* and the corresponding label placements are shown in the following table.

| n | Label Placement |
|---|---|
| 1 | Left |
| 2 | Right |

**WAXIS=**n

> specifies line thickness for axes and frame. The default value is 1.

**WFIT=**n

> specifies line thickness for fitted curves. The default value is 1.

**WGRID=**n

> specifies line thickness for grids. The default value is 1.

**WREFL=**n

> specifies line thickness for reference lines. The default value is 1.

## ODS Graphics

The following options are available if ODS Graphics is enabled.

**HCL**

> computes and draws confidence limits for the predicted probabilities in the horizontal direction.

**HLOWER=**value

> specifies the lower limit on the lifetime axis scale. The HLOWER= option specifies *value* as the lower lifetime axis tick mark. The tick mark interval and the upper axis limit are determined automatically.

**HUPPER=**value

> specifies *value* as the upper lifetime axis tick mark. The tick mark interval and the lower axis limit are determined automatically.

**HREF < (INTERSECT) > =**value-list

> requests reference lines perpendicular to the horizontal axis be drawn at horizontal axis values in the *value-list*. If (INTERSECT) is specified, a second reference line perpendicular to the vertical axis is drawn that intersects the fit line at the same point as the horizontal axis reference line. If a horizontal axis reference line label is specified with the HREFLABELS= option, the intersecting vertical axis reference line is labeled with the vertical axis value. See also the CHREF=, HREFLABELS=, and LHREF= options.

**HREFLABELS=**'label1' ... 'labeln'

**HREFLABEL=**'label1' ... 'labeln'

**HREFLAB=**'label1' ... 'labeln'

    specifies labels for the lines requested by the HREF= option. The number of labels must equal the number of lines. Enclose each label in quotes. Labels can be up to 16 characters.

**ITPRINTEM**

    displays the iteration history for the Turnbull algorithm.

**MAXITEM=**n1 < ,n2 >

    specifies the maximum number of iterations for the Turnbull algorithm. Iteration history will be displayed in increments of *n2* if requested with the ITPRINTEM option. See the section "Arbitrarily Censored Data" on page 3942 for details.

**NOCENPLOT**

    suppresses the plotting of censored data points.

**NOCONF**

    suppresses the default confidence bands on the probability plot.

**NODATA**

    suppresses plotting of the estimated empirical probability plot.

**NOFIT**

    suppresses the fitted probability (percentile) line and confidence bands.

**NOFRAME**

    suppresses the frame around plotting areas.

**NOGRID**

    suppresses grid lines.

**NOPOLISH**

    suppresses setting small interval probabilities to zero in the Turnbull algorithm.

**NPINTERVALS=**interval type

    specifies one of the two kinds of confidence limits for the estimated cumulative probabilities, pointwise (NPINTERVALS=POINT) or simultaneous (NPINTERVALS=SIMUL), requested by the PPOUT option to be displayed in the tabular output.

**PCTLIST=**value-list

    specifies the list of percentages for which to compute percentile estimates; *value-list* must be a list of values separated by blanks or commas. Each value in the list must be between 0 and 100.

**PLOWER=**value

    specifies the lower limit on the probability axis scale. The PLOWER= option specifies *value* as the lower probability axis tick mark. The tick mark interval and the upper axis limit are determined automatically.

**PRINTPROBS**
> displays intervals and associated probabilities for the Turnbull algorithm.

**PUPPER=***value*
> specifies the upper limit on the probability axis scale. The PUPPER= option specifies *value* as the upper probability axis tick mark. The tick mark interval and the lower axis limit are determined automatically.

**PPOS=***plotting-position-type*
> specifies the plotting position type. See the section "Probability Plotting" on page 3940 for details.

| PPOS= | Method |
|---------|--------|
| **EXPRANK** | Expected ranks |
| **MEDRANK** | Median ranks |
| **MEDRANK1** | Median ranks (exact formula) |
| **KM** | Kaplan-Meier |
| **MKM** | Modified Kaplan-Meier (default) |

**PPOUT**
> specifies that a table of the cumulative probabilities plotted on the probability plot be displayed. Kaplan-Meier estimates of the cumulative probabilities are also displayed, along with standard errors and confidence limits. The confidence limits can be pointwise or simultaneous, as specified by the NPINTERVALS= option.

**PROBLIST=***value-list*
> specifies the list of initial values for the Turnbull algorithm.

**ROTATE**
> requests probability plots with probability scale on the horizontal axis.

**SQUARE**
> makes the layout of the probability plots square.

**TOLLIKE=***value*
> specifies the criterion for convergence in the Turnbull algorithm.

**TOLPROB=***value*
> specifies the criterion for setting the interval probability to zero in the Turnbull algorithm.

**VREF< (INTERSECT) >=***value-list*
> requests reference lines perpendicular to the vertical axis be drawn at vertical axis values in the *value-list*. If (INTERSECT) is specified, a second reference line perpendicular to the horizontal axis is drawn that intersects the fit line at the same point as the vertical axis reference line. If a vertical axis reference line label is specified with the VREFLABELS= option, the intersecting horizontal axis reference line is labeled with the horizontal axis value. See also the CVREF=, LVREF=, and VREFLABELS= options.

**VREFLABELS=***'label1' . . . 'labeln'*
**VREFLABEL=***'label1' . . . 'labeln'*
**VREFLAB=***'label1' . . . 'labeln'*
> specifies labels for the lines requested by the VREF= option. The number of labels must equal the number of lines. Enclose each label in quotes. Labels can be up to 16 characters.

## SLICE Statement

> **SLICE** *model-effect* < / *options* > ;

The SLICE statement provides a general mechanism for performing a partitioned analysis of the LS-means for an interaction. This analysis is also known as an analysis of simple effects.

The SLICE statement uses the same options as the LSMEANS statement, which are summarized in Table 19.21. For details about the syntax of the SLICE statement, see the section "SLICE Statement" on page 498 in Chapter 19, "Shared Concepts and Topics."

## STORE Statement

> **STORE** < **OUT=** >*item-store-name* < / **LABEL=**'*label*' > ;

The STORE statement requests that the procedure save the context and results of the statistical analysis. The resulting item store has a binary file format that cannot be modified. The contents of the item store can be processed with the PLM procedure.

For details about the syntax of the STORE statement, see the section "STORE Statement" on page 501 in Chapter 19, "Shared Concepts and Topics."

## TEST Statement

> **TEST** < *model-effects* > < / *options* > ;

The TEST statement enables you to perform chi-square tests for model effects that test Type I, Type II, or Type III hypotheses. By default, the Type III tests are performed. For more information, see Chapter 19, "Shared Concepts and Topics."

## WEIGHT Statement

> **WEIGHT** *variable* ;

If you want to use weights for each observation in the input data set, place the weights in a variable in the data set and specify the name in a WEIGHT statement. The values of the WEIGHT variable can be nonintegral and are not truncated. Observations with nonpositive or missing values for the weight variable do not contribute to the fit of the model. The WEIGHT variable multiplies the contribution to the log likelihood for each observation.

# Details: LIFEREG Procedure

## Missing Values

Any observation with missing values for the dependent variable is not used in the model estimation unless it is one and only one of the values in an interval specification. Also, if one of the explanatory variables or the censoring variable is missing, the observation is not used. For any observation to be used in the estimation of a model, only the variables needed in that model have to be nonmissing. Predicted values are computed for all observations with no missing explanatory variable values. If the censoring variable is missing, the CENSORED= variable in the OUT= SAS data set is also missing.

## Model Specification

Main effects as well as interaction terms are allowed in the model specification, similar to the GLM procedure. For numeric variables, a main effect is a linear term equal to the value of the variable unless the variable appears in the CLASS statement. For variables listed in the CLASS statement, PROC LIFEREG creates indicator variables (variables taking the values zero or one) for every level of the variable except the last level. If there is no intercept term, the first CLASS variable has indicator variables created for all levels including the last level. The levels are ordered according to the ORDER= option. Estimates of a main effect depend upon other effects in the model and, therefore, are adjusted for the presence of other effects in the model.

## Computational Method

By default, the LIFEREG procedure computes initial values for the parameters by using ordinary least squares (OLS) and ignoring censoring. This might not be the best set of starting values for a given set of data. For example, if there are extreme values in your data, the OLS fit might be excessively influenced by the extreme observations, causing an overflow or convergence problems. See Example 51.3 for one way to deal with convergence problems.

You can specify the INITIAL= option in the MODEL statement to override these starting values. You can also specify the INTERCEPT=, SCALE=, and SHAPE= options to set initial values of the intercept, scale, and shape parameters. For models with multilevel interaction effects, it is a little difficult to use the INITIAL= option to provide starting values for all parameters. In this case, you can use the INEST= data set. See the section "INEST= Data Set" on page 3945 for details. The INEST= data set overrides all previous specifications for starting values of parameters.

The rank of the design matrix $\mathbf{X}$ is estimated before the model is fit. Columns of $\mathbf{X}$ that are judged linearly dependent on other columns have the corresponding parameters set to zero. The test for linear dependence is controlled by the SINGULAR= option in the MODEL statement. Variables are included in the model in the order in which they are listed in the MODEL statement with the continuous variables included in the model before any classification variables.

The log-likelihood function is maximized by means of a ridge-stabilized Newton-Raphson algorithm. The maximized value of the log likelihood can take positive or negative values, depending on the specified model and the values of the maximum likelihood estimates of the model parameters.

If convergence of the maximum likelihood estimates is attained, a Type III chi-square test statistic is computed for each effect, testing whether there is any contribution from any of the levels of the effect. This statistic is computed as a quadratic form in the appropriate parameter estimates by using the corresponding submatrix of the asymptotic covariance matrix estimate. See Chapter 42, "The GLM Procedure," and Chapter 15, "The Four Types of Estimable Functions," for more information about Type III estimable functions. The asymptotic covariance matrix is computed as the inverse of the observed information matrix. Note that if the NOINT option is specified and CLASS variables are used, the first CLASS variable contains a contribution from an intercept term. The results are displayed in an ODS table named "Type3Analysis." Chi-square tests for individual parameters are Wald tests based on the observed information matrix and the parameter estimates. If an effect has a single degree of freedom in the parameter estimates table, the chi-square test for this parameter is equivalent to the Type III test for this effect.

Before SAS 8.2, a multiple-degree-of-freedom statistic was computed for each effect to test for contribution from any level of the effect. In general, the Type III test statistic in a main-effect-only model (no interaction terms) will be equal to the previously computed effect statistic, unless there are collinearities among the effects. If there are collinearities, the Type III statistic will adjust for them, and the value of the Type III statistic and the number of degrees of freedom might not be equal to those of the previous effect statistic.

Suppose there are $n$ observations from the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}$ (or $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{O} + \sigma\boldsymbol{\epsilon}$ if there is an offset variable), where $\mathbf{X}$ is an $n \times k$ matrix of covariate values (including the intercept), $\mathbf{y}$ is a vector of responses, $\mathbf{O}$ is a vector of offset variable values, and $\boldsymbol{\epsilon}$ is a vector of errors with survival function $S$, cumulative distribution function $F$, and probability density function $f$. That is, $S(t) = \Pr(\epsilon_i > t)$, $F(t) = \Pr(\epsilon_i \leq t)$, and $f(t) = dF(t)/dt$, where $\epsilon_i$ is a component of the error vector. Then, if all the responses are observed, the log likelihood, $L$, can be written as

$$L = \sum \log\left(\frac{f(u_i)}{\sigma}\right)$$

where $u_i = \frac{1}{\sigma}(y_i - \mathbf{x}_i'\boldsymbol{\beta})$.

If some of the responses are left, right, or interval censored, the log likelihood can be written as

$$L = \sum \log\left(\frac{f(u_i)}{\sigma}\right) + \sum \log(S(u_i)) + \sum \log(F(u_i)) + \sum \log(F(u_i) - F(v_i))$$

with the first sum over uncensored observations, the second sum over right-censored observations, the third sum over left-censored observations, the last sum over interval-censored observations, and

$$v_i = \frac{1}{\sigma}(z_i - \mathbf{x}_i'\boldsymbol{\beta})$$

where $z_i$ is the lower end of a censoring interval.

If the response is specified in the binomial format, *events/trials*, then the log-likelihood function is

$$L = \sum r_i \log(P_i) + (n_i - r_i)\log(1 - P_i)$$

where $r_i$ is the number of events and $n_i$ is the number of trials for the $i$th observation. In this case, $P_i = 1 - F(-\mathbf{x}_i'\boldsymbol{\beta})$. For the symmetric distributions, logistic and normal, this is the same as $F(\mathbf{x}_i'\boldsymbol{\beta})$. Additional

information about censored and limited dependent variable models can be found in Kalbfleisch and Prentice (1980) and Maddala (1983).

The estimated covariance matrix of the parameter estimates is computed as the negative inverse of $\mathbf{I}$, which is the information matrix of second derivatives of $L$ with respect to the parameters evaluated at the final parameter estimates. If $\mathbf{I}$ is not positive definite, a positive-definite submatrix of $\mathbf{I}$ is inverted, and the remaining rows and columns of the inverse are set to zero. If some of the parameters, such as the scale and intercept, are restricted, the corresponding elements of the estimated covariance matrix are set to zero. The standard error estimates for the parameter estimates are taken as the square roots of the corresponding diagonal elements.

For restrictions placed on the intercept, scale, and shape parameters, one-degree-of-freedom Lagrange multiplier test statistics are computed. These statistics are computed as

$$\chi^2 = \frac{g^2}{V}$$

where $g$ is the derivative of the log likelihood with respect to the restricted parameter at the restricted maximum and

$$V = \mathbf{I}_{11} - \mathbf{I}_{12}\mathbf{I}_{22}^{-1}\mathbf{I}_{21}$$

where the 1 subscripts refer to the restricted parameter and the 2 subscripts refer to the unrestricted parameters. The information matrix is evaluated at the restricted maximum. These statistics are asymptotically distributed as chi-squares with one degree of freedom under the null hypothesis that the restrictions are valid, provided that some regularity conditions are satisfied. See Rao (1973, p. 418) for a more complete discussion. It is possible for these statistics to be missing if the observed information matrix is not positive definite. Higher-degree-of-freedom tests for multiple restrictions are not currently computed.

A Lagrange multiplier test statistic is computed to test this constraint. Notice that this test statistic is comparable to the Wald test statistic for testing that the scale is one. The Wald statistic is the result of squaring the difference of the estimate of the scale parameter from one and dividing this by the square of its estimated standard error.

## Supported Distributions

For most distributions, the baseline survival function ($S$) and the probability density function($f$) are listed for the additive random disturbance ($y_0$ or $\log(T_0)$) with location parameter $\mu$ and scale parameter $\sigma$. See the section "Overview: LIFEREG Procedure" on page 3878 for more information. These distributions apply when the log of the response is modeled (this is the default analysis). The corresponding survival function ($G$) and its density function ($g$) are given for the untransformed baseline distribution ($T_0$).

For the normal and logistic distributions, the response is not log transformed by PROC LIFEREG, and the survival functions and probability density functions listed apply to the untransformed response.

For example, for the WEIBULL distribution, $S(w)$ and $f(w)$ are the survival function and the probability density function for the extreme-value distribution (distribution of the log of the response), while $G(t)$ and $g(t)$ are the survival function and the probability density function of a Weibull distribution (using the untransformed response).

The chosen baseline functions define the meaning of the intercept, scale, and shape parameters. Only the gamma distribution has a free shape parameter in the following parameterizations. Notice that some of the

distributions do not have mean zero and that $\sigma$ is not, in general, the standard deviation of the baseline distribution.

For the Weibull distribution, the accelerated failure time model is also a proportional-hazards model. However, the parameterization for the covariates differs by a multiple of the scale parameter from the parameterization commonly used for the proportional hazards model.

The distributions supported in the LIFEREG procedure follow. If there are no covariates in the model, $\mu =$ Intercept in the output; otherwise, $\mu = \mathbf{x}'\boldsymbol{\beta}$. $\sigma =$ Scale in the output.

## Exponential

$$
\begin{aligned}
S(w) &= \exp(-\exp(w - \mu)) \\
f(w) &= \exp(w - \mu)\exp(-\exp(w - \mu)) \\
G(t) &= \exp(-\alpha t) \\
g(t) &= \alpha \exp(-\alpha t)
\end{aligned}
$$

where $\exp(-\mu) = \alpha$.

## Generalized Gamma

$S(w) = S'(u)$, $f(w) = \sigma^{-1} f'(u)$, $G(t) = G'(v)$, $g(t) = \frac{v}{t\sigma} g'(v)$, $u = \frac{w - \mu}{\sigma}$, $v = \exp(\frac{\log(t) - \mu}{\sigma})$, and

$$
S'(u) =
\begin{cases}
1 - \dfrac{\Gamma(\delta^{-2}, \delta^{-2}\exp(\delta u))}{\Gamma(\delta^{-2})} & \text{if } \delta > 0 \\[2ex]
\dfrac{\Gamma(\delta^{-2}, \delta^{-2}\exp(\delta u))}{\Gamma(\delta^{-2})} & \text{if } \delta < 0
\end{cases}
$$

$$
f'(u) = \frac{|\delta|}{\Gamma\left(\delta^{-2}\right)} \left(\delta^{-2}\exp(\delta u)\right)^{\delta^{-2}} \exp\left(-\exp(\delta u)\delta^{-2}\right)
$$

$$
G'(v) =
\begin{cases}
1 - \dfrac{\Gamma(\delta^{-2}, \delta^{-2}v^{\delta})}{\Gamma(\delta^{-2})} & \text{if } \delta > 0 \\[2ex]
\dfrac{\Gamma(\delta^{-2}, \delta^{-2}v^{\delta})}{\Gamma(\delta^{-2})} & \text{if } \delta < 0
\end{cases}
$$

$$
g'(v) = \frac{|\delta|}{v\Gamma\left(\delta^{-2}\right)} \left(\delta^{-2}v^{\delta}\right)^{\delta^{-2}} \exp\left(-v^{\delta}\delta^{-2}\right)
$$

where $\Gamma(a)$ denotes the complete gamma function, $\Gamma(a, z)$ denotes the incomplete gamma function, and $\delta$ is a free shape parameter. The $\delta$ parameter is called Shape by PROC LIFEREG. See Lawless (2003, p. 240), and Klein and Moeschberger (1997, p. 386) for a description of the generalized gamma distribution.

## Logistic

$$S(w) = \left(1 + \exp\left(\frac{w-\mu}{\sigma}\right)\right)^{-1}$$

$$f(w) = \frac{\exp\left(\frac{w-\mu}{\sigma}\right)}{\sigma\left(1 + \exp\left(\frac{w-\mu}{\sigma}\right)\right)^2}$$

## Log-Logistic

$$S(w) = \left(1 + \exp\left(\frac{w-\mu}{\sigma}\right)\right)^{-1}$$

$$f(w) = \frac{\exp\left(\frac{w-\mu}{\sigma}\right)}{\sigma\left(1 + \exp\left(\frac{w-\mu}{\sigma}\right)\right)^2}$$

$$G(t) = \frac{1}{1 + \alpha t^\gamma}$$

$$g(t) = \frac{\alpha\gamma t^{\gamma-1}}{(1 + \alpha t^\gamma)^2}$$

where $\gamma = 1/\sigma$ and $\alpha = \exp(-\mu/\sigma)$.

## Lognormal

$$S(w) = 1 - \Phi\left(\frac{w-\mu}{\sigma}\right)$$

$$f(w) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{w-\mu}{\sigma}\right)^2\right)$$

$$G(t) = 1 - \Phi\left(\frac{\log(t)-\mu}{\sigma}\right)$$

$$g(t) = \frac{1}{\sqrt{2\pi}\sigma t}\exp\left(-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right)$$

where $\Phi$ is the cumulative distribution function for the normal distribution.

## Normal

$$S(w) = 1 - \Phi\left(\frac{w - \mu}{\sigma}\right)$$

$$f(w) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{w - \mu}{\sigma}\right)^2\right)$$

where $\Phi$ is the cumulative distribution function for the normal distribution.

## Weibull

$$S(w) = \exp\left(-\exp\left(\frac{w - \mu}{\sigma}\right)\right)$$

$$f(w) = \frac{1}{\sigma}\exp\left(\frac{w - \mu}{\sigma}\right)\exp\left(-\exp\left(\frac{w - \mu}{\sigma}\right)\right)$$

$$G(t) = \exp\left(-\alpha t^\gamma\right)$$

$$g(t) = \gamma\alpha t^{\gamma - 1}\exp\left(-\alpha t^\gamma\right)$$

where $\sigma = 1/\gamma$ and $\alpha = \exp(-\mu/\sigma)$.

If your parameterization is different from the ones shown here, you can still use the procedure to fit your model. For example, a common parameterization for the Weibull distribution is

$$g(t; \lambda, \beta) = \left(\frac{\beta}{\lambda}\right)\left(\frac{t}{\lambda}\right)^{\beta - 1}\exp\left(-\left(\frac{t}{\lambda}\right)^\beta\right)$$

$$G(t; \lambda, \beta) = \exp\left(-\left(\frac{t}{\lambda}\right)^\beta\right)$$

so that $\lambda = \exp(\mu)$ and $\beta = 1/\sigma$.

Again note that the expected value of the baseline log response is, in general, not zero and that the distributions are not symmetric in all cases. Thus, for a given set of covariates, $x$, the expected value of the log response is not always $x'\beta$.

Some relations among the distributions are as follows:

- The gamma with Shape=1 is a Weibull distribution.

- The gamma with Shape=0 is a lognormal distribution.

- The Weibull with Scale=1 is an exponential distribution.

## Predicted Values

For a given set of covariates, $\mathbf{x}$ (including the intercept term), the $p$th quantile of the log response, $y_p$, is given by

$$y_p = \mathbf{x}'\boldsymbol{\beta} + \sigma u_p$$

if no offset variable has been specified, or

$$y_p = \mathbf{x}'\boldsymbol{\beta} + o + \sigma u_p$$

for a given value o of an offset variable, where $u_p$ is the $p$th quantile of the baseline distribution. The estimated quantile is computed by replacing the unknown parameters with their estimates, including any shape parameters on which the baseline distribution might depend. The estimated quantile of the original response is obtained by taking the exponential of the estimated log quantile unless the NOLOG option is specified in the preceding MODEL statement.

The following table shows how $u_p$ is computed from the baseline distribution $F(u)$:

**Table 51.13** Baseline Probability Functions and $u_p$

| Distribution | $F(u)$ | $u_p$ |
|---|---|---|
| Exponential | $1 - \exp(-\exp(u))$ | $\log(-\log(1-p))$ |
| Generalized Gamma | $\begin{cases} \dfrac{\Gamma(\delta^{-2}, \delta^{-2}\exp(\delta u))}{\Gamma(\delta^{-2})} & \text{if } \delta > 0 \\ 1 - \dfrac{\Gamma(\delta^{-2}, \delta^{-2}\exp(\delta u))}{\Gamma(\delta^{-2})} & \text{if } \delta < 0 \end{cases}$ | $F^{-1}(p)$ |
| Logistic | $1 - (1+\exp(u))^{-1}$ | $\log(p/(1-p))$ |
| Log-logistic | $1 - (1+\exp(u))^{-1}$ | $\log(p/(1-p))$ |
| Lognormal | $\Phi(u)$ | $\Phi^{-1}(p)$ |
| Normal | $\Phi(u)$ | $\Phi^{-1}(p)$ |
| Weibull | $1 - \exp(-\exp(u))$ | $\log(-\log(1-p))$ |

For the generalized gamma distribution, $u_p$ is computed numerically.

The standard errors of the quantile estimates are computed using the estimated covariance matrix of the parameter estimates and a Taylor series expansion of the quantile estimate. The standard error is computed as

$$\text{STD} = \sqrt{\mathbf{z}'\mathbf{V}\mathbf{z}}$$

where $\mathbf{V}$ is the estimated covariance matrix of the parameter vector $(\boldsymbol{\beta}', \sigma, \delta)'$, and $\mathbf{z}$ is the vector

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \hat{u}_p \\ \hat{\sigma}\dfrac{\partial u_p}{\partial \delta} \end{bmatrix}$$

where $\delta$ is the vector of the shape parameters. Unless the NOLOG option is specified, this standard error estimate is converted into a standard error estimate for $\exp(y_p)$ as $\exp(\hat{y}_p)\text{STD}$. It might be more desirable

to compute confidence limits for the log response and convert them back to the original response variable than to use the standard error estimates for $\exp(y_p)$ directly. See Example 51.1 for a 90% confidence interval of the response constructed by exponentiating a confidence interval for the log response.

The variable CDF is computed as

$$\text{CDF}_i = F(u_i)$$

where the residual is defined by

$$u_i = \left( \frac{y_i - \mathbf{x}_i' \mathbf{b}}{\hat{\sigma}} \right)$$

and $F$ is the baseline cumulative distribution function. If the data are interval-censored, then the cumulative distribution function, $\text{CDF}_i = F(u_i)$, is evaluated at the lower interval endpoint.

## Confidence Intervals

Confidence intervals are computed for all model parameters and are reported in the "Analysis of Parameter Estimates" table. The confidence coefficient can be specified with the ALPHA=$\alpha$ MODEL statement option, resulting in a $(1 - \alpha) \times 100\%$ two-sided confidence coefficient. The default confidence coefficient is 95%, corresponding to $\alpha = 0.05$.

### Regression Parameters

A two-sided $(1 - \alpha) \times 100\%$ confidence interval $[\beta_{iL}, \quad \beta_{iU}]$ for the regression parameter $\beta_i$ is based on the asymptotic normality of the maximum likelihood estimator $\hat{\beta}_i$ and is computed by

$$\beta_{iL} = \hat{\beta}_i - z_{1-\alpha/2}(\text{SE}_{\hat{\beta}_i})$$

$$\beta_{iU} = \hat{\beta}_i + z_{1-\alpha/2}(\text{SE}_{\hat{\beta}_i})$$

where $\text{SE}_{\hat{\beta}_i}$ is the estimated standard error of $\hat{\beta}_i$, and $z_p$ is the $p \times 100\%$ percentile of the standard normal distribution.

### Scale Parameter

A two-sided $(1 - \alpha) \times 100\%$ confidence interval $[\sigma_L, \quad \sigma_U]$ for the scale parameter $\sigma$ in the location-scale model is based on the asymptotic normality of the logarithm of the maximum likelihood estimator $\log(\hat{\sigma})$, and is computed by

$$\sigma_L = \hat{\sigma} / \exp[z_{1-\alpha/2}(\text{SE}_{\hat{\sigma}})/\hat{\sigma}]$$

$$\sigma_U = \hat{\sigma} \exp[z_{1-\alpha/2}(\text{SE}_{\hat{\sigma}})/\hat{\sigma}]$$

See Meeker and Escobar (1998) for more information.

## Weibull Scale and Shape Parameters

The Weibull distribution scale parameter $\eta$ and shape parameter $\beta$ are obtained by transforming the extreme-value location parameter $\mu$ and scale parameter $\sigma$:

$$\eta = \exp(\mu)$$

$$\beta = 1/\sigma$$

Consequently, two-sided $(1 - \alpha) \times 100\%$ confidence intervals for the Weibull scale and shape parameters are computed as

$$[\eta_L, \quad \eta_U] = [\exp(\mu_L), \quad \exp(\mu_U)]$$

$$[\beta_L, \quad \beta_U] = [1/\sigma_U, \quad 1/\sigma_L]$$

## Gamma Shape Parameter

A two-sided $(1 - \alpha) \times 100\%$ confidence interval for the three-parameter gamma shape parameter $\delta$ is computed by

$$[\delta_L, \quad \delta_U] = [\hat{\delta} - z_{1-\alpha/2}(\mathrm{SE}_{\hat{\delta}}), \quad \hat{\delta} + z_{1-\alpha/2}(\mathrm{SE}_{\hat{\delta}})]$$

## Fit Statistics

Suppose that the model contains $p$ parameters and that $n$ observations are used in model fitting. The fit criteria displayed by the LIFEREG procedure are calculated as follows:

- –2 log likelihood:

  $$-2\log(\mathrm{L})$$

  where L is the maximized likelihood for the model.

- Akaike's information criterion:

  $$\mathrm{AIC} = -2\log(\mathrm{L}) + 2p$$

- corrected Akaike's information criterion:

  $$\mathrm{AICC} = \mathrm{AIC} + \frac{2p(p + 1)}{n - p - 1}$$

- Bayesian information criterion:

  $$\mathrm{BIC} = -2\log(\mathrm{L}) + p\log(n)$$

If you specify the Weibull, exponential, lognormal, log-logistic, or gamma distribution, then maximum likelihood estimates of model parameters are computed by maximizing the log likelihood of the distribution of the logarithm of the response. This is equivalent to computing maximum likelihood parameter estimates based on the response on the original, rather than log, scale. If you specify the Weibull, exponential, lognormal, log-logistic, or gamma distribution, then fit statistics based on the maximized log likelihood $\log(L)$ of the log of the response are reported in the "Fit Statistics" table. Fit criteria computed in this way cannot be meaningfully compared with fit criteria that are based on the log likelihood of the unlogged response. If you specify the normal or logistic distribution, or if you specify the NOLOG option in the MODEL statement, then the fit criteria reported in the "Fit Statistics" table are based on the response on the original, rather than log, scale.

In addition to the "Fit Statistics" table described previously, if you specify the Weibull, exponential, lognormal, log-logistic, or gamma distribution, fit criteria that are based on the distribution of the response on the original scale, rather than the log of the response, are reported in the "Fit Statistics (Unlogged Response)" table.

When comparing models, you should compare fit criteria based on the log likelihood that is computed by using the response on the same scale, either always based on the log of the response or always based on the response on the original scale.

See Akaike (1981, 1979) for details of AIC and BIC. See Simonoff (2003) for a discussion of using AIC, AICC, and BIC in statistical modeling.

## Probability Plotting

Probability plots are useful tools for the display and analysis of lifetime data. Probability plots use an inverse distribution scale so that a cumulative distribution function (CDF) plots as a straight line. A nonparametric estimate of the CDF of the lifetime data will plot approximately as a straight line, thus providing a visual assessment of goodness of fit.

You can use the PROBPLOT statement in PROC LIFEREG to create probability plots of data that are complete, right censored, interval censored, or a combination of censoring types (arbitrarily censored). A line representing the maximum likelihood fit from the MODEL statement and pointwise parametric confidence bands for the cumulative probabilities are also included in the plot.

A random variable $Y$ belongs to a *location-scale* family of distributions if its CDF $F$ is of the form

$$\Pr\{Y \leq y\} = F(y) = G\left(\frac{y - \mu}{\sigma}\right)$$

where $\mu$ is the location parameter and $\sigma$ is the scale parameter. Here, $G$ is a CDF that cannot depend on any unknown parameters, and $G$ is the CDF of $Y$ if $\mu = 0$ and $\sigma = 1$. For example, if $Y$ is a normal random variable with mean $\mu$ and standard deviation $\sigma$,

$$G(u) = \Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du$$

and

$$F(y) = \Phi\left(\frac{y - \mu}{\sigma}\right)$$

The normal, extreme-value, and logistic distributions are location-scale models. The three-parameter gamma distribution is a location-scale model if the shape parameter $\delta$ is fixed. If $T$ has a lognormal, Weibull, or log-logistic distribution, then $\log(T)$ has a distribution that is a location-scale model. These distributions are said to be of type log-location-scale. Probability plots are constructed for lognormal, Weibull, and log-logistic distributions by using $\log(T)$ instead of $T$ in the plots.

Let $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$ be ordered observations of a random sample with distribution function $F(y)$. A probability plot is a plot of the points $y_{(i)}$ against $m_i = G^{-1}(a_i)$, where $a_i = \hat{F}(y_i)$ is an estimate of the CDF $F(y_{(i)}) = G\left(\frac{y_{(i)} - \mu}{\sigma}\right)$. The nonparametric CDF estimates $a_i$ are sometimes called *plotting positions*. The axis on which the points $m_i$ are plotted is usually labeled with a probability scale (the scale of $a_i$).

If $F$ is one of the location-scale distributions, then $y$ is the lifetime; otherwise, the log of the lifetime is used to transform the distribution to a location-scale model.

If the data actually have the stated distribution, then $\hat{F} \approx F$,

$$m_i = G^{-1}(\hat{F}(y_i)) \approx G^{-1}\left(G\left(\frac{y_{(i)} - \mu}{\sigma}\right)\right) = \frac{y_{(i)} - \mu}{\sigma}$$

and points $(y_{(i)}, m_i)$ should fall approximately in a straight line.

There are several ways to compute the nonparametric CDF estimates used in probability plots from lifetime data. These are discussed in the next two sections.

## Complete and Right-Censored Data

The censoring times must be taken into account when you compute plotting positions for right-censored data. The modified Kaplan-Meier method described in the following section is the default method for computing nonparametric CDF estimates for display on probability plots. See Abernethy (1996), Meeker and Escobar (1998), and Nelson (1982) for discussions of the methods described in the following sections.

### *Expected Ranks, Kaplan-Meier, and Modified Kaplan-Meier Methods*

Let $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$ be ordered observations of a random sample including failure times and censor times. Order the data in increasing order. Label all the data with reverse ranks $r_i$, with $r_1 = n, \ldots, r_n = 1$. For the lifetime (not censoring time) corresponding to reverse rank $r_i$, compute the survival function estimate

$$S_i = \left[\frac{r_i}{r_i + 1}\right] S_{i-1}$$

with $S_0 = 1$. The expected rank plotting position is computed as $a_i = 1 - S_i$. The option PPOS=EXPRANK specifies the expected rank plotting position.

For the Kaplan-Meier method,

$$S_i = \left[\frac{r_i - 1}{r_i}\right] S_{i-1}$$

The Kaplan-Meier plotting position is then computed as $a_i' = 1 - S_i$. The option PPOS=KM specifies the Kaplan-Meier plotting position.

For the modified Kaplan-Meier method, use

$$S_i' = \frac{S_i + S_{i-1}}{2}$$

where $S_i$ is computed from the Kaplan-Meier formula with $S_0 = 1$. The plotting position is then computed as $a_i'' = 1 - S_i'$. The option PPOS=MKM specifies the modified Kaplan-Meier plotting position. If the PPOS option is not specified, the modified Kaplan-Meier plotting position is used as the default method.

For complete samples, $a_i = i/(n+1)$ for the expected rank method, $a_i' = i/n$ for the Kaplan-Meier method, and $a_i'' = (i-0.5)/n$ for the modified Kaplan-Meier method. If the largest observation is a failure for the Kaplan-Meier estimator, then $F_n = 1$ and the point is not plotted.

### *Median Ranks*

Let $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$ be ordered observations of a random sample including failure times and censor times. A failure order number $j_i$ is assigned to the $i$th failure: $j_i = j_{i-1} + \Delta$, where $j_0 = 0$. The increment $\Delta$ is initially 1 and is modified when a censoring time is encountered in the ordered sample. The new increment is computed as

$$\Delta = \frac{(n+1) - \text{ previous failure order number}}{1 + \text{ number of items beyond previous censored item}}$$

The plotting position is computed for the $i$th failure time as

$$a_i = \frac{j_i - 0.3}{n + 0.4}$$

For complete samples, the failure order number $j_i$ is equal to $i$, the order of the failure in the sample. In this case, the preceding equation for $a_i$ is an approximation of the median plotting position computed as the median of the $i$th-order statistic from the uniform distribution on (0, 1). In the censored case, $j_i$ is not necessarily an integer, but the preceding equation still provides an approximation to the median plotting position. The PPOS=MEDRANK option specifies the median rank plotting position.

## Arbitrarily Censored Data

The LIFEREG procedure can create probability plots for data that consist of combinations of exact, left-censored, right-censored, and interval-censored lifetimes—that is, arbitrarily censored data. The LIFEREG procedure uses an iterative algorithm developed by Turnbull (1976) to compute a nonparametric maximum likelihood estimate of the cumulative distribution function for the data. Since the technique is maximum likelihood, standard errors of the cumulative probability estimates are computed from the inverse of the associated Fisher information matrix. This algorithm is an example of the expectation-maximization (EM) algorithm. The default initial estimate assigns equal probabilities to each interval. You can specify different initial values with the PROBLIST= option. Convergence is determined if the change in the log likelihood between two successive iterations is less than delta, where the default value of delta is $10^{-8}$. You can specify a different value for delta with the TOLLIKE= option. Iterations will be terminated if the algorithm does not converge after a fixed number of iterations. The default maximum number of iterations is 1000. Some data might require more iterations for convergence. You can specify the maximum allowed number of iterations with the MAXITEM= option in the PROBPLOT statement. The iteration history of the log likelihood is displayed if you specify the ITPRINTEM option. The iteration history of the estimated interval probabilities are also displayed if you specify both options ITPRINTEM and PRINTPROBS.

If an interval probability is smaller than a tolerance ($10^{-6}$ by default) after convergence, the probability is set to zero, the interval probabilities are renormalized so that they add to one, and iterations are restarted. Usually the algorithm converges in just a few more iterations. You can change the default value of the tolerance with the TOLPROB= option. You can specify the NOPOLISH option to avoid setting small probabilities to zero and restarting the algorithm.

If you specify the ITPRINTEM option, a table summarizing the Turnbull estimate of the interval probabilities is displayed. The columns labeled "Reduced Gradient" and "Lagrange Multiplier" are used in checking final convergence of the maximum likelihood estimate. The Lagrange multipliers must all be greater than or equal to zero, or the solution is not maximum likelihood. See Gentleman and Geyer (1994) for more details of the convergence checking. Also see Meeker and Escobar (1998, Chapter 3) for more information.

See Example 51.6 for an illustration.

## Nonparametric Confidence Intervals

You can use the PPOUT option in the PROBPLOT statement to create a table containing the nonparametric CDF estimates computed by the selected method, Kaplan-Meier CDF estimates, standard errors of the Kaplan-Meier estimator, and nonparametric confidence limits for the CDF. The confidence limits are either pointwise or simultaneous, depending on the value of the NPINTERVALS= option in the PROBPLOT statement. The method used in the LIFEREG procedure for computation of approximate pointwise and simultaneous confidence intervals for cumulative failure probabilities relies on the Kaplan-Meier estimator of the cumulative distribution function of failure time and approximate standard deviation of the Kaplan-Meier estimator. For the case of arbitrarily censored data, the Turnbull algorithm, discussed previously, provides an extension of the Kaplan-Meier estimator. Both the Kaplan-Meier and the Turnbull estimators provide an estimate of the standard error of the CDF estimator, $\mathrm{se}_{\hat{F}}$, that is used in computing confidence intervals.

### *Pointwise Confidence Intervals*
Approximate $(1 - \alpha)100\%$ pointwise confidence intervals are computed as in Meeker and Escobar (1998, Section 3.6) as

$$[F_L, \quad F_U] = \left[ \frac{\hat{F}}{\hat{F} + (1 - \hat{F})w}, \quad \frac{\hat{F}}{\hat{F} + (1 - \hat{F})/w} \right]$$

where

$$w = \exp\left[ \frac{z_{1-\alpha/2}\mathrm{se}_{\hat{F}}}{(\hat{F}(1 - \hat{F}))} \right]$$

where $z_p$ is the $p$th quantile of the standard normal distribution.

### *Simultaneous Confidence Intervals*
Approximate $(1 - \alpha)100\%$ simultaneous confidence bands valid over the lifetime interval $(t_a, t_b)$ are computed as the "Equal Precision" case of Nair (1984) and Meeker and Escobar (1998, Section 3.8) as

$$[F_L, \quad F_U] = \left[ \frac{\hat{F}}{\hat{F} + (1 - \hat{F})w}, \quad \frac{\hat{F}}{\hat{F} + (1 - \hat{F})/w} \right]$$

where

$$w = \exp\left[\frac{e_{a,b,1-\alpha/2}\mathrm{se}_{\hat{F}}}{(\hat{F}(1-\hat{F}))}\right]$$

where the factor $x = e_{a,b,1-\alpha/2}$ is the solution of

$$x \exp(-x^2/2) \log\left[\frac{(1-a)b}{(1-b)a}\right]/\sqrt{8\pi} = \alpha/2$$

The time interval $(t_a, t_b)$ over which the bands are valid depends in a complicated way on the constants $a$ and $b$ defined in Nair (1984), $0 < a < b < 1$. The constants $a$ and $b$ are chosen by default so that the confidence bands are valid between the lowest and highest times corresponding to failures in the case of multiply censored data, or to the lowest and highest intervals for which probabilities are computed for arbitrarily censored data. You can optionally specify $a$ and $b$ directly with the NPINTERVALS=SIMULTANEOUS($a$, $b$) option in the PROBPLOT statement.

## Parametric Confidence Intervals

Pointwise parametric confidence bands are displayed in a probability plot, unless you specify the NOCONF option in the PROBPLOT statement. Two kinds of confidence intervals are available for display in a probability plot: confidence limits for the estimated cumulative distribution function (CDF) and confidence limits for estimated distribution percentiles.

### Confidence Limits for the Estimated CDF

If the distribution is of type log-location-scale, let $y = \log(t)$ where $t$ is the value of time at which the confidence limits are to be computed. If the distribution is of type location-scale, let $y$ be the value at which you want to evaluate confidence limits for the estimated CDF $\hat{F}(y)$. Let

$$\hat{u} = \frac{y - \mathbf{x}'\hat{\boldsymbol{\beta}}}{\hat{\sigma}}$$

where the column vector $\mathbf{x}$ of covariate values is determined by the rules summarized in the section "XDATA= Data Set" on page 3947. If an offset variable is specified, the mean of the offset variable values is included in $\mathbf{x}'\boldsymbol{\beta}$.

The CDF estimate is given by

$$\hat{F}(y) = G(\hat{u})$$

where $G$ is the baseline distribution. The approximate standard error of $\hat{F}(y)$ is computed as in Meeker and Escobar (1998, Section 8.4.3) as

$$\mathrm{SE}_{\hat{F}} = \frac{g(\hat{u})}{\hat{\sigma}}\left[\mathrm{Var}(\mathbf{x}'\hat{\boldsymbol{\beta}}) + 2\hat{u}\mathrm{Cov}(\mathbf{x}'\hat{\boldsymbol{\beta}}, \hat{\sigma}) + \hat{u}^2\mathrm{Var}(\hat{\sigma})\right]^{\frac{1}{2}}$$

where $g$ is the probability density function corresponding to $G$. Two-sided $(1-\alpha) \times 100\%$ confidence limits are given by

$$[F_L, \ F_U] = \left[ \frac{\hat{F}}{\hat{F} + (1 - \hat{F}) \times w}, \ \frac{\hat{F}}{\hat{F} + (1 - \hat{F})/w} \right]$$

where

$$w = \exp\left[ \frac{z_{1-\alpha/2}\mathrm{SE}_{\hat{F}}}{\hat{F}(1 - \hat{F})} \right]$$

and $z_p$ is the $p \times 100\%$ percentile of the standard normal distribution. The quantities $\mathrm{Var}(\mathbf{x}'\hat{\boldsymbol{\beta}})$, $\mathrm{Cov}(\mathbf{x}'\hat{\boldsymbol{\beta}}, \hat{\sigma})$, and $\mathrm{Var}(\hat{\sigma})$ are computed based on the covariance matrix of the estimated parameter vector $(\hat{\boldsymbol{\beta}}, \hat{\sigma})$.

### Confidence Limits for Percentiles

If the HCL option is specified in the PROBPLOT statement, confidence limits based on estimated distribution percentiles instead of the default CDF limits are displayed in the probability plot.

For location-scale distributions, the estimated $p \times 100\%$ percentile of the distribution $F$ is given by

$$y_p = \mathbf{x}'\hat{\boldsymbol{\beta}} + G^{-1}(p)\hat{\sigma}$$

where $G$ is the baseline distribution and the column vector $\mathbf{x}$ of covariate values is determined by the rules summarized in the section "XDATA= Data Set" on page 3947. The standard error of $y_p$ is estimated by $\mathrm{SE}_y = z'\Sigma z$ where $\mathbf{z} = (\mathbf{x}', G^{-1}(p))'$ and $\Sigma$ is the covariance matrix of the parameter estimates $(\hat{\boldsymbol{\beta}}', \hat{\sigma})'$. Two-sided $(1 - \alpha) \times 100\%$ confidence limits for $y_p$ are given by

$$[y_L, \ y_U] = [y_p - z_{1-\alpha/2}\mathrm{SE}_y, \ y_p + z_{1-\alpha/2}\mathrm{SE}_y]$$

For distributions of type log-location-scale, the confidence limits are computed as

$$[t_L = \exp(y_L), \ t_U = \exp(y_U)]$$

For example, if $T$ has the Weibull distribution, $G$ is the standardized extreme value distribution, $[y_L, \ y_U]$ are confidence limits for the $p \times 100\%$ percentile of the extreme value distribution for $\log(T)$, and $[t_L = \exp(y_L), \ t_U = \exp(y_U)]$ are confidence limits for the $p \times 100\%$ percentile of the Weibull distribution for $T$.

## INEST= Data Set

If specified, the INEST= data set specifies initial estimates for all the parameters in the model. The INEST= data set must contain the intercept variable (named Intercept) and all independent variables in the MODEL statement.

If BY processing is used, the INEST= data set should also include the BY variables, and there must be at least one observation for each BY group. If there is more than one observation in one BY group, the first observation read is used for that BY group.

If the INEST= data set also contains the _TYPE_ variable, only observations with _TYPE_ value 'PARMS' are used as starting values. Combining the INEST= data set and the MAXITER= option in the MODEL

statement, partial scoring can be done, such as predicting on a validation data set by using the model built from a training data set.

You can specify starting values for the iterative algorithm in the INEST= data set. This data set overwrites the INITIAL= option in the MODEL statement, which is a little difficult to use for models including multilevel interaction effects. The INEST= data set has the same structure as the OUTEST= data set but is not required to have all the variables or observations that appear in the OUTEST= data set. One simple use of the INEST= option is passing the previous OUTEST= data set directly to the next model as an INEST= data set, assuming that the two models have the same parameterization. See Example 51.3 for an illustration.

## OUTEST= Data Set

The OUTEST= data set contains parameter estimates and the log likelihood for the model. You can specify a label in the MODEL statement to distinguish between the estimates for different models fit with the LIFEREG procedure. If the COVOUT option is specified, the OUTEST= data set also contains the estimated covariance matrix of the parameter estimates. Note that, if the LIFEREG procedure does not converge, the parameter estimates are set to missing in the OUTEST data set.

The OUTEST= data set contains all variables specified in the MODEL statement and the BY statement. One observation consists of parameter values for the model with the dependent variable having the value –1. If the COVOUT option is specified, there are additional observations containing the rows of the estimated covariance matrix. For these observations, the dependent variable contains the parameter estimate for the corresponding row variable. The following variables are also added to the data set:

| | |
|---|---|
| _MODEL_ | a character variable containing the label of the MODEL statement, if present. Otherwise, the variable's value is blank. |
| _NAME_ | a character variable containing the name of the dependent variable for the parameter estimates observations or the name of the row for the covariance matrix estimates |
| _TYPE_ | a character variable containing the type of the observation, either PARMS for parameter estimates or COV for covariance estimates |
| _DIST_ | a character variable containing the name of the distribution modeled |
| _LNLIKE_ | a numeric variable containing the last computed value of the log likelihood |
| INTERCEPT | a numeric variable containing the intercept parameter estimates and covariances |
| _SCALE_ | a numeric variable containing the scale parameter estimates and covariances |
| _SHAPE1_ | a numeric variable containing the first shape parameter estimates and covariances if the specified distribution has additional shape parameters |

Any BY variables specified are also added to the OUTEST= data set.

## XDATA= Data Set

The XDATA= data set is used for plotting the predicted probability when there are covariates specified in a MODEL statement and a probability plot is specified with a PROBPLOT statement. See Example 51.4 for an illustration.

The XDATA= data set is an input SAS data set that contains values for all the independent variables in the MODEL statement and variables in the CLASS statement. The XDATA= data set has the same structure as the DATA= data set but is not required to have all the variables or observations that appear in the DATA= data set.

The XDATA= data set must contain all the independent variables in the MODEL statement and variables in the CLASS statement. Even though variables in the CLASS statement might not be used, valid values are required for these variables in the XDATA= data set. Missing values are not allowed. Missing values are not allowed in the XDATA= data set for any of the independent variables, either. Missing values are allowed for the dependent variables and other variables if they are included in the XDATA= data set.

If BY processing is used, the XDATA= data set should also include the BY variables, and there must be at least one valid observation for each BY group. If there is more than one valid observation in a BY group, the last one read is used for that BY group.

If there is no XDATA= data set in the PROC LIFEREG statement, by default, the LIFEREG procedure will use the overall mean for effects containing a continuous variable (or variables) and the highest level of a single classification variable as reference level. The rules are summarized as follows:

- If the effect contains a continuous variable (or variables), the overall mean of this effect (not the variables) is used.

- If the effect is a single classification variable, the highest level of the variable is used.

## Computational Resources

Let $p$ be the number of parameters estimated in the model. The minimum working space (in bytes) needed is

$$16p^2 + 100p$$

However, if sufficient space is available, the input data set is also kept in memory; otherwise, the input data set is reread for each evaluation of the likelihood function and its derivatives, with the resulting execution time of the procedure substantially increased.

Let $n$ be the number of observations used in the model estimation. Each evaluation of the likelihood function and its first and second derivatives requires $O(np^2)$ multiplications and additions, $n$ individual function evaluations for the log density or log distribution function, and $n$ evaluations of the first and second derivatives of the function. The calculation of each updating step from the gradient and Hessian requires $O(p^3)$ multiplications and additions. The $O(v)$ notation means that, for large values of the argument, $v$, $O(v)$ is approximately a constant times $v$.

# Bayesian Analysis

## Gibbs Sampling

This section provides details about Bayesian analysis by Gibbs sampling in the location-scale models for survival data available in PROC LIFEREG. See the section "Gibbs Sampler" on page 135 for a general discussion of Gibbs sampling. PROC LIFEREG fits parametric location-scale survival models. That is, the probability density of the response $Y$ can expressed in the general form

$$f(y) = g\left(\frac{y - \mu}{\sigma}\right)$$

where $Y = \log(T)$ for lifetimes $T$. The function $g$ determines the specific distribution. The location parameter $\mu_i$ is modeled through regression parameters as $\mu_i = x_i'\beta$. The LIFEREG procedure can provide Bayesian estimates of the regression parameters and $\sigma$. The OUTPUT and PROBPLOT statements, if specified, are ignored. The PLOTS=PROBPLOT option in the PROC LIFEREG statement and the CORRB and COVB options in the MODEL statement are also ignored.

For the Weibull distribution, you can specify that Gibbs sampling be performed on the Weibull shape parameter $\beta = \sigma^{-1}$ instead of the scale parameter $\sigma$ by specifying a prior distribution for the shape parameter with the WEIBULLSHAPEPRIOR= option. In addition, if there are no covariates in the model, you can specify Gibbs sampling on the Weibull scale parameter $\alpha = \exp(\mu)$, where $\mu$ is the intercept term, with the WEIBULLSCALEPRIOR= option.

In the case of the exponential distribution with no covariates, you can specify Gibbs sampling on the exponential scale parameter $\alpha = \exp(\mu)$, where $\mu$ is the intercept term, with the EXPSCALEPRIOR= option.

Let $\theta = (\theta_1, \ldots, \theta_k)'$ be the parameter vector. For location-scale models, the $\theta_i$'s are the regression coefficients $\beta_i$'s and the scale parameter $\sigma$. In the case of the three-parameter gamma distribution, there is an additional gamma shape parameter $\tau$. Let $L(D|\theta)$ be the likelihood function, where $D$ is the observed data. Let $\pi(\theta)$ be the prior distribution. The full conditional distribution of $[\theta_i|\theta_j, i \neq j]$ is proportional to the joint distribution; that is,

$$\pi(\theta_i|\theta_j, i \neq j, D) \propto L(D|\theta)p(\theta)$$

For instance, the one-dimensional conditional distribution of $\theta_1$ given $\theta_j = \theta_j^*, 2 \leq j \leq k$, is computed as

$$\pi(\theta_1|\theta_j = \theta_j^*, 2 \leq j \leq k, D) = L(D|(\theta = (\theta_1, \theta_2^*, \ldots, \theta_k^*)')p(\theta = (\theta_1, \theta_2^*, \ldots, \theta_k^*)')$$

Suppose you have a set of arbitrary starting values $\{\theta_1^{(0)}, \ldots, \theta_k^{(0)}\}$. Using the ARMS (adaptive rejection Metropolis sampling) algorithm of Gilks and Wild (1992) and Gilks, Best, and Tan (1995), you can do the following:

draw $\theta_1^{(1)}$ from $[\theta_1|\theta_2^{(0)}, \ldots, \theta_k^{(0)}]$

draw $\theta_2^{(1)}$ from $[\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)}]$

. . .

draw $\theta_k^{(1)}$ from $[\theta_k|\theta_1^{(1)}, \ldots, \theta_{k-1}^{(1)}]$

This completes one iteration of the Gibbs sampler. After one iteration, you have $\{\theta_1^{(1)}, \ldots, \theta_k^{(1)}\}$. After $n$ iterations, you have $\{\theta_1^{(n)}, \ldots, \theta_k^{(n)}\}$. PROC LIFEREG implements the ARMS algorithm based on a program provided by Gilks (2003) to draw a sample from a full conditional distribution. See the section "Assessing Markov Chain Convergence" on page 139 for information about assessing the convergence of the chain of posterior samples.

You can output these posterior samples into a SAS data set. The following option in the BAYES statement outputs the posterior samples into the SAS data set Post: OUTPOST=Post. The data set also includes the variables LogPost and LogLike, which represent the log of the posterior distribution and the log of the likelihood, respectively.

## Priors for Model Parameters

The model parameters are the regression coefficients and the dispersion parameter (or the precision or scale), if the model has one. The priors for the dispersion parameter and the priors for the regression coefficients are assumed to be independent, while you can have a joint multivariate normal prior for the regression coefficients.

### Scale and Shape Parameters

**Gamma Prior**   The gamma distribution $G(a, b)$ has a PDF

$$f_{a,b}(u) = \frac{b(bu)^{a-1} e^{-bu}}{\Gamma(a)}, \qquad u > 0$$

where $a$ is the shape parameter and $b$ is the inverse-scale parameter. The mean is $\frac{a}{b}$ and the variance is $\frac{a}{b^2}$.

**Improper Prior**   The joint prior density is given by

$$p(u) \propto u^{-1}, \qquad u > 0$$

### Regression Coefficients

Let $\boldsymbol{\beta}$ be the regression coefficients.

**Normal Prior**   Assume $\boldsymbol{\beta}$ has a multivariate normal prior with mean vector $\boldsymbol{\beta}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$. The joint prior density is given by

$$p(\boldsymbol{\beta}) \propto e^{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)' \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)}$$

**Uniform Prior**   The joint prior density is given by

$$p(\boldsymbol{\beta}) \propto 1$$

## Posterior Distribution

Denote the observed data by $D$.

The posterior distribution is

$$\pi(\boldsymbol{\theta}|D) \propto L_P(D|\boldsymbol{\theta}) p(\boldsymbol{\theta})$$

where $L_P(D|\boldsymbol{\theta})$ is the likelihood function with regression coefficients and any additional parameters, such as scale or shape, $\boldsymbol{\theta}$ as parameters; and $p(\boldsymbol{\theta})$ is the joint prior distribution of the parameters.

## Deviance Information Criterion

Let $\boldsymbol{\theta}_i$ be the model parameters at iteration $i$ of the Gibbs sampler, and let $LL(\boldsymbol{\theta}_i)$ be the corresponding model log likelihood. PROC LIFEREG computes the following fit statistics defined by Spiegelhalter et al. (2002):

- effective number of parameters:

$$p_D = \overline{LL(\boldsymbol{\theta})} - LL(\bar{\boldsymbol{\theta}})$$

- deviance information criterion (DIC):

$$DIC = \overline{LL(\boldsymbol{\theta})} + p_D$$

where

$$\overline{LL(\boldsymbol{\theta})} = \frac{1}{n}\sum_{i=1}^{n} LL(\boldsymbol{\theta}_i)$$

$$\bar{\boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\theta}_i$$

and $n$ is the number of Gibbs samples.

## Starting Values of the Markov Chains

When the BAYES statement is specified, PROC LIFEREG generates one Markov chain containing the approximate posterior samples of the model parameters. Additional chains are produced when the Gelman-Rubin diagnostics are requested. Starting values (or initial values) can be specified in the INITIAL= data set in the BAYES statement. If INITIAL= option is not specified, PROC LIFEREG picks its own initial values for the chains.

Denote $[x]$ as the integral value of x. Denote $\hat{s}(X)$ as the estimated standard error of the estimator $X$.

### Regression Coefficients and Gamma Shape Parameter

For the first chain that the summary statistics and regression diagnostics are based on, the default initial values are estimates of the mode of the posterior distribution. If the INITIALMLE option is specified, the initial values are the maximum likelihood estimates; that is,

$$\beta_i^{(0)} = \hat{\beta}_i$$

Initial values for the $r$th chain ($r \geq 2$) are given by

$$\beta_i^{(0)} = \hat{\beta}_i \pm \left(2 + \left[\frac{r}{2}\right]\right)\hat{s}(\hat{\beta}_i)$$

with the plus sign for odd $r$ and minus sign for even $r$.

### *Scale, Exponential Scale, Weibull Scale, or Weibull Shape Parameter* $\lambda$

Let $\lambda$ be the parameter sampled.

For the first chain that the summary statistics and diagnostics are based on, the initial values are estimates of the mode of the posterior distribution; or the maximum likelihood estimates if the INITIALMLE option is specified; that is,

$$\lambda^{(0)} = \hat{\lambda}$$

The initial values of the $r$th chain ($r \geq 2$) are given by

$$\lambda^{(0)} = \hat{\lambda} e^{\pm\left([\frac{r}{2}]+2\right)\hat{s}(\hat{\lambda})}$$

with the plus sign for odd $r$ and minus sign for even $r$.

## OUTPOST= Output Data Set

The OUTPOST= data set contains the generated posterior samples. There are 2+$n$ variables, where $n$ is the number of model parameters. The variable Iteration represents the iteration number and the variable LogPost contains the log posterior likelihood values. The other $n$ variables represent the draws of the Markov chain for the model parameters.

# Displayed Output for Classical Analysis

For each model, PROC LIFEREG displays the following.

## Model Information

The "Model Information" table displays the two-level name of the input data set, the distribution name, and the name and label of the dependent variable; the name and label of the censor indicator variable, for right-censored data; if you specify the WEIGHT statement, the name and label of the weight variable; and the maximum value of the log likelihood.

## Number of Observations

The "Number of Observations" table displays the number of observations read from the input data set, and the number of observations used in the analysis.

## Class Level Information

The "Class Level Information" table displays the levels of classification variables if you specify a CLASS statement.

## Fit Statistics

The "Fit Statistics" table displays the negative of twice the log likelihood, Akaike's information criterion (AIC), the corrected Akaike's information criterion (AICC), and the Bayesian information criterion (BIC). If the specified distribution is Weibull, lognormal, log-logistic, or gamma, the fit criteria are based on the log likelihood for the log of the response, rather than for the response on the original scale.

### Fit Statistics (Unlogged Response)

If the specified distribution is Weibull, lognormal, log-logistic, or gamma, the "Fit Statistics (Unlogged Response)" table displays fit criteria that are based on the log likelihood for the response on the original, rather than log, scale. The negative of twice the log likelihood, Akaike's information criterion (AIC), the corrected Akaike's information criterion (AICC), and the Bayesian information criterion (BIC) are displayed.

### Type III Analysis of Effects

The "Type III Analysis of Effects" table displays, for each effect in the model, the effect name, the degrees of freedom associated with the type III contrast for the effect, the chi-square statistic for the contrast, and the *p*-value for the statistic.

### Analysis of Maximum Likelihood Parameter Estimates

The "Analysis of Maximum Likelihood Parameter Estimates" table displays the parameter name, the degrees of freedom for each parameter, the maximum likelihood estimate of each parameter, the estimated standard error of the parameter estimator, confidence limits for each parameter, a chi-square statistic for testing whether the parameter is zero, and the associated *p*-value for the statistic.

### Lagrange Multiplier Statistics

If there are constrained parameters in the model, such as the scale or intercept, then the "Lagrange Multiplier Statistics" table displays a Lagrange multiplier test for the constraint.

---

## Displayed Output for Bayesian Analysis

If a Bayesian analysis is requested with a BAYES statement, the displayed output includes the following.

### Model Information

The "Model Information" table displays the two-level name of the input data set, the number of burn-in iterations, the number of iterations after the burn-in, the number of thinning iterations, the distribution name, and the name and label of the dependent variable; the name and label of the censor indicator variable, for right-censored data; if you specify the WEIGHT statement, the name and label of the weight variable; and the maximum value of the log likelihood.

### Class Level Information

The "Class Level Information" table displays the levels of classification variables if you specify a CLASS statement.

### Maximum Likelihood Estimates

The "Analysis of Maximum Likelihood Parameter Estimates" table displays the maximum likelihood estimate of each parameter, the estimated standard error of the parameter estimator, and confidence limits for each parameter.

## Coefficient Prior

The "Coefficient Prior" table displays the prior distribution of the regression coefficients.

## Independent Prior Distributions for Model Parameters

The "Independent Prior Distributions for Model Parameters" table displays the prior distributions of additional model parameters (scale, exponential scale, Weibull scale, Weibull shape, gamma shape).

## Initial Values and Seeds

The "Initial Values and Seeds" table displays the initial values and random number generator seeds for the Gibbs chains.

## Fit Statistics

The "Fit Statistics" table displays the deviance information criterion (DIC) and the effective number of parameters.

## Posterior Summaries

The "Posterior Summaries" table contains the size of the sample, the mean, the standard deviation, and the quartiles for each model parameter.

## Posterior Intervals

The "Posterior Intervals" table contains the HPD intervals and the credible intervals for each model parameter.

## Correlation Matrix of the Posterior Samples

The "Correlation Matrix of the Posterior Samples" table is produced if you include the CORR suboption in the SUMMARY= option in the BAYES statement. This table displays the sample correlation of the posterior samples.

## Covariance Matrix of the Posterior Samples

The "Covariance Matrix of the Posterior Samples" table is produced if you include the COV suboption in the SUMMARY= option in the BAYES statement. This table displays the sample covariance of the posterior samples.

## Autocorrelations of the Posterior Samples

The "Autocorrelations of the Posterior Samples" table displays the lag1, lag5, lag10, and lag50 autocorrelations for each parameter.

## Gelman and Rubin Diagnostics

The "Gelman and Rubin Diagnostics" table is produced if you include the GELMAN suboption in the DIAGNOSTIC= option in the BAYES statement. This table displays the estimate of the potential scale reduction factor and its 97.5% upper confidence limit for each parameter.

## Geweke Diagnostics

The "Geweke Diagnostics" table displays the Geweke statistic and its *p*-value for each parameter.

## Raftery and Lewis Diagnostics

The "Raftery Diagnostics" tables is produced if you include the RAFTERY suboption in the DIAGNOS-TIC= option in the BAYES statement. This table displays the Raftery and Lewis diagnostics for each variable.

## Heidelberger and Welch Diagnostics

The "Heidelberger and Welch Diagnostics" table is displayed if you include the HEIDELBERGER suboption in the DIAGNOSTIC= option in the BAYES statement. This table shows the results of a stationary test and a halfwidth test for each parameter.

## Effective Sample Size

The "Effective Sample Size" table displays, for each parameter, the effective sample size, the correlation time, and the efficiency.

## Monte Carlo Standard Errors

The "Monte Carlo Standard Errors" table displays, for each parameter, the Monte Carlo standard error, the posterior sample standard deviation, and the ratio of the two.

## ODS Table Names

PROC LIFEREG assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed separately in Table 51.14 for a maximum likelihood analysis and in Table 51.15 for a Bayesian analysis. For more information about ODS, see Chapter 20, "Using the Output Delivery System."

**Table 51.14** ODS Tables Produced in PROC LIFEREG for a Classical Analysis

| ODS Table Name | Description | Statement | Option |
|---|---|---|---|
| ClassLevels | Classification variable levels | CLASS | Default* |
| ConvergenceStatus | Convergence status | MODEL | Default |
| CorrB | Parameter estimate correlation matrix | MODEL | CORRB |
| CovB | Parameter estimate covariance matrix | MODEL | COVB |
| IterEM | Iteration history for Turnbull algorithm | PROBPLOT | ITPRINTEM |
| FitStatistics | Fit statistics | MODEL | Default |
| FitStatisticsUL | Fit statistics for unlogged response | MODEL | DISTRIBUTION=WEIBULL, LOGNORMAL, LLO-GISTIC, or GAMMA |
| IterHistory | Iteration history | MODEL | ITPRINT |
| LagrangeStatistics | Lagrange statistics | MODEL | NOINT \| NOSCALE |
| LastGrad | Last evaluation of the gradient | MODEL | ITPRINT |

**Table 51.14**   *continued*

| ODS Table Name | Description | Statement | Option |
|---|---|---|---|
| LastHess | Last evaluation of the Hessian | MODEL | ITPRINT |
| ModelInfo | Model information | MODEL | Default |
| NObs | Number of observations | MODEL | Default |
| ParameterEstimates | Parameter estimates | MODEL | Default |
| ParmInfo | Parameter indices | MODEL | Default |
| ProbabilityEstimates | Nonparametric CDF estimates | PROBPLOT | PPOUT |
| TConvergenceStatus | Convergence status for Turnbull algorithm | PROBPLOT | Default |
| Turnbull | Probability estimates from Turnbull algorithm | PROBPLOT | ITPRINTEM |
| Type3Analysis | Type 3 tests | MODEL | Default* |

* Depending on the data.

**Table 51.15**   ODS Tables Produced in PROC LIFEREG for a Bayesian Analysis

| ODS Table Name | Description | Statement | Option |
|---|---|---|---|
| AutoCorr | Autocorrelations of the posterior samples | BAYES | Default |
| ClassLevels | Classification variable levels | CLASS | Default* |
| CoeffPrior | Prior distribution of the regression coefficients | BAYES | Default |
| ConvergenceStatus | Convergence status of maximum likelihood estimation | MODEL | Default |
| Corr | Correlation matrix of the posterior samples | BAYES | SUMMARY=CORR |
| ESS | Effective sample size | BAYES | Default |
| FitStatistics | Fit statistics | BAYES | Default |
| Gelman | Gelman and Rubin convergence diagnostics | BAYES | DIAG=GELMAN |
| Geweke | Geweke convergence diagnostics | BAYES | Default |
| Heidelberger | Heidelberger and Welch convergence diagnostics | BAYES | DIAG=HEIDELBERGER |
| InitialValues | Initial values of the Markov chains | BAYES | Default |
| MCError | Monte Carlo standard errors | BAYES | DIAG=MCSE |
| ModelInfo | Model information | MODEL | Default |
| NObs | Number of observations | MODEL | Default |
| ParameterEstimates | Maximum likelihood estimates of model parameters | MODEL | Default |
| ParmPrior | Prior distribution for scale and shape | BAYES | Default |
| PostIntervals | HPD and equal-tail intervals of the posterior samples | BAYES | Default |
| PosteriorSample | Posterior samples (for output data set only) | BAYES | |

**Table 51.15** *continued*

| ODS Table Name | Description | Statement | Option |
|---|---|---|---|
| PostSummaries | Summary statistics of the posterior samples | BAYES | Default |
| Raftery | Raftery and Lewis convergence diagnostics | BAYES | DIAG=RAFTERY |

* Depending on the data.

## ODS Graphics

Statistical procedures use ODS Graphics to create graphs as part of their output. ODS Graphics is described in detail in Chapter 21, "Statistical Graphics Using ODS."

Before you create graphs, ODS Graphics must be enabled (for example, by specifying the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 600 in Chapter 21, "Statistical Graphics Using ODS."

The overall appearance of graphs is controlled by ODS styles. Styles and other aspects of using ODS Graphics are discussed in the section "A Primer on ODS Statistical Graphics" on page 599 in Chapter 21, "Statistical Graphics Using ODS."

Some graphs are produced by default; other graphs are produced by using statements and options.

### ODS Graph Names

PROC LIFEREG assigns a name to each graph it creates using ODS. You can use these names to reference the graphs when using ODS. The names of the graphs that PROC LIFEREG generates are listed in Table 51.16, along with the required statements and options.

**Table 51.16** Graphs Produced by PROC LIFEREG

| ODS Graph Name | Description | Statement | Option |
|---|---|---|---|
| ADPanel | Autocorrelation function and density panel | BAYES | PLOTS=(AUTOCORR DENSITY) |
| AutocorrPanel | Autocorrelation function panel | BAYES | PLOTS= AUTOCORR |
| AutocorrPlot | Autocorrelation function plot | BAYES | PLOTS(UNPACK)=AUTOCORR |
| ProbPlot | Probability plot | PROBPLOT | Default |
| TAPanel | Trace and autocorrelation function panel | BAYES | PLOTS=(TRACE AUTOCORR) |
| TADPanel | Trace, autocorrelation, and density function panel | BAYES | Default |
| TDPanel | Trace and density panel | BAYES | PLOTS=(TRACE DENSITY) |
| TracePanel | Trace panel | BAYES | PLOTS=TRACE |
| TracePlot | Trace plot | BAYES | PLOTS(UNPACK)=TRACE |

# Examples: LIFEREG Procedure

## Example 51.1: Motorette Failure

This example fits a Weibull model and a lognormal model to the example given in Kalbfleisch and Prentice (1980, p. 5). An output data set called models is specified to contain the parameter estimates. By default, the natural log of the variable time is used by the procedure as the response. After this log transformation, the Weibull model is fit using the extreme-value baseline distribution, and the lognormal is fit using the normal baseline distribution.

Since the extreme-value and normal distributions do not contain any shape parameters, the variable SHAPE1 is missing in the models data set. An additional output data set, out, is created that contains the predicted quantiles and their standard errors for values of the covariate corresponding to temp=130 and temp=150. This is done with the control variable, which is set to 1 for only two observations.

Using the standard error estimates obtained from the output data set, approximate 90% confidence limits for the predicted quantities are then created in a subsequent DATA step for the log response. The logs of the predicted values are obtained because the values of the P= variable in the OUT= data set are in the same units as the original response variable, time. The standard errors of the quantiles of log(time) are approximated (using a Taylor series approximation) by the standard deviation of time divided by the mean value of time. These confidence limits are then converted back to the original scale by the exponential function.

The following statements produce Output 51.1.1:

```
title 'Motorette Failures With Operating Temperature as a Covariate';
data motors;
   input time censor temp @@;
   if _N_=1 then
      do;
         temp=130;
         time=.;
         control=1;
         z=1000/(273.2+temp);
         output;
         temp=150;
         time=.;
         control=1;
         z=1000/(273.2+temp);
         output;
      end;
   if temp>150;
   control=0;
   z=1000/(273.2+temp);
   output;
   datalines;
 8064 0 150 8064 0 150 8064 0 150 8064 0 150 8064 0 150
 8064 0 150 8064 0 150 8064 0 150 8064 0 150 8064 0 150
 1764 1 170 2772 1 170 3444 1 170 3542 1 170 3780 1 170
 4860 1 170 5196 1 170 5448 0 170 5448 0 170 5448 0 170
  408 1 190  408 1 190 1344 1 190 1344 1 190 1440 1 190
```

```
  1680 0 190 1680 0 190 1680 0 190 1680 0 190 1680 0 190
   408 1 220  408 1 220  504 1 220  504 1 220  504 1 220
   528 0 220  528 0 220  528 0 220  528 0 220  528 0 220
 ;

 proc print data=motors;
 run;
```

**Output 51.1.1** Motorette Failure Data

```
          Motorette Failures With Operating Temperature as a Covariate

              Obs    time    censor    temp    control      z

               1       .        0       130       1       2.48016
               2       .        0       150       1       2.36295
               3     1764       1       170       0       2.25632
               4     2772       1       170       0       2.25632
               5     3444       1       170       0       2.25632
               6     3542       1       170       0       2.25632
               7     3780       1       170       0       2.25632
               8     4860       1       170       0       2.25632
               9     5196       1       170       0       2.25632
              10     5448       0       170       0       2.25632
              11     5448       0       170       0       2.25632
              12     5448       0       170       0       2.25632
              13      408       1       190       0       2.15889
              14      408       1       190       0       2.15889
              15     1344       1       190       0       2.15889
              16     1344       1       190       0       2.15889
              17     1440       1       190       0       2.15889
              18     1680       0       190       0       2.15889
              19     1680       0       190       0       2.15889
              20     1680       0       190       0       2.15889
              21     1680       0       190       0       2.15889
              22     1680       0       190       0       2.15889
              23      408       1       220       0       2.02758
              24      408       1       220       0       2.02758
              25      504       1       220       0       2.02758
              26      504       1       220       0       2.02758
              27      504       1       220       0       2.02758
              28      528       0       220       0       2.02758
              29      528       0       220       0       2.02758
              30      528       0       220       0       2.02758
              31      528       0       220       0       2.02758
              32      528       0       220       0       2.02758
```

The following statements produce Output 51.1.2 and Output 51.1.3:

```
proc lifereg data=motors outest=modela covout;
   a: model time*censor(0)=z;
      output out=outa quantiles=.1 .5 .9 std=std p=predtime
         control=control;
run;

proc lifereg data=motors outest=modelb covout;
   b: model time*censor(0)=z / dist=lnormal;
      output out=outb quantiles=.1 .5 .9 std=std p=predtime
         control=control;
run;
```

**Output 51.1.2** Motorette Failure: Model A

```
        Motorette Failures With Operating Temperature as a Covariate

                          The LIFEREG Procedure

                           Model Information

                Data Set                        WORK.MOTORS
                Dependent Variable               Log(time)
                Censoring Variable                  censor
                Censoring Value(s)                       0
                Number of Observations                  30
                Noncensored Values                      17
                Right Censored Values                   13
                Left Censored Values                     0
                Interval Censored Values                 0
                Number of Parameters                     3
                Name of Distribution               Weibull
                Log Likelihood                -22.95148315


                     Type III Analysis of Effects

                                      Wald
                Effect          DF   Chi-Square    Pr > ChiSq

                  z              1     99.5239       <.0001


             Analysis of Maximum Likelihood Parameter Estimates

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

       Intercept     1 -11.8912  1.9655 -15.7435  -8.0389  36.60   <.0001
       z             1   9.0383  0.9060   7.2626  10.8141  99.52   <.0001
       Scale         1   0.3613  0.0795   0.2347   0.5561
       Weibull Shape 1   2.7679  0.6091   1.7982   4.2605
```

**Output 51.1.3** Motorette Failure: Model B

```
         Motorette Failures With Operating Temperature as a Covariate

                          The LIFEREG Procedure

                            Model Information

                Data Set                      WORK.MOTORS
                Dependent Variable            Log(time)
                Censoring Variable               censor
                Censoring Value(s)                    0
                Number of Observations               30
                Noncensored Values                   17
                Right Censored Values                13
                Left Censored Values                  0
                Interval Censored Values              0
                Number of Parameters                  3
                Name of Distribution          Lognormal
                Log Likelihood            -24.47381031


                      Type III Analysis of Effects

                                        Wald
                Effect           DF   Chi-Square    Pr > ChiSq

                   z             1      42.0001        <.0001


            Analysis of Maximum Likelihood Parameter Estimates

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

      Intercept      1 -10.4706  2.7719 -15.9034  -5.0377  14.27    0.0002
      z              1   8.3221  1.2841   5.8052  10.8389  42.00    <.0001
      Scale          1   0.6040  0.1107   0.4217   0.8652
```

The following statements produce Output 51.1.4:

```
data models;
   set modela modelb;
run;

proc print data=models;
   id _model_;
   title 'Fitted Models';
run;
```

**Output 51.1.4** Motorette Failure: Fitted Models

```
                            Fitted Models

      _MODEL_      _NAME_       _TYPE_     _DIST_       _STATUS_        _LNLIKE_

         a         time         PARMS      Weibull      0 Converged     -22.9515
         a         Intercept    COV        Weibull      0 Converged     -22.9515
         a         z            COV        Weibull      0 Converged     -22.9515
         a         Scale        COV        Weibull      0 Converged     -22.9515
         b         time         PARMS      Lognormal    0 Converged     -24.4738
         b         Intercept    COV        Lognormal    0 Converged     -24.4738
         b         z            COV        Lognormal    0 Converged     -24.4738
         b         Scale        COV        Lognormal    0 Converged     -24.4738

      _MODEL_         time      Intercept        z           _SCALE_

         a          -1.0000     -11.8912      9.03834        0.36128
         a         -11.8912       3.8632     -1.77878        0.03448
         a           9.0383      -1.7788      0.82082       -0.01488
         a           0.3613       0.0345     -0.01488        0.00632
         b          -1.0000     -10.4706      8.32208        0.60403
         b         -10.4706       7.6835     -3.55566        0.03267
         b           8.3221      -3.5557      1.64897       -0.01285
         b           0.6040       0.0327     -0.01285        0.01226
```

The following statements produce Output 51.1.5:

```
data out;
   set outa outb;
run;

data out1;
   set out;
   ltime=log(predtime);
   stde=std/predtime;
   upper=exp(ltime+1.64*stde);
   lower=exp(ltime-1.64*stde);
run;

title 'Quantile Estimates and Confidence Limits';
proc print data=out1;
   id temp;
run;
title;
```

**Output 51.1.5** Motorette Failure: Quantile Estimates and Confidence Limits

```
                Quantile Estimates and Confidence Limits

                                P
            c                   r
          c  o              _   e
          e  n          P   d              l              u          l
    t  t  n  t          R   t          t       s          p          o
    e  i  s  r          O   i      s   i       t          p          w
    m  m  o  o          B   m      t   m       d          e          e
    p  e  r  l   z      _   e      d   e       e          r          r

   130 . 0 1 2.48016 0.1 16519.27  5999.85  9.7123 0.36320  29969.51  9105.47
   130 . 0 1 2.48016 0.5 32626.65  9874.33 10.3929 0.30265  53595.71 19861.63
   130 . 0 1 2.48016 0.9 50343.22 15044.35 10.8266 0.29884  82183.49 30838.80
   150 . 0 1 2.36295 0.1  5726.74  1569.34  8.6529 0.27404   8976.12  3653.64
   150 . 0 1 2.36295 0.5 11310.68  2299.92  9.3335 0.20334  15787.62  8103.28
   150 . 0 1 2.36295 0.9 17452.49  3629.28  9.7672 0.20795  24545.37 12409.24
   130 . 0 1 2.48016 0.1 12033.19  5482.34  9.3954 0.45560  25402.68  5700.09
   130 . 0 1 2.48016 0.5 26095.68 11359.45 10.1695 0.43530  53285.36 12779.95
   130 . 0 1 2.48016 0.9 56592.19 26036.90 10.9436 0.46008 120349.65 26611.42
   150 . 0 1 2.36295 0.1  4536.88  1443.07  8.4200 0.31808   7643.71  2692.83
   150 . 0 1 2.36295 0.5  9838.86  2901.15  9.1941 0.29487  15957.38  6066.36
   150 . 0 1 2.36295 0.9 21336.97  7172.34  9.9682 0.33615  37029.72 12294.62
```

## Example 51.2: Computing Predicted Values for a Tobit Model

The LIFEREG procedure can be used to perform a Tobit analysis. The Tobit model, described by Tobin (1958), is a regression model for left-censored data assuming a normally distributed error term. The model parameters are estimated by maximum likelihood. PROC LIFEREG provides estimates of the parameters of the distribution of the *uncensored* data. See Greene (1993) and Maddala (1983) for a more complete discussion of censored normal data and related distributions. This example shows how you can use PROC LIFEREG and the DATA step to compute two of the three types of predicted values discussed there.

Consider a continuous random variable Y and a constant C. If you were to sample from the distribution of Y but discard values less than (greater than) C, the distribution of the remaining observations would be *truncated* on the left (right). If you were to sample from the distribution of Y and report values less than (greater than) C as C, the distribution of the sample would be left (right) *censored*.

The probability density function of the truncated random variable $Y'$ is given by

$$f_{Y'}(y) = \frac{f_Y(y)}{\Pr(Y > C)} \quad \text{for} \quad y > C$$

where $f_Y(y)$ is the probability density function of Y. PROC LIFEREG cannot compute the proper likelihood function to estimate parameters or predicted values for a truncated distribution. Suppose the model being fit is specified as follows:

$$Y_i^* = x_i'\beta + \epsilon_i$$

where $\epsilon_i$ is a normal error term with zero mean and standard deviation $\sigma$.

Define the censored random variable $Y_i$ as

$$Y_i \quad = \quad 0 \ \text{if} \ Y_i^* \leq 0$$
$$Y_i \quad = \quad Y_i^* \ \text{if} \ Y_i^* > 0$$

This is the Tobit model for left-censored normal data. $Y_i^*$ is sometimes called the *latent variable*. PROC LIFEREG estimates parameters of the distribution of $Y_i^*$ by maximum likelihood.

You can use the LIFEREG procedure to compute predicted values based on the mean functions of the latent and observed variables. The mean of the latent variable $Y_i^*$ is $x_i'\beta$, and you can compute values of the mean for different settings of $x_i$ by specifying XBETA=*variable-name* in an OUTPUT statement. Estimates of $x_i'\beta$ for each observation will be written to the OUT= data set. Predicted values of the observed variable $Y_i$ can be computed based on the mean

$$E(Y_i) = \Phi\left(\frac{x_i'\beta}{\sigma}\right)(x_i'\beta + \sigma\lambda_i)$$

where

$$\lambda_i = \frac{\phi(x_i'\beta/\sigma)}{\Phi(x_i'\beta/\sigma)}$$

$\phi$ and $\Phi$ represent the normal probability density and cumulative distribution functions.

Although the distribution of $\epsilon_i$ in the Tobit model is often assumed normal, you can use other distributions for the Tobit model in the LIFEREG procedure by specifying a distribution with the DISTRIBUTION= option in the MODEL statement. One distribution that should be mentioned is the logistic distribution. For this distribution, the MLE has bounded influence function with respect to the response variable, but not the design variables. If you believe your data have outliers in the response direction, you might try this distribution for some robust estimation of the Tobit model.

With the logistic distribution, the predicted values of the observed variable $Y_i$ can be computed based on the mean of $Y_i^*$,

$$E(Y_i) = \sigma \ln(1 + \exp(x_i'\beta/\sigma))$$

The following table shows a subset of the Mroz (1987) data set. In these data, Hours is the number of hours the wife worked outside the household in a given year, Yrs_Ed is the years of education, and Yrs_Exp is the years of work experience. A Tobit model will be fit to the hours worked with years of education and experience as covariates.

| Hours | Yrs_Ed | Yrs_Exp |
|-------|--------|---------|
| 0     | 8      | 9       |
| 0     | 8      | 12      |
| 0     | 9      | 10      |
| 0     | 10     | 15      |
| 0     | 11     | 4       |
| 0     | 11     | 6       |
| 1000  | 12     | 1       |
| 1960  | 12     | 29      |
| 0     | 13     | 3       |
| 2100  | 13     | 36      |
| 3686  | 14     | 11      |
| 1920  | 14     | 38      |
| 0     | 15     | 14      |
| 1728  | 16     | 3       |
| 1568  | 16     | 19      |
| 1316  | 17     | 7       |
| 0     | 17     | 15      |

If the wife was not employed (worked 0 hours), her hours worked will be left censored at zero. In order to accommodate left censoring in PROC LIFEREG, you need two variables to indicate censoring status of observations. You can think of these variables as lower and upper endpoints of interval censoring. If there is no censoring, set both variables to the observed value of Hours. To indicate left censoring, set the lower endpoint to missing and the upper endpoint to the censored value, zero in this case.

The following statements create a SAS data set with the variables Hours, Yrs_Ed, and Yrs_Exp from the preceding data. A new variable, Lower, is created such that Lower=. if Hours=0 and Lower=Hours if Hours>0.

```
data subset;
   input Hours Yrs_Ed Yrs_Exp @@;
   if Hours eq 0
      then Lower=.;
      else Lower=Hours;
   datalines;
0 8 9 0 8 12 0 9 10 0 10 15 0 11 4 0 11 6
1000 12 1 1960 12 29 0 13 3 2100 13 36
3686 14 11 1920 14 38 0 15 14 1728 16 3
1568 16 19 1316 17 7 0 17 15
;
```

The following statements fit a normal regression model to the left-censored Hours data with Yrs_Ed and Yrs_Exp as covariates. You need the estimated standard deviation of the normal distribution to compute the predicted values of the censored distribution from the preceding formulas. The data set OUTEST contains the standard deviation estimate in a variable named _SCALE_. You also need estimates of $x_i'\beta$. These are contained in the data set OUT as the variable Xbeta.

```
proc lifereg data=subset outest=OUTEST(keep=_scale_);
   model (lower, hours) = yrs_ed yrs_exp / d=normal;
   output out=OUT xbeta=Xbeta;
run;
```

Output 51.2.1 shows the results of the model fit. These tables show parameter estimates for the uncensored, or latent variable, distribution.

**Output 51.2.1** Parameter Estimates from PROC LIFEREG

```
                        The LIFEREG Procedure

                          Model Information

                Data Set                       WORK.SUBSET
                Dependent Variable                   Lower
                Dependent Variable                   Hours
                Number of Observations                  17
                Noncensored Values                       8
                Right Censored Values                    0
                Left Censored Values                     9
                Interval Censored Values                 0
                Number of Parameters                     4
                Name of Distribution                Normal
                Log Likelihood                  -74.9369977


           Analysis of Maximum Likelihood Parameter Estimates


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

     Intercept  1 -5598.64 2850.248 -11185.0 -12.2553    3.86    0.0495
     Yrs_Ed     1 373.1477 191.8872  -2.9442 749.2397    3.78    0.0518
     Yrs_Exp    1  63.3371  38.3632 -11.8533 138.5276    2.73    0.0987
     Scale      1 1582.870 442.6732 914.9433 2738.397
```

The following statements combine the two data sets created by PROC LIFEREG to compute predicted values for the censored distribution. The OUTEST= data set contains the estimate of the standard deviation from the uncensored distribution, and the OUT= data set contains estimates of $x_i' \beta$.

```
data predict;
   drop lambda _scale_ _prob_;
   set out;
   if _n_ eq 1 then set outest;
   lambda = pdf('NORMAL',Xbeta/_scale_)
            / cdf('NORMAL',Xbeta/_scale_);
   Predict = cdf('NORMAL', Xbeta/_scale_)
             * (Xbeta + _scale_*lambda);
   label Xbeta='MEAN OF UNCENSORED VARIABLE'
         Predict = 'MEAN OF CENSORED VARIABLE';
run;
```

Output 51.2.2 shows the original variables, the predicted means of the uncensored distribution, and the predicted means of the censored distribution.

**Output 51.2.2** Predicted Means from PROC LIFEREG

| Hours | Lower | Yrs_Ed | Yrs_Exp | MEAN OF UNCENSORED VARIABLE | MEAN OF CENSORED VARIABLE |
|-------|-------|--------|---------|------------------------------|----------------------------|
| 0 | . | 8 | 9 | −2043.42 | 73.46 |
| 0 | . | 8 | 12 | −1853.41 | 94.23 |
| 0 | . | 9 | 10 | −1606.94 | 128.10 |
| 0 | . | 10 | 15 | −917.10 | 276.04 |
| 0 | . | 11 | 4 | −1240.67 | 195.76 |
| 0 | . | 11 | 6 | −1113.99 | 224.72 |
| 1000 | 1000 | 12 | 1 | −1057.53 | 238.63 |
| 1960 | 1960 | 12 | 29 | 715.91 | 1052.94 |
| 0 | . | 13 | 3 | −557.71 | 391.42 |
| 2100 | 2100 | 13 | 36 | 1532.42 | 1672.50 |
| 3686 | 3686 | 14 | 11 | 322.14 | 805.58 |
| 1920 | 1920 | 14 | 38 | 2032.24 | 2106.81 |
| 0 | . | 15 | 14 | 885.30 | 1170.39 |
| 1728 | 1728 | 16 | 3 | 561.74 | 951.69 |
| 1568 | 1568 | 16 | 19 | 1575.13 | 1708.24 |
| 1316 | 1316 | 17 | 7 | 1188.23 | 1395.61 |
| 0 | . | 17 | 15 | 1694.93 | 1809.97 |

## Example 51.3: Overcoming Convergence Problems by Specifying Initial Values

This example illustrates the use of parameter initial value specification to help overcome convergence difficulties.

The following statements create a SAS data set.

```
data raw;
   input censor x c1 @@;
   datalines;
0 16 0.00    0 17 0.00    0 18 0.00
0 17 0.04    0 18 0.04    0 18 0.04
0 23 0.40    0 22 0.40    0 22 0.40
0 33 4.00    0 34 4.00    0 35 4.00
1 54 40.00   1 54 40.00   1 54 40.00
1 54 400.00 1 54 400.00 1 54 400.00
;
```

Output 51.3.1 shows the contents of the data set raw.

**Output 51.3.1** Contents of the Data Set

```
           Obs     censor      x        c1

            1         0       16       0.00
            2         0       17       0.00
            3         0       18       0.00
            4         0       17       0.04
            5         0       18       0.04
            6         0       18       0.04
            7         0       23       0.40
            8         0       22       0.40
            9         0       22       0.40
           10         0       33       4.00
           11         0       34       4.00
           12         0       35       4.00
           13         1       54      40.00
           14         1       54      40.00
           15         1       54      40.00
           16         1       54     400.00
           17         1       54     400.00
           18         1       54     400.00
```

The following SAS statements request that a Weibull regression model be fit to the data:

```
title 'OLS (Default) Initial Values';
proc lifereg data=raw;
   model x*censor(1) = c1 / distribution = Weibull itprint;
run;
```

Convergence was not attained in 50 iterations for this model, as the following messages to the log indicate:

```
WARNING: Convergence was not attained in 50 iterations. You might want to
         increase the maximum number of iterations (MAXITER= option) or
         change the convergence criteria (CONVERGE = value) in the MODEL
         statement.
WARNING: The procedure is continuing in spite of the above warning. Results
         shown are based on the last maximum likelihood iteration. Validity
         of the model fit is questionable.
```

The first line (iter=0) of the iteration history table, shown in Output 51.3.2, shows the default initial ordinary least squares (OLS) estimates of the parameters.

**Output 51.3.2** Initial Least Squares

```
                         OLS (Default) Initial Values

                           The LIFEREG Procedure

                  Iteration History for Parameter Estimates

   Iter    Ridge    Loglikelihood       Intercept            c1           Scale

     0        0        -22.891088     3.2324769714    0.0020664542    0.3995754195
     1        0        -16.427074     3.5337141598    0.0028713635    0.3283544365
     2        0        -13.216768     3.4480787541    0.0052801225    0.3816964358
     3        0         -5.0786635    3.1966395335    0.0191439929    0.2325418958
     4        0         -2.0018885    3.1848047525    0.0275425402    0.1963590539
     5        0         -0.1814984    3.1478989655    0.0374731819    0.2103607621
     6        0          2.90712131   3.0858183316    0.0659946149    0.1818245261
     7    0.063          2.9991781    3.1014479187    0.0661096622    0.1648677081
     8    0.063          3.01557837   3.0995493638    0.0662333056    0.1670552505
     9    0.063          3.0301815    3.0992317977    0.0663580659    0.1669529486
    10    0.063          3.0448013    3.0989901232    0.0664827053    0.1667371524
    11    0.063          3.05941254   3.0987507448    0.0666071514    0.1665197313
    12    0.063          3.07401474   3.0985118143    0.0667314052    0.1663026517
    13    0.063          3.08860788   3.0982732928     0.066855467    0.1660859472
    14    0.063          3.10319193   3.0980351787    0.0669793371    0.1658696184
    15    0.063          3.11776689   3.0977974713    0.0671030156    0.1656536651
    16    0.063          3.13233272   3.0975601698    0.0672265029    0.1654380873
    17    0.063           3.1468894   3.0973232737    0.0673497993     0.165222885
    18    0.063          3.16143692   3.0970867821    0.0674729049    0.1650080579
    19    0.063          3.17597526   3.0968506943     0.06759582     0.1647936061
    20    0.063          3.19050439   3.0966150098    0.0677185449    0.1645795293
    21    0.063           3.2050243   3.0963797277    0.0678410799    0.1643658275
    22    0.063          3.21953496   3.0961448474    0.0679634252    0.1641525006
    23    0.063          3.23403635   3.0959103682     0.068085581    0.1639395483
    24    0.063          3.24852845   3.0956762896    0.0682075476    0.1637269705
    25    0.063          3.26301123   3.0954426107    0.0683293253    0.1635147672
    26    0.063          3.27748468    3.095209331    0.0684509143     0.163302938
    27    0.063          3.29194878   3.0949764498    0.0685723149    0.1630914829
    28    0.063           3.3064035   3.0947439665    0.0686935273    0.1628804017
    29    0.063          3.32084881   3.0945118805    0.0688145517    0.1626696942
    30    0.063           3.3352847   3.0942801911    0.0689353885    0.1624593601
    31    0.063          3.34971114   3.0940488977    0.0690560378    0.1622493994
    32    0.063          3.36412812   3.0938179997       0.0691765    0.1620398118
    33    0.063           3.3785356   3.0935874965    0.0692967752    0.1618305971
    34    0.063          3.39293356   3.0933573875    0.0694168637     0.161621755
    35    0.063          3.40732199    3.093127672    0.0695367658    0.1614132855
    36    0.063          3.42170085   3.0928983495    0.0696564816    0.1612051882
    37    0.063          3.43607013   3.0926694194    0.0697760116    0.1609974629
    38    0.063          3.45042979   3.0924408811    0.0698953558    0.1607901095
    39    0.063          3.46477983    3.092212734    0.0700145146    0.1605831276
    40    0.063           3.4791202   3.0919849776    0.0701334882     0.160376517
    41    0.063           3.4934509   3.0917576112    0.0702522768    0.1601702775
    42    0.063          3.50777188   3.0915306343    0.0703708808    0.1599644088
    43    0.063          3.52208314   3.0913040464    0.0704893002    0.1597589108
    44    0.063          3.53638465   3.0910778468    0.0706075354     0.159553783
    45    0.063          3.55067637   3.0908520349    0.0707255867    0.1593490254
    46    0.063           3.5649583   3.0906266104    0.0708434542    0.1591446376
    47    0.063          3.57923039   3.0904015725    0.0709611382    0.1589406193
    48    0.063          3.59349263   3.0901769207    0.0710786389    0.1587369703
    49    0.063            3.607745   3.0899526546    0.0711959567    0.1585336903
    50    0.063          3.62198746   3.0897287734    0.0713130916    0.1583307791
```

The log-logistic distribution is more robust to large values of the response than the Weibull distribution, so one approach to improving the convergence performance is to fit a log-logistic distribution, and if this converges, use the resulting parameter estimates as initial values in a subsequent fit of a model with the Weibull distribution.

The following statements fit a log-logistic distribution to the data:

```
proc lifereg data=raw;
   model x*censor(1) = c1 / distribution = llogistic;
run;
```

The algorithm converges, and the maximum likelihood estimates for the log-logistic distribution are shown in Output 51.3.3

**Output 51.3.3** Estimates from the Log-Logistic Distribution

```
                      OLS (Default) Initial Values

                         The LIFEREG Procedure

              Analysis of Maximum Likelihood Parameter Estimates

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

      Intercept      1   2.8983   0.0318   2.8360   2.9606 8309.43   <.0001
      c1             1   0.1592   0.0133   0.1332   0.1852  143.85   <.0001
      Scale          1   0.0498   0.0122   0.0308   0.0804
```

The following statements refit the Weibull model by using the maximum likelihood estimates from the log-logistic fit as initial values:

```
proc lifereg data=raw outest=outest;
   model x*censor(1) = c1 / itprint distribution = weibull
                          intercept=2.898 initial=0.16 scale=0.05;
   output out=out xbeta=xbeta;
run;
```

Examination of the resulting output in Output 51.3.4 shows that the convergence problem has been solved by specifying different initial values.

**Output 51.3.4** Final Estimates from the Weibull Distribution

```
                        OLS (Default) Initial Values

                          The LIFEREG Procedure

                            Model Information

                Data Set                        WORK.RAW
                Dependent Variable               Log(x)
                Censoring Variable               censor
                Censoring Value(s)                    1
                Number of Observations               18
                Noncensored Values                   12
                Right Censored Values                 6
                Left Censored Values                  0
                Interval Censored Values              0
                Number of Parameters                  3
                Name of Distribution            Weibull
                Log Likelihood             11.232023272


      Algorithm converged.


              Analysis of Maximum Likelihood Parameter Estimates

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

      Intercept       1   2.9699  0.0326   2.9059   3.0338 8278.86     <.0001
      c1              1   0.1435  0.0165   0.1111   0.1758   75.43     <.0001
      Scale           1   0.0844  0.0189   0.0544   0.1308
      Weibull Shape   1  11.8526  2.6514   7.6455  18.3749
```

As an example of an alternative way of specifying initial values, the following invocation of PROC LIF-EREG, using the INEST= data set to provide starting values for the three parameters, is equivalent to the previous invocation:

```
data in;
   input  intercept c1 scale;
   datalines;
2.898 0.16 0.05
;

proc lifereg data=raw inest=in outest=outest;
   model x*censor(1) = c1 / itprint distribution = weibull;
   output out=out xbeta=xbeta;
run;
```

# Example 51.4: Analysis of Arbitrarily Censored Data with Interaction Effects

The artificial data in this example are from a study of the natural recovery time of mice after injection of a certain toxin. Twenty mice were grouped by sex (sex: 1 = Male, 2 = Female) with equal sizes. Their ages (in days) were recorded at the injection. Their recovery times (in minutes) were also recorded. Toxin density in blood was used to decide whether a mouse recovered. Mice were checked at two times for recovery. If a mouse had recovered at the first time, the observation is left censored, and no further measurement is made. The variable time1 is set to missing and time2 is set to the measurement time to indicate left censoring. If a mouse had not recovered at the first time, it was checked later at a second time. If it had recovered by the second measurement time, the observation is interval censored, and the variable time1 is set to the first measurement time and time2 is set to the second measurement time. If there was no recovery at the second measurement, the observation is right censored, and time1 is set to the second measurement time and time2 is set to missing to indicate right censoring.

The following statements create a SAS data set containing the data from the experiment:

```
title 'Natural Recovery Time';
data mice;
   input sex age time1 time2;
   datalines;
1  57  631   631
1  45  .     170
1  54  227   227
1  43  143   143
1  64  916   .
1  67  691   705
1  44  100   100
1  59  730   .
1  47  365   365
1  74  1916  1916
2  79  1326  .
2  75  837   837
2  84  1200  1235
2  54  .     365
2  74  1255  1255
2  71  1823  .
2  65  537   637
2  33  583   683
2  77  955   .
2  46  577   577
;
```

The following SAS statements create the SAS data sets xrow1 and xrow2:

```
data xrow1;
   input sex age time1 time2;
   datalines;
1  50  .  .
;

data xrow2;
   input sex age time1 time2;
   datalines;
2  60.6  .  .
;
```

The following SAS statements fit a Weibull model with age, sex, and an age-by-sex interaction term as covariates, and create a plot of predicted probabilities against recovery time for the fixed values of age and sex specified in the SAS data set xrow1:

```
ods graphics on;
proc lifereg data=mice xdata=xrow1;
   class sex;
   model (time1, time2) = age sex age*sex / dist=Weibull;

   probplot  / nodata
       plower=.5
       vref(intersect) = 75
       vreflab = '75 Percent';
   inset;
run;
```

Standard output is shown in Output 51.4.1. Tables containing general model information, Type III tests for the main effects and interaction terms, and parameter estimates are created.

**Output 51.4.1** Parameter Estimates for the Interaction Model

```
                        Natural Recovery Time

                        The LIFEREG Procedure

                         Model Information

            Data Set                         WORK.MICE
            Dependent Variable               Log(time1)
            Dependent Variable               Log(time2)
            Number of Observations                  20
            Noncensored Values                       9
            Right Censored Values                    5
            Left Censored Values                     2
            Interval Censored Values                 4
            Number of Parameters                     5
            Name of Distribution               Weibull
            Log Likelihood                 -25.91033295


                   Type III Analysis of Effects

                                       Wald
            Effect            DF    Chi-Square     Pr > ChiSq

            age                1      33.8496        <.0001
            sex                1      14.0245        0.0002
            age*sex            1      10.7196        0.0011


         Analysis of Maximum Likelihood Parameter Estimates

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

    Intercept        1   5.4110   0.5549   4.3234   6.4986   95.08   <.0001
    age              1   0.0250   0.0086   0.0081   0.0419    8.42   0.0037
    sex            1 1  -3.9808   1.0630  -6.0643  -1.8974   14.02   0.0002
    sex            2 0   0.0000      .        .        .        .        .
    age*sex        1 1   0.0613   0.0187   0.0246   0.0980   10.72   0.0011
    age*sex        2 0   0.0000      .        .        .        .        .
    Scale            1   0.4087   0.0900   0.2654   0.6294
    Weibull Shape    1   2.4468   0.5391   1.5887   3.7682
```

The following two plots display the predicted probability against the recovery time for two different populations. Output 51.4.2 is created with the PROBPLOT statement with the option XDATA= xrow1, which specifies the population with sex = 1, age = 50. Output 51.4.3 is created with the PROBPLOT statement with the option XDATA= xrow2, which specifies the population with sex = 2, age = 60.6. These are the default values that the LIFEREG procedure would use for the probability plot if the XDATA= option had not been specified. Reference lines are used to display specified predicted probability points and their relative locations in the plot.

**Output 51.4.2** Probability Plot for Recovery Time with sex = 1, age = 50
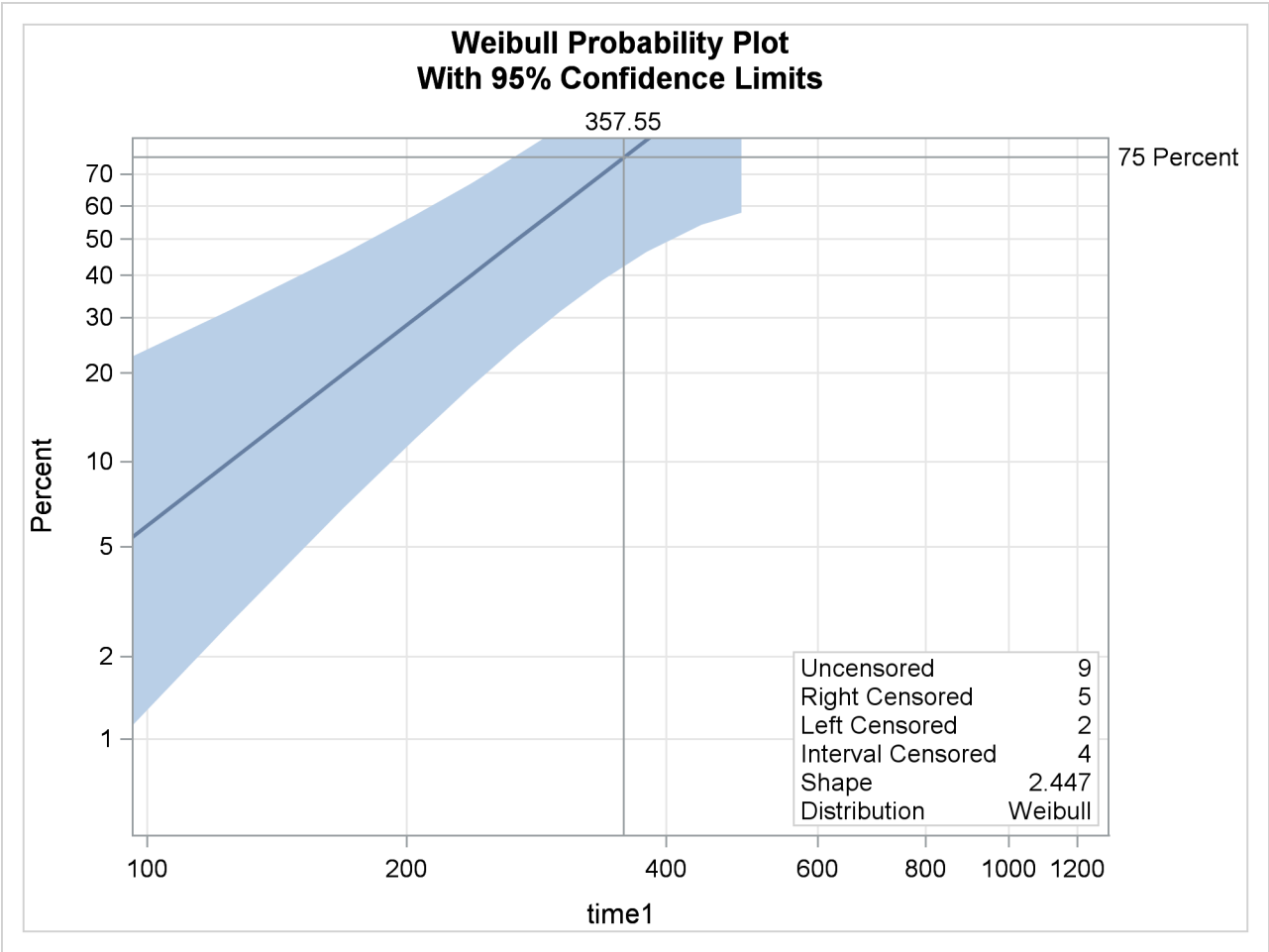
The following SAS statements fit a Weibull model with age, sex, and an age-by-sex interaction term as covariates, and create the plot of predicted probabilities against recovery time shown in Output 51.4.3, for the fixed values of age and sex specified in the SAS data set xrow2:

```
proc lifereg data=mice xdata=xrow2;
   class sex;
   model (time1, time2) = age sex age*sex / dist=Weibull;

   probplot  / nodata
       plower=.5
       vref(intersect) = 75
       vreflab = '75 Percent';
   inset;
run;
title;
ods graphics off;
```

**Output 51.4.3** Probability Plot for Recovery Time with sex = 2, age = 60.6
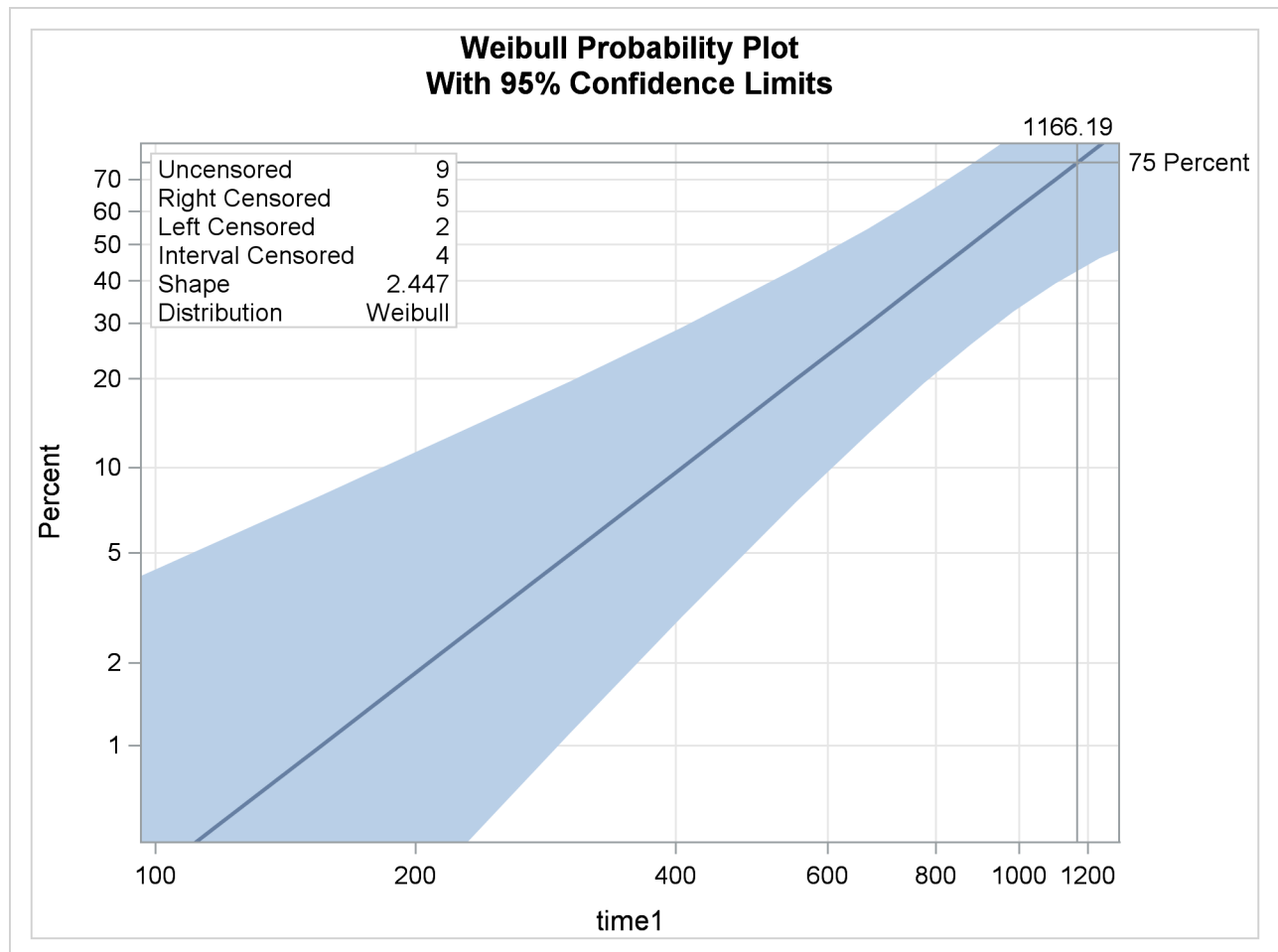
## Example 51.5: Probability Plotting—Right Censoring

The following statements create a SAS data set containing observed and right-censored lifetimes of 70 diesel engine fans (Nelson 1982):

```
data Fan;
   input Lifetime Censor@@;
   Lifetime = Lifetime / 1000;
   datalines;
 450 0    460 1   1150 0   1150 0   1560 1
1600 0   1660 1   1850 1   1850 1   1850 1
1850 1   1850 1   2030 1   2030 1   2030 1
2070 0   2070 0   2080 0   2200 1   3000 1
3000 1   3000 1   3000 1   3100 0   3200 1
3450 0   3750 1   3750 1   4150 1   4150 1
4150 1   4150 1   4300 1   4300 1   4300 1
4300 1   4600 0   4850 1   4850 1   4850 1
4850 1   5000 1   5000 1   5000 1   6100 1
6100 0   6100 1   6100 1   6300 1   6450 1
6450 1   6700 1   7450 1   7800 1   7800 1
8100 1   8100 1   8200 1   8500 1   8500 1
8500 1   8750 1   8750 0   8750 1   9400 1
9900 1  10100 1  10100 1  10100 1  11500 1
;
```

Some of the fans had not failed at the time the data were collected, and the unfailed units have right-censored lifetimes. The variable LIFETIME represents either a failure time or a censoring time, in thousands of hours. The variable CENSOR is equal to 0 if the value of LIFETIME is a failure time, and it is equal to 1 if the value is a censoring time. The following statements use the LIFEREG procedure to produce the probability plot with an inset for the engine lifetimes:

```
ods graphics on;
proc lifereg data=Fan;
   model Lifetime*Censor( 1 ) = / d = Weibull;
   probplot
   ppout
   npintervals=simul;
   inset;
run;
ods graphics off;
```

The resulting graphical output is shown in Output 51.5.1. The estimated CDF, a line representing the maximum likelihood fit, and pointwise parametric confidence bands are plotted in the body of Output 51.5.1. The values of right-censored observations are plotted along the bottom of the graph. The "Cumulative Probability Estimates" table is also created in Output 51.5.2.

**Output 51.5.1** Probability Plot for the Fan Data
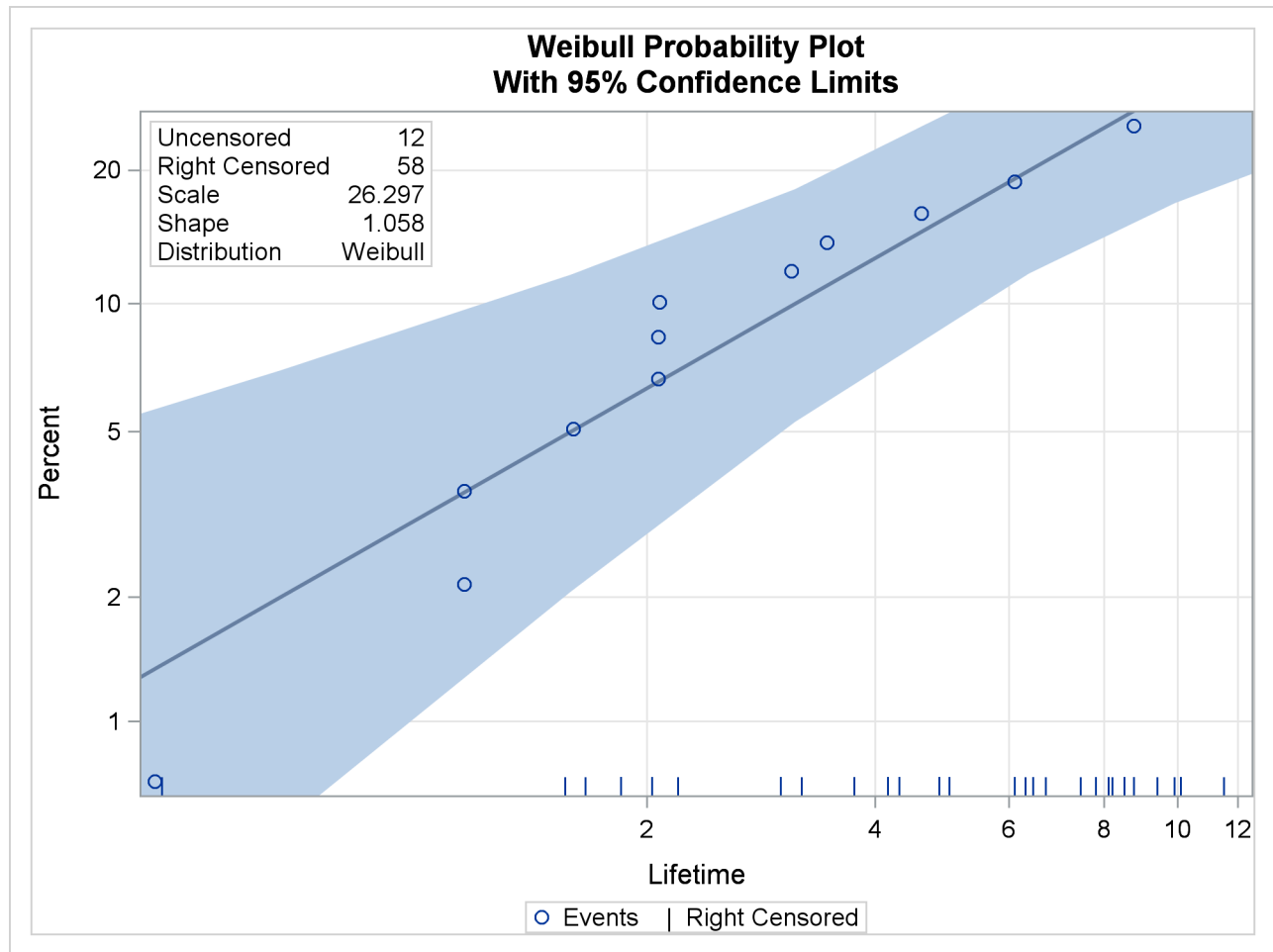


**Output 51.5.2** CDF Estimates

| | | Simultaneous 95% Confidence Limits | | Kaplan– | Kaplan– Meier |
|---|---|---|---|---|---|
| Lifetime | Cumulative Probability | Lower | Upper | Meier Estimate | Standard Error |
| 0.45 | 0.0071 | 0.0007 | 0.2114 | 0.0143 | 0.0142 |
| 1.15 | 0.0215 | 0.0033 | 0.2114 | 0.0288 | 0.0201 |
| 1.15 | 0.0360 | 0.0073 | 0.2168 | 0.0433 | 0.0244 |
| 1.6 | 0.0506 | 0.0125 | 0.2304 | 0.0580 | 0.0282 |
| 2.07 | 0.0666 | 0.0190 | 0.2539 | 0.0751 | 0.0324 |
| 2.07 | 0.0837 | 0.0264 | 0.2760 | 0.0923 | 0.0361 |
| 2.08 | 0.1008 | 0.0344 | 0.2972 | 0.1094 | 0.0392 |
| 3.1 | 0.1189 | 0.0436 | 0.3223 | 0.1283 | 0.0427 |
| 3.45 | 0.1380 | 0.0535 | 0.3471 | 0.1477 | 0.0460 |
| 4.6 | 0.1602 | 0.0653 | 0.3844 | 0.1728 | 0.0510 |
| 6.1 | 0.1887 | 0.0791 | 0.4349 | 0.2046 | 0.0581 |
| 8.75 | 0.2488 | 0.0884 | 0.6391 | 0.2930 | 0.0980 |

Cumulative Probability Estimates

## Example 51.6: Probability Plotting—Arbitrary Censoring

Table 51.17 contains microprocessor failure data (Nelson 1990). Units were inspected at predetermined time intervals. The data consist of inspection interval endpoints (in hours) and the number of units failing in each interval. A missing (.) lower endpoint indicates left censoring, and a missing upper endpoint indicates right censoring. These can be thought of as semi-infinite intervals with a lower (upper) endpoint of negative (positive) infinity for left (right) censoring.

**Table 51.17**   Interval-Censored Data

| Lower Endpoint | Upper Endpoint | Number Failed |
|---|---|---|
| . | 6 | 6 |
| 6 | 12 | 2 |
| 24 | 48 | 2 |
| 24 | . | 1 |
| 48 | 168 | 1 |
| 48 | . | 839 |
| 168 | 500 | 1 |
| 168 | . | 150 |
| 500 | 1000 | 2 |
| 500 | . | 149 |
| 1000 | 2000 | 1 |
| 1000 | . | 147 |
| 2000 | . | 122 |

The following SAS statements create the SAS data set Micro:

```
data Micro;
   input t1 t2 f;
   datalines;
. 6 6
6 12 2
12 24 0
24 48 2
24 .  1
48 168 1
48 .   839
168 500 1
168 .   150
500 1000 2
500 .    149
1000 2000 1
1000 . 147
2000 . 122
;
```

The following SAS statements compute the nonparametric Turnbull estimate of the cumulative distribution function and create a lognormal probability plot:

```
ods graphics on;
proc lifereg data=Micro;
   model ( t1 t2 ) = / d=lognormal intercept=25 scale=5;
   weight f;
   probplot
   pupper = 10
   itprintem
   printprobs
   maxitem = (1000,25)
   ppout;
   inset;
run;
ods graphics off;
```

The two initial values INTERCEPT=25 and SCALE=5 in the MODEL statement are used to aid convergence in the model-fitting algorithm.

The following tables are created by the PROBPLOT statement in addition to the standard tabular output from the MODEL statement. Output 51.6.1 shows the iteration history for the Turnbull estimate of the CDF for the microprocessor data. With both options ITPRINTEM and PRINTPROBS specified in the PROBPLOT statement, this table contains the log likelihoods and interval probabilities for every 25th iteration and the last iteration. It would contain only the log likelihoods if the option PRINTPROBS were not specified.

**Output 51.6.1** Iteration History for the Turnbull Estimate

```
                      The LIFEREG Procedure

        Iteration History for the Turnbull Estimate of the CDF
  Iteration   Loglikelihood       (., 6)        (6, 12)       (24, 48)      (48, 168)
                               (168, 500)     (500, 1000)   (1000, 2000)     (2000, .)

      0        -1133.4051         0.125          0.125         0.125          0.125
                                  0.125          0.125         0.125          0.125
     25         -104.16622      0.00421644     0.00140548    0.00140648     0.00173338
                                0.00237846     0.00846094    0.04565407     0.93474475
     50         -101.15151      0.00421644     0.00140548    0.00140648     0.00173293
                                0.00234891     0.00727679    0.01174486     0.96986811
     75         -101.06641      0.00421644     0.00140548    0.00140648     0.00173293
                                0.00234891     0.00727127    0.00835638     0.9732621
    100         -101.06534      0.00421644     0.00140548    0.00140648     0.00173293
                                0.00234891     0.00727125    0.00801814     0.97360037
    125         -101.06533      0.00421644     0.00140548    0.00140648     0.00173293
                                0.00234891     0.00727125    0.00798438     0.97363413
    130         -101.06533      0.00421644     0.00140548    0.00140648     0.00173293
                                0.00234891     0.00727125    0.007983       0.97363551
```

The table in Output 51.6.2 summarizes the Turnbull estimates of the interval probabilities, the reduced gradients, and Lagrange multipliers as described in the section "Arbitrarily Censored Data" on page 3942.

**Output 51.6.2** Summary for the Turnbull Algorithm

| Lower Lifetime | Upper Lifetime | Probability | Reduced Gradient | Lagrange Multiplier |
|---|---|---|---|---|
| . | 6 | 0.0042 | 0 | 0 |
| 6 | 12 | 0.0014 | 0 | 0 |
| 24 | 48 | 0.0014 | 0 | 0 |
| 48 | 168 | 0.0017 | 0 | 0 |
| 168 | 500 | 0.0023 | 0 | 0 |
| 500 | 1000 | 0.0073 | −7.219342E−9 | 0 |
| 1000 | 2000 | 0.0080 | −0.037063236 | 0 |
| 2000 | . | 0.9736 | 0.0003038877 | 0 |

Output 51.6.3 shows the final estimate of the CDF, along with standard errors and nonparametric confidence limits. Two kinds of nonparametric confidence limits, pointwise or simultaneous, are available. The default is the pointwise nonparametric confidence limits. You can specify the simultaneous nonparametric confidence limits by using the NPINTERVALS=SIMUL option.

**Output 51.6.3** Final CDF Estimates for Turnbull Algorithm

| | | Cumulative Probability Estimates | | | |
|---|---|---|---|---|---|
| | | | Pointwise 95% Confidence Limits | | |
| Lower Lifetime | Upper Lifetime | Cumulative Probability | Lower | Upper | Standard Error |
|---|---|---|---|---|---|
| 6 | 6 | 0.0042 | 0.0019 | 0.0094 | 0.0017 |
| 12 | 24 | 0.0056 | 0.0028 | 0.0112 | 0.0020 |
| 48 | 48 | 0.0070 | 0.0038 | 0.0130 | 0.0022 |
| 168 | 168 | 0.0088 | 0.0047 | 0.0164 | 0.0028 |
| 500 | 500 | 0.0111 | 0.0058 | 0.0211 | 0.0037 |
| 1000 | 1000 | 0.0184 | 0.0094 | 0.0357 | 0.0063 |
| 2000 | 2000 | 0.0264 | 0.0124 | 0.0553 | 0.0101 |

Output 51.6.4 shows the CDF estimates, maximum likelihood fit, and pointwise parametric confidence limits plotted on a lognormal probability plot.

**Output 51.6.4** Lognormal Probability Plot for the Microprocessor Data
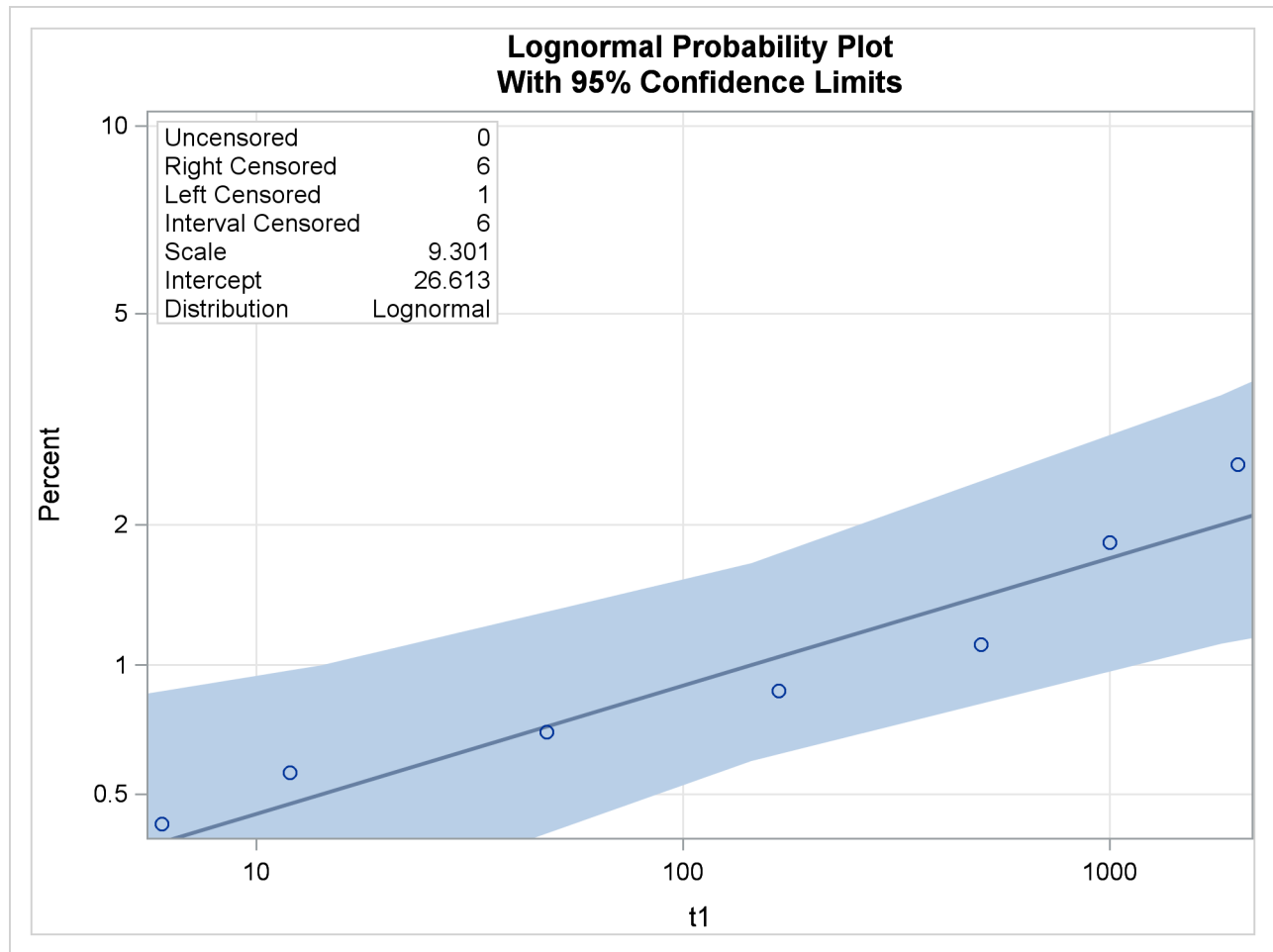


## Example 51.7: Bayesian Analysis of Clinical Trial Data

Consider the data on melanoma patients from a clinical trial described in Ibrahim, Chen, and Sinha (2001). A partial listing of the data is shown in Output 51.7.1.

The survival time is modeled by a Weibull regression model with three covariates. An analysis of the right-censored survival data is performed with PROC LIFEREG to obtain Bayesian estimates of the regression coefficients by using the following SAS statements:

```
ods graphics on;
proc lifereg data=e1684;
   class Sex;
   model Survtime*Survcens(1)=Age Sex Perform / dist=Weibull;
   bayes WeibullShapePrior=gamma seed=9999;
run;
ods graphics off;
```

**Output 51.7.1** Clinical Trial Data

| Obs | survtime | survcens | age | sex | perform |
|-----|----------|----------|---------|-----|---------|
| 1 | 1.57808 | 2 | 35.9945 | 1 | 0 |
| 2 | 1.48219 | 2 | 41.9014 | 1 | 0 |
| 3 | 7.33425 | 1 | 70.2164 | 2 | 0 |
| 4 | 0.65479 | 2 | 58.1753 | 2 | 1 |
| 5 | 2.23288 | 2 | 33.7096 | 1 | 0 |
| 6 | 9.38356 | 1 | 47.9726 | 1 | 0 |
| 7 | 3.27671 | 2 | 31.8219 | 2 | 0 |
| 8 | 0.00000 | 1 | 72.3644 | 2 | 0 |
| 9 | 0.80274 | 2 | 40.7151 | 2 | 0 |
| 10 | 9.64384 | 1 | 32.9479 | 1 | 0 |
| 11 | 1.66575 | 2 | 35.9205 | 1 | 0 |
| 12 | 0.94247 | 2 | 40.5068 | 2 | 0 |
| 13 | 1.68767 | 2 | 57.0384 | 1 | 0 |
| 14 | 5.94247 | 2 | 63.1452 | 1 | 0 |
| 15 | 2.34247 | 2 | 62.0630 | 1 | 0 |
| 16 | 0.89863 | 2 | 56.5342 | 1 | 1 |
| 17 | 9.03288 | 1 | 22.9945 | 2 | 0 |
| 18 | 9.63014 | 1 | 18.4712 | 1 | 0 |
| 19 | 0.52603 | 2 | 41.2521 | 1 | 0 |
| 20 | 1.82192 | 2 | 29.5178 | 1 | 0 |

Maximum likelihood estimates of the model parameters shown in Output 51.7.2 are displayed by default.

**Output 51.7.2** Maximum Likelihood Parameter Estimates

The LIFEREG Procedure

Bayesian Analysis

Analysis of Maximum Likelihood Parameter Estimates

| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | |
|-----------|-----|----------|----------|---------|--------|
| Intercept | 1 | 2.4402 | 0.3716 | 1.7119 | 3.1685 |
| age | 1 | −0.0115 | 0.0070 | −0.0253 | 0.0023 |
| sex | 1 1 | −0.1170 | 0.1978 | −0.5046 | 0.2707 |
| sex | 2 0 | 0.0000 | . | . | . |
| perform | 1 | 0.2905 | 0.3222 | −0.3411 | 0.9220 |
| Scale | 1 | 1.2537 | 0.0824 | 1.1021 | 1.4260 |
| Weibull Shape | 1 | 0.7977 | 0.0524 | 0.7012 | 0.9073 |

Since no prior distributions for the regression coefficients were specified, the default uniform improper distributions shown in the "Uniform Prior for Regression Coefficients" table in Output 51.7.3 are used. The specified gamma prior for the Weibull shape parameter is also shown in Output 51.7.3.

**Output 51.7.3** Model Parameter Priors

```
                        The LIFEREG Procedure

                          Bayesian Analysis

                Uniform Prior for Regression Coefficients

                        Parameter     Prior

                        Intercept     Constant
                        age           Constant
                        sex1          Constant
                        perform       Constant


            Independent Prior Distributions for Model Parameters

                    Prior
Parameter           Distribution                 Hyperparameters

Weibull Shape       Gamma           Shape       0.001    Inverse Scale       0.001
```

Fit statistics, descriptive statistics, interval statistics, and the sample parameter correlation matrix for the posterior sample are displayed in the tables in Output 51.7.4. Since noninformative prior distributions for the regression coefficients were used, the mean and standard deviations of the posterior distributions for the model parameters are close to the maximum likelihood estimates and standard errors.

**Output 51.7.4** Posterior Sample Statistics

```
                            Fit Statistics

            DIC (smaller is better)                875.251
            pD (effective number of parameters)      4.984


                        The LIFEREG Procedure

                          Bayesian Analysis

                         Posterior Summaries

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

Intercept        10000    2.4668    0.3862     2.1989    2.4621    2.7256
age              10000   -0.0115    0.00733   -0.0163   -0.0115   -0.00652
sex1             10000   -0.1255    0.2004    -0.2584   -0.1247    0.00817
perform          10000    0.3304    0.3317     0.1071    0.3188    0.5470
WeibShape        10000    0.7834    0.0518     0.7481    0.7815    0.8178
```

**Output 51.7.4** *continued*

```
                      Posterior Intervals

   Parameter     Alpha      Equal-Tail Interval        HPD Interval

   Intercept     0.050       1.7279    3.2368       1.7234    3.2264
   age           0.050      -0.0260    0.00263     -0.0261    0.00244
   sex1          0.050      -0.5197    0.2676      -0.5260    0.2583
   perform       0.050      -0.2898    1.0072      -0.3200    0.9726
   WeibShape     0.050       0.6846    0.8905       0.6805    0.8849


                    Posterior Correlation Matrix

                                                              Weib
   Parameter      Intercept       age       sex1     perform   Shape

   Intercept        1.0000     -.9018     -.3099     -.0888   -.1140
   age             -.9018      1.0000     -.0259     -.0363    0.0493
   sex1            -.3099      -.0259      1.0000     0.1248    0.0371
   perform         -.0888      -.0363      0.1248     1.0000   -.0355
   WeibShape       -.1140      0.0493      0.0371     -.0355    1.0000
```

The default diagnostic statistics are displayed in Output 51.7.5. See the section "Assessing Markov Chain Convergence" on page 139 for more details on Bayesian convergence diagnostics.

**Output 51.7.5** Convergence Diagnostics

```
                    The LIFEREG Procedure

                      Bayesian Analysis

                  Posterior Autocorrelations

   Parameter      Lag 1      Lag 5      Lag 10     Lag 50

   Intercept     0.0564     0.0030     0.0082     0.0234
   age          -0.0079    -0.0184    -0.0015     0.0239
   sex1          0.6293     0.0700     0.0055    -0.0199
   perform       0.6514     0.0773     0.0397    -0.0123
   WeibShape     0.0719    -0.0083    -0.0062     0.0112


                    Geweke Diagnostics

           Parameter              z      Pr > |z|

           Intercept          0.4962      0.6198
           age               -0.4119      0.6804
           sex1              -0.2519      0.8011
           perform           -0.1049      0.9165
           WeibShape         -0.6573      0.5110
```

**Output 51.7.5** *continued*

```
                    Effective Sample Sizes

                                 Autocorrelation
          Parameter        ESS            Time      Efficiency

          Intercept      7476.1          1.3376        0.7476
          age           10000.0          1.0000        1.0000
          sex1           2482.1          4.0288        0.2482
          perform        2174.0          4.5998        0.2174
          WeibShape      8538.8          1.1711        0.8539
```

Trace, autocorrelation, and density plots for the seven model parameters are shown in Output 51.7.6 through Output 51.7.10. These plots show no indication that the Markov chains have not converged. See the sections "Assessing Markov Chain Convergence" on page 139 and "Visual Analysis via Trace Plots" on page 139 for more information about assessing the convergence of the chain of posterior samples.

**Output 51.7.6** Diagnostic Plots
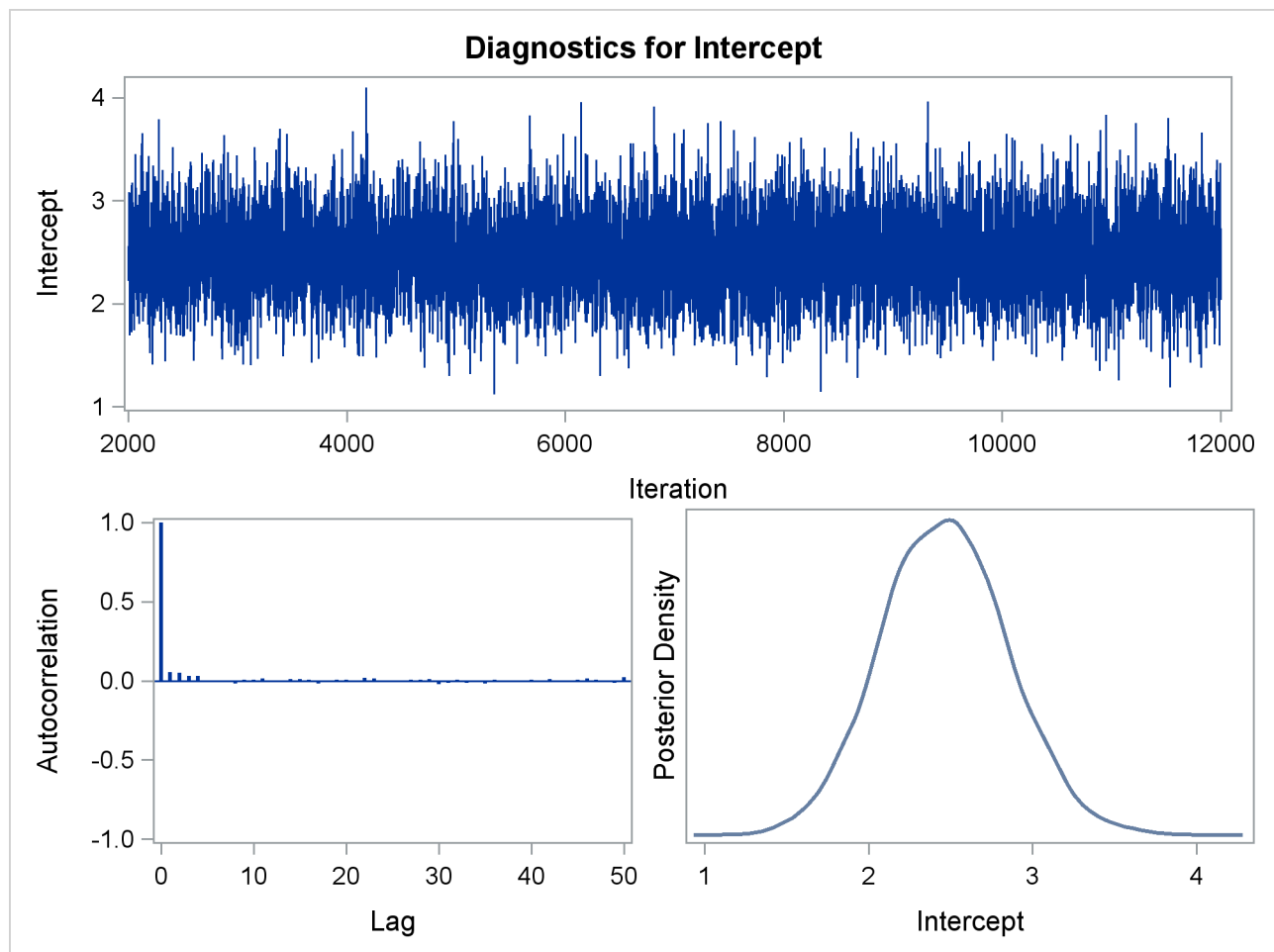
**Output 51.7.7** Diagnostic Plots
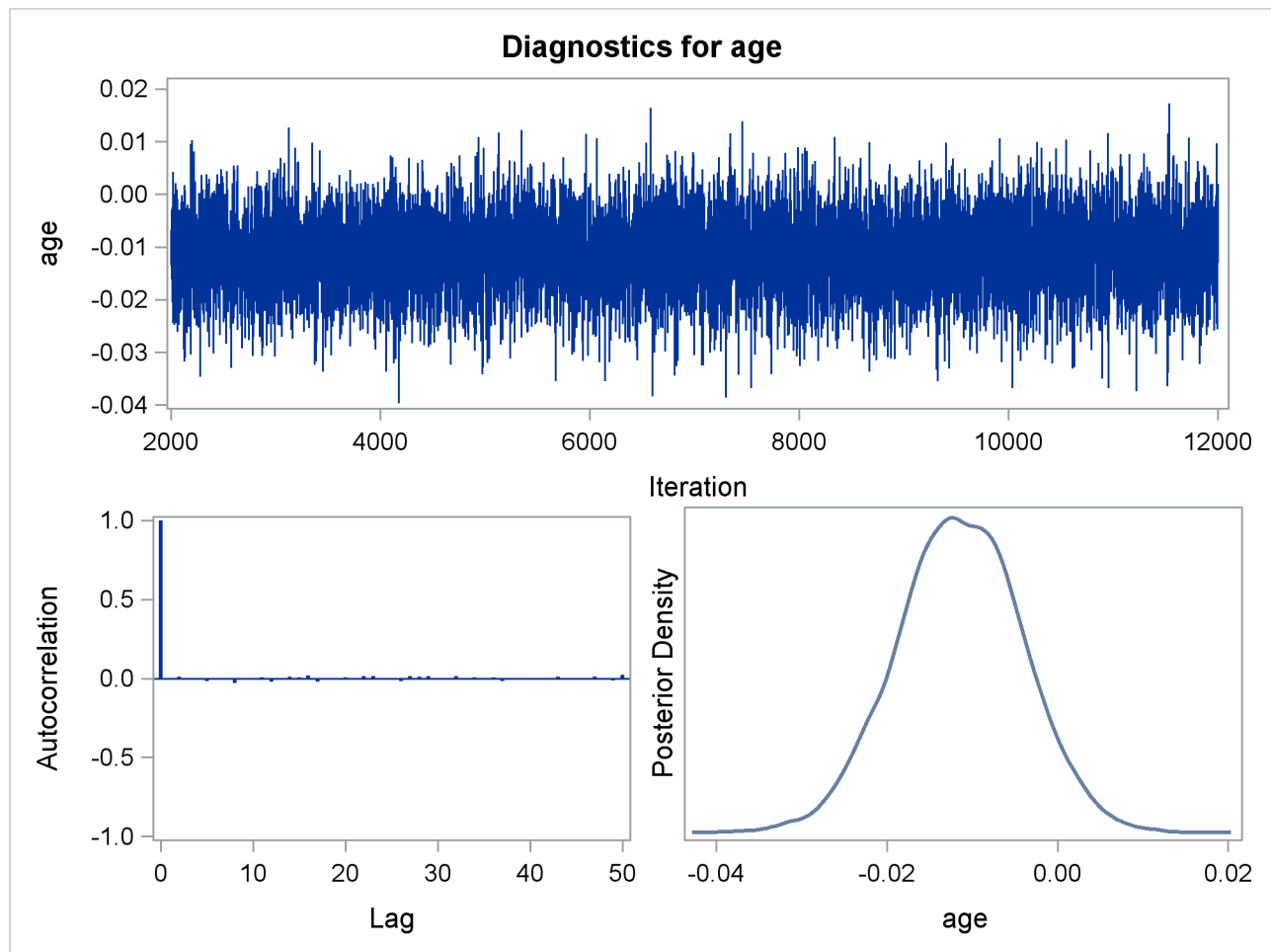
**Output 51.7.8** Diagnostic Plots
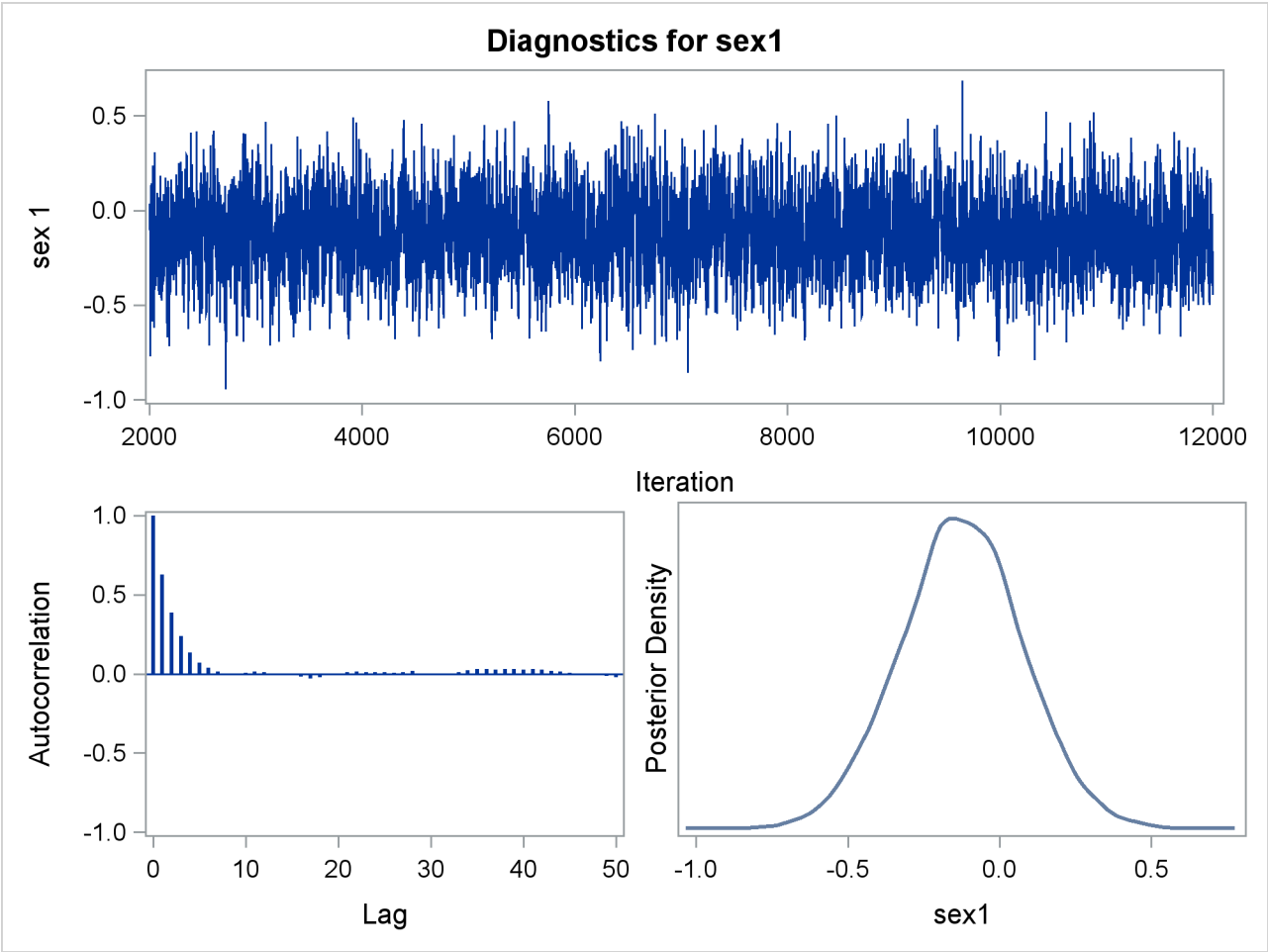
**Output 51.7.9** Diagnostic Plots
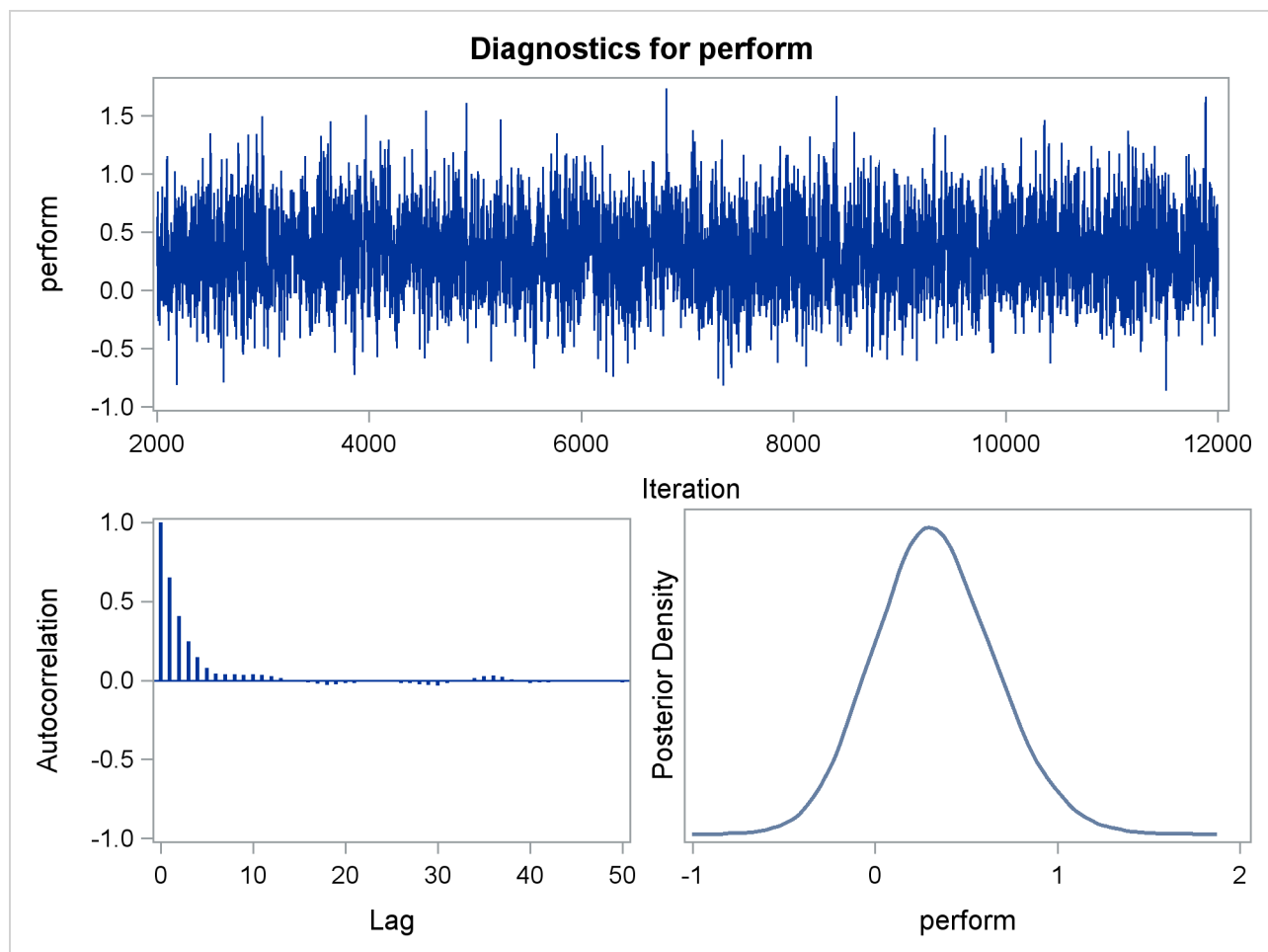
**Output 51.7.10** Diagnostic Plots
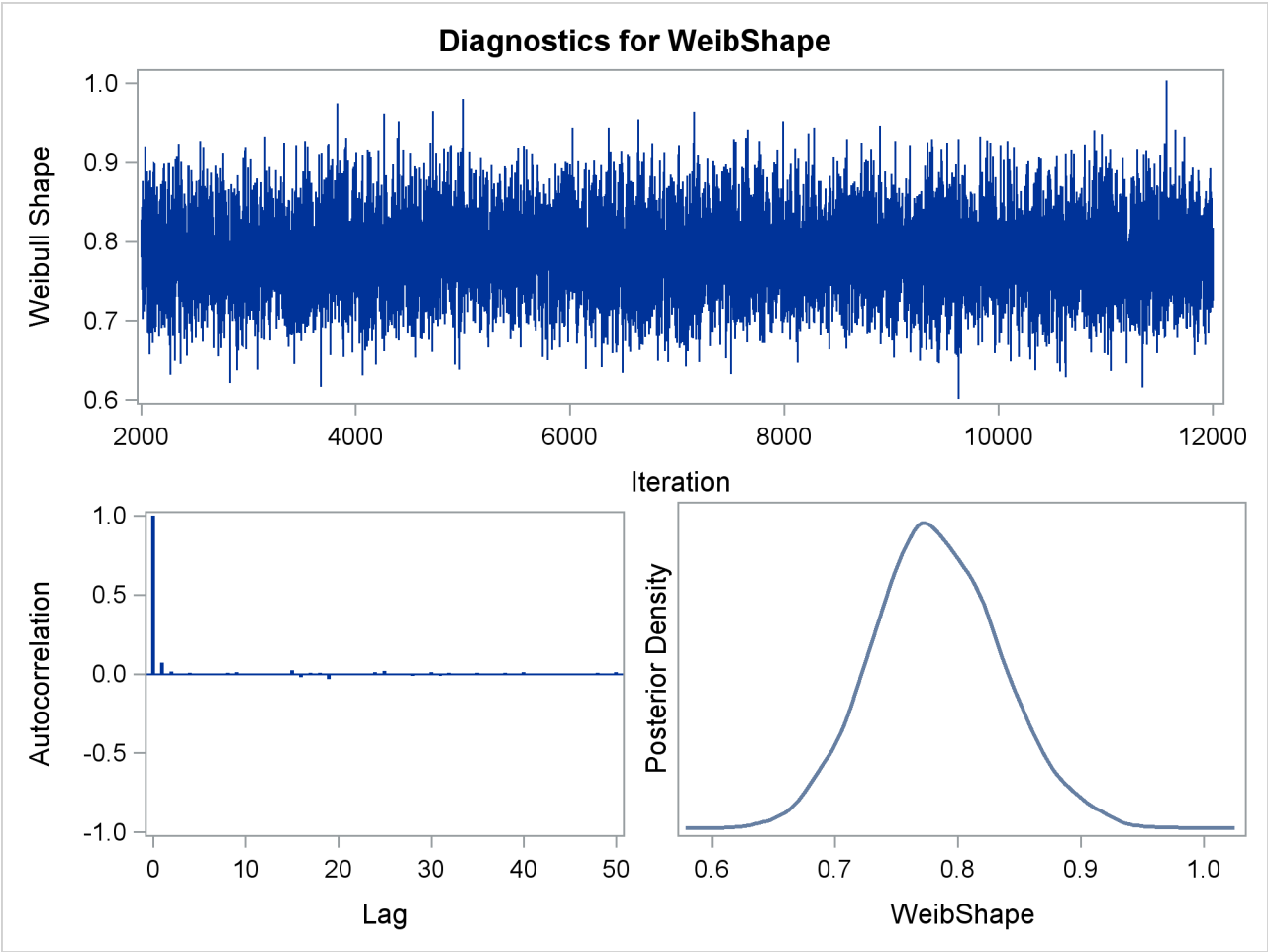
## Example 51.8: Model Postfitting Analysis

PROC LIFEREG enables you to make model-based inferences. This example uses the larynx cancer data (Klein and Moeschberger 1997) to illustrate usage of the LSMEANS, LSMESTIMATE, and EFFECTPLOT statements for model postfitting analysis.

The survival time is modeled by a proportional odds model with two covariates: patient age and cancer stage (1, 2, 3, 4). The following statements use PROC LIFEREG to fit this model:

```
ods graphics on;

proc sort data=Larynx;
   by DESCENDING Stage;
run;

proc lifereg data=Larynx order=data;
   class Stage;
   model Time*Death(0) = Age Stage / dist = llogistic;
   lsmeans Stage / diff adjust=Sidak;
   effectplot / noobs;
run;
```

The LSMEANS statement compares pairwise differences in survival times among the four different cancer stages, while adjusting for age. The ADJUST=SIDAK option uses the Sidak method to control the overall Type I error rate of these comparisons. Results are displayed in Output 51.8.1.

**Output 51.8.1** LS-Means Differences between Disease Stages

```
                        The LIFEREG Procedure


                 Differences of Stage Least Squares Means
                 Adjustment for Multiple Comparisons: Sidak


                              Standard
    Stage    _Stage    Estimate      Error    z Value    Pr > |z|     Adj P

      4        3       -0.9604      0.4379      -2.19      0.0283     0.1581
      4        2       -1.6404      0.4931      -3.33      0.0009     0.0053
      4        1       -1.7661      0.4257      -4.15     <.0001      0.0002
      3        2       -0.6800      0.4316      -1.58      0.1151     0.5199
      3        1       -0.8057      0.3539      -2.28      0.0228     0.1292
      2        1       -0.1257      0.4152      -0.30      0.7621     0.9998
```

All the LS-means differences and their significance are displayed by the mean-mean scatter plot in Output 51.8.2.

**Output 51.8.2** Plot of Pairwise LS-Means Differences
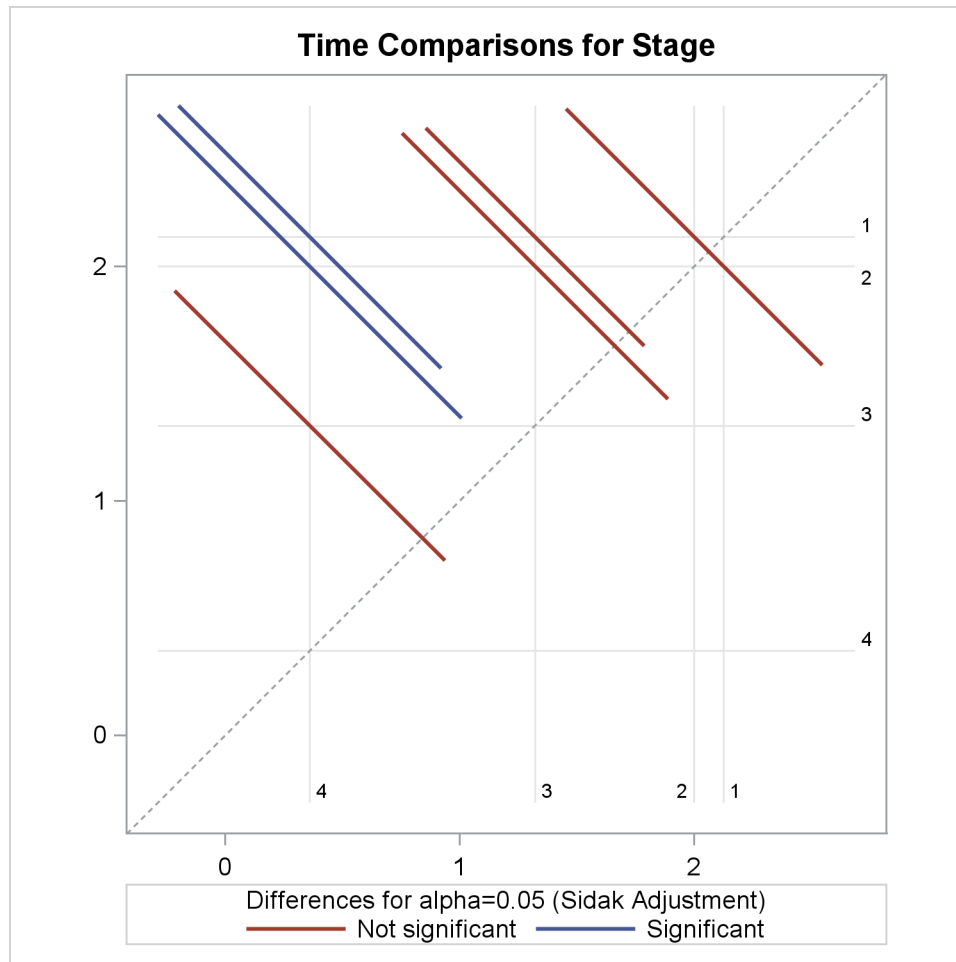


Suppose you want to jointly test whether the effects of stages 2, 3, and 4 are different from stage 1. The following LSMESTIMATE statement contrasts the LS-means of stages 2, 3, and 4 against the LS-means of stage 1:

```
proc lifereg data=Larynx order=data;
   class Stage year;
   model Time*Death(0) = Age Stage / dist = llogistic;
   lsmestimate Stage 'Stage 4 vs 1' 1 0 0 -1,
                     'Stage 3 vs 1' 0 1 0 -1,
                     'Stage 2 vs 1' 0 0 1 -1 / cl adjust=Sidak;
run;
```

The CL option produces 95% confidence limits, including both unadjusted ones and those adjusted for multiple comparisons according to the ADJUST= option. Results are displayed in Output 51.8.3.

**Output 51.8.3** Custom LS-Means Tests and Relative Odds

```
                        The LIFEREG Procedure

                    Least Squares Means Estimates
                  Adjustment for Multiplicity: Sidak


                            Standard
Effect   Label            Estimate     Error   z Value  Pr > |z|   Adj P    Alpha

Stage    Stage 4 vs 1    -1.7661     0.4257    -4.15     <.0001   0.0001    0.05
Stage    Stage 3 vs 1    -0.8057     0.3539    -2.28     0.0228   0.0668    0.05
Stage    Stage 2 vs 1    -0.1257     0.4152    -0.30     0.7621   0.9865    0.05

                    Least Squares Means Estimates
                  Adjustment for Multiplicity: Sidak


                                                Adj          Adj
        Effect   Label            Lower       Upper        Lower        Upper

        Stage    Stage 4 vs 1    -2.6004    -0.9319      -2.7825      -0.7498
        Stage    Stage 3 vs 1    -1.4993    -0.1122      -1.6507       0.03921
        Stage    Stage 2 vs 1    -0.9395     0.6881      -1.1171       0.8657
```

As displayed in Output 51.8.4, the EFFECTPLOT statement generates a plot of age effects on survival time on a natural logarithm scale by four disease stages.

**Output 51.8.4** Age Effects by Disease Stages
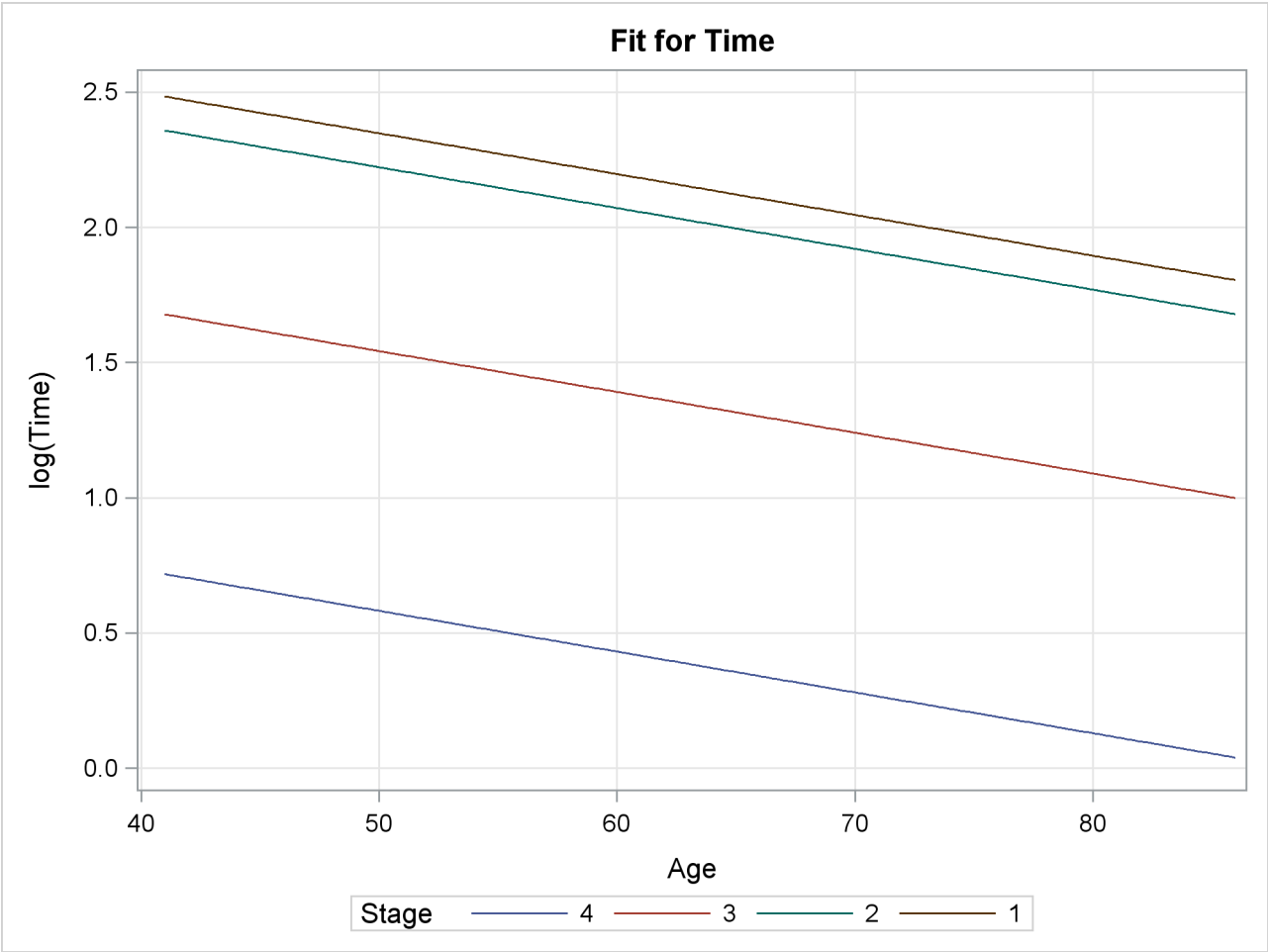
You can also perform the preceding analysis for a Bayesian model. The following statements generate posterior samples from a Bayesian model and request an LS-means analysis to compare the stage effects:

```
proc lifereg data=Larynx order=data;
   class Stage;
   model Time*Death(0) = Age Stage / dist = llogistic;
   bayes seed=100 nmc=500 nbi=500 diagnostic=none outpost=OOO;
   lsmeans Stage / diff exp;
   lsmestimate Stage 'Stage 4 vs 1' 1 0 0 -1,
                     'Stage 3 vs 1' 0 1 0 -1,
                     'Stage 2 vs 1' 0 0 1 -1
                        / cl plots=boxplot(orient=horizontal);
run;
```

Because no prior distributions for the regression coefficients were specified, the default uniform improper distributions shown in the "Uniform Prior for Regression Coefficients" table in Output 51.8.5 are used. The specified gamma prior for the scale parameter is also shown in Output 51.8.5.

**Output 51.8.5** Model Parameter Priors

```
                      The LIFEREG Procedure

                       Bayesian Analysis

             Uniform Prior for Regression Coefficients

                       Parameter      Prior

                       Intercept      Constant
                       Age            Constant
                       Stage4         Constant
                       Stage3         Constant
                       Stage2         Constant


           Independent Prior Distributions for Model Parameters

                   Prior
   Parameter     Distribution                  Hyperparameters

    Scale          Gamma          Shape      0.001    Inverse Scale      0.001
```

Under the Bayesian framework, the LS-means differences are treated as random variables for which posterior samples are readily available according to the linear relationship of LS-means and the regression coefficients. Output 51.8.6 lists the sample mean, standard deviation, and percentiles for each LS-means difference.

**Output 51.8.6** LS-Means Differences between Disease Stages

```
              Sample Differences of Stage Least Squares Means

                                    Standard   ----------Percentiles---------
  Stage   _Stage       N    Estimate  Deviation     25th       50th       75th

    4       3         500    -0.9307    0.4752    -1.2743    -0.9446    -0.6086
    4       2         500    -1.6591    0.5327    -2.0161    -1.6573    -1.2861
    4       1         500    -1.8001    0.4321    -2.0951    -1.7943    -1.5491
    3       2         500    -0.7284    0.4828    -1.0488    -0.7219    -0.3975
    3       1         500    -0.8694    0.3727    -1.1199    -0.8541    -0.6149
    2       1         500    -0.1410    0.4413    -0.4126    -0.1417     0.1363

              Sample Differences of Stage Least Squares Means

                                    Standard   --------Percentiles for
                                     Error of        Exponentiated---------
  Stage   _Stage   Exponentiated  Exponentiated     25th       50th       75th

    4       3           0.4426       0.232690    0.2796     0.3888     0.5441
    4       2           0.2181       0.115808    0.1332     0.1907     0.2763
    4       1           0.1815       0.082975    0.1231     0.1663     0.2124
    3       2           0.5410       0.268735    0.3504     0.4858     0.6720
    3       1           0.4488       0.168055    0.3263     0.4257     0.5407
    2       1           0.9585       0.462376    0.6619     0.8679     1.1461
```

The LSMESTIMATE statement produces summary statistics of the posterior samples for the specified LS-means contrasts. Results are presented in Output 51.8.7; they are very similar to the results based on maximum likelihood in Output 51.8.3.

**Output 51.8.7** Summary Statistics of Custom LS-Means Differences

```
                  Sample Least Squares Means Estimates

                                     Standard   ---------Percentiles--------
  Effect   Label            N    Estimate  Deviation     25th      50th      75th

  Stage    Stage 4 vs 1    500    -1.8001    0.4321    -2.0951   -1.7943   -1.5491
  Stage    Stage 3 vs 1    500    -0.8694    0.3727    -1.1199   -0.8541   -0.6149
  Stage    Stage 2 vs 1    500    -0.1410    0.4413    -0.4126   -0.1417    0.1363

                  Sample Least Squares Means Estimates

                                           Lower       Upper
              Effect   Label       Alpha    HPD         HPD

              Stage    Stage 4 vs 1  0.05   -2.6279     -0.8897
              Stage    Stage 3 vs 1  0.05   -1.6033     -0.2031
              Stage    Stage 2 vs 1  0.05   -1.1401      0.6252
```

The PLOTS= option uses ODS Graphics to display the Bayesian samples. A box plot is presented in Output 51.8.8.

**Output 51.8.8** Box Plot of Sampled LS-Means Differences



Distribution of Stage Least Squares Means Estimates

---
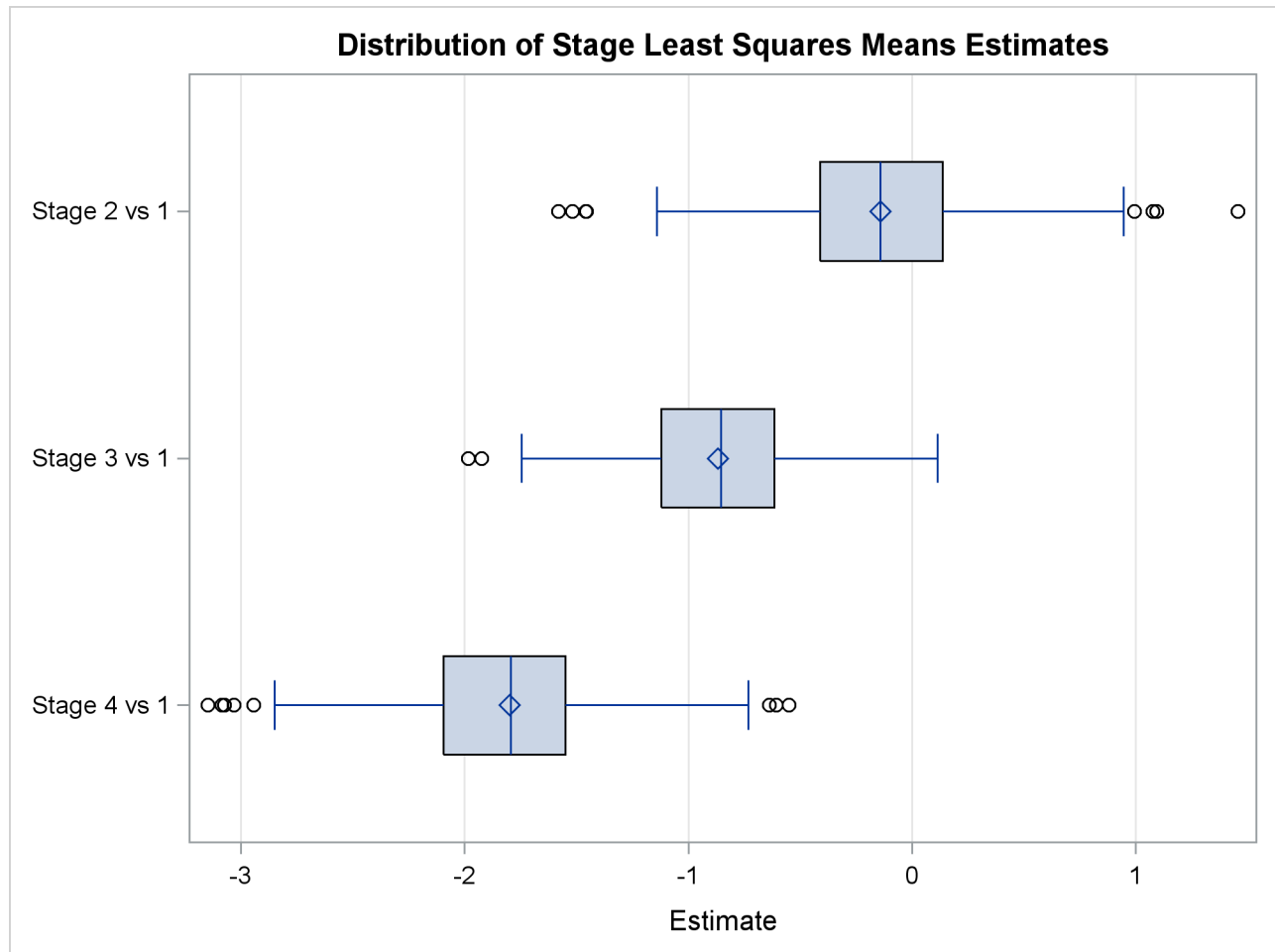
# References

Abernethy, R. B. (1996), *The New Weibull Handbook*, Second Edition, North Palm Beach, FL: Robert B. Abernethy.

Akaike, H. (1979), "A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting," *Biometrika*, 66, 237–242.

Akaike, H. (1981), "Likelihood of a Model and Information Criteria," *Journal of Econometrics*, 16, 3–14.

Cox, D. R. and Oakes, D. (1984), *Analysis of Survival Data*, London: Chapman & Hall.

Gentleman, R. and Geyer, C. J. (1994), "Maximum Likelihood for Interval Censored Data: Consistency and Computation," *Biometrika*, 81, 618–623.

Gilks, W. (2003), "Adaptive Metropolis Rejection Sampling (ARMS)," software from MRC Biostatistics Unit, Cambridge, UK, http://www.maths.leeds.ac.uk/~wally.gilks/adaptive.rejection/web_page/Welcome.html.

Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995), "Adaptive Rejection Metropolis Sampling with Gibbs Sampling," *Applied Statistics*, 44, 455–472.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall.

Gilks, W. R. and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics*, 41, 337–348.

Greene, W. H. (1993), *Econometric Analysis*, Second Edition, New York: Macmillan.

Ibrahim, J. G., Chen, M. H., and Sinha, D. (2001), *Bayesian Survival Analysis*, New York: Springer-Verlag.

Kalbfleisch, J. D. and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data*, New York: John Wiley & Sons.

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

Lawless, J. F. (2003), *Statistical Model and Methods for Lifetime Data*, Second Edition, New York: John Wiley & Sons.

Maddala, G. S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, New York: Cambridge University Press.

Meeker, W. Q. and Escobar, L. A. (1998), *Statistical Methods for Reliability Data*, New York: John Wiley & Sons.

Mroz, T. A. (1987), "The Sensitivity of an Empirical Model of Married Women's Work to Economic and Statistical Assumptions," *Econometrica*, 55, 765–799.

Nair, V. N. (1984), "Confidence Bands for Survival Functions with Censored Data: A Comparative Study," *Technometrics*, 26, 265–275.

Nelson, W. (1982), *Applied Life Data Analysis*, New York: John Wiley & Sons.

Nelson, W. (1990), *Accelerated Testing: Statistical Models, Test Plans, and Data Analyses*, New York: John Wiley & Sons.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, Second Edition, New York: John Wiley & Sons.

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit," *Journal of the Royal Statistical Society, Series B*, 64(4), 583–616, with discussion.

Tobin, J. (1958), "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 26, 24–36.

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

# Subject Index

# Syntax Index

# Your Turn

We welcome your feedback.

- If you have comments about this book, please send them to **yourturn@sas.com**. Include the full title and page numbers (if applicable).
- If you have comments about the software, please send them to **suggest@sas.com**.

# SAS® Publishing Delivers!

**Whether you are new to the work force or an experienced professional, you need to distinguish yourself in this rapidly changing and competitive job market. SAS® Publishing provides you with a wide range of resources to help you set yourself apart. Visit us online at support.sas.com/bookstore.**

## SAS® Press

Need to learn the basics? Struggling with a programming problem? You'll find the expert answers that you need in example-rich books from SAS Press. Written by experienced SAS professionals from around the world, SAS Press books deliver real-world insights on a broad range of topics for all skill levels.

**support.sas.com/saspress**

## SAS® Documentation

To successfully implement applications using SAS software, companies in every industry and on every continent all turn to the one source for accurate, timely, and reliable information: SAS documentation. We currently produce the following types of reference documentation to improve your work experience:

- Online help that is built into the software.
- Tutorials that are integrated into the product.
- Reference documentation delivered in HTML and PDF – **free** on the Web.
- Hard-copy books.

**support.sas.com/publishing**

## SAS® Publishing News

Subscribe to SAS Publishing News to receive up-to-date information about all new SAS titles, author podcasts, and new Web site features via e-mail. Complete instructions on how to subscribe, as well as access to past issues, are available at our Web site.

**support.sas.com/spn**

§sas | THE POWER TO KNOW®