Chapter 36

The LIFEREG Procedure

Chapter Table of Contents

Chapter 36
The LIFEREG Procedure

Overview

The LIFEREG procedure fits parametric models to failure time data that can be right, left, 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, loglogistic, and gamma distributions.

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

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

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), β is a vector of unknown regression parameters, σ is an unknown scale parameter, and ϵ is a vector of errors assumed to come from a known distribution (such as the standard normal distribution). The distribution may 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 using 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 $\exp(\mathbf{x}'\beta)$, where $\mathbf{x}$ is the vector of covariate values and β 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 $\mathbf{x}$, the event time is $T = \exp(\mathbf{x}'\beta)T_0$. If $y = \log(T)$ and $y_0 = \log(T_0)$, then

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

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

In terms of survival or exceedance probabilities, this model is

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

The probability on the left-hand side of the equal sign is evaluated given the value $\mathbf{x}$ 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 Distribution Function evaluated at $\exp(-\mathbf{x}'\beta)t$.

Usually, an intercept parameter and a scale parameter are allowed in the model. 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 power the event time, respectively. That is, if

$$\log(T) = \mu + \sigma \log(T_0)$$

then

$$T = \exp(\mu) T_0^{\sigma}$$

Although it is possible to fit these models to the original response variable 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 only be used if you want to fit a distribution appropriate for the untransformed response, the extreme value instead of the Weibull, for example.

The 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. Likewise, values that are extremely large even after the log transformation have a strong influence in fitting the extreme value (Weibull) and normal distributions. 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 using the observed information matrix. In small samples, these approximations may be poor. Refer to Lawless (1982) 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 σ. Therefore, it may be more accurate to construct confidence intervals for $\log(\sigma)$ and transform these into confidence intervals for σ. 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. Refer to Cox and Oakes (1984) for a discussion of the merits of some possible test methods including score, Wald, and likelihood ratio tests. It is believed that likelihood ratio tests are generally more reliable in 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 non-censored 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 (Tobin 1958) also fits into this class of models but with data usually censored on the left. The data considered by Tobin 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

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

which is a regression model with left censoring.

Getting Started

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, censor is a binary variable indicating censored values, and x1 and x2 are two linearly independent variables. The following statements perform a typical accelerated failure time model analysis. Note that no higher-order effects such as interactions are allowed in the covariables list.

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

PROC LIFEREG can operate on interval-censored data. The model syntax for specifying the censored interval is

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

You can also express the response with *events/trials* syntax, 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). The first five records of the data set headache are shown below.

Obs	minutes	group	censor
1	11	1	0
2	12	1	0
3	19	1	0
4	19	1	0
5	19	1	0

Figure 36.1. Headache Data

The following statements invoke the LIFEREG procedure:

```
proc lifereg;
   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 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 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 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.

```
                        The LIFEREG Procedure

                       Class Level Information

                     Name       Levels    Values

                     group           2    1 2


                          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
              Name of Distribution              WEIBULL
              Log Likelihood                -9.37930239
```

Figure 36.2. Model Fitting Information from the LIFEREG Procedure

Figure 36.2 displays the class level information and model fitting information. There are 30 noncensored 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 different models.

```
                           The LIFEREG Procedure

                      Analysis of Parameter Estimates

                                  Standard
Variable   DF   Estimate      Error Chi-Square Pr > ChiSq Label

Intercept   1    3.30912    0.05885  3161.7000     <.0001 Intercept
group       1                           6.0540     0.0139
            1   -0.19330    0.07856     6.0540     0.0139 1
            0          0          0      .          .     2
Scale       1    0.21219    0.03036                       Extreme value scale
```

Figure 36.3. Model Parameter Estimates from the LIFEREG Procedure

The table of parameter estimates is displayed in Figure 36.3. 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

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

Note that the Weibull shape parameter for this model is the reciprocal of the extreme value scale parameter estimate shown in Figure 36.3 ($1/0.21219 = 4.7128$).

The following statements produce a graph of the cumulative distribution values versus the variable minutes. The LEGEND1 statement defines the appearance of the legend that displays on the plot. The two AXIS statements define the appearance of the plot axes. The SYMBOL statements control the plotting symbol, color, and method of smoothing.

```
legend1 frame cframe=ligr cborder=black
   position=center value=(justify=center);

axis1 label=(angle=90 rotate=0 'Estimated CDF') minor=none;
axis2 minor=none;

symbol1 c=white i=spline;
symbol2 c=yellow i=spline;

proc sort data=new;
   by prob;

proc gplot data=new;
   plot prob*minutes=group/ frame cframe=ligr
       legend=legend1 vaxis=axis1 haxis=axis2;
run;
```

The SORT procedure sorts the data set new by the variable prob. Then the GPLOT procedure plots the variable prob versus the variable minutes using the grouping variable as the identification variable. The LEGEND=, VAXIS=, and HAXIS= options specify the previously defined legend and axis statements.

Figure 36.4 displays the estimated cumulative distribution function for each group.

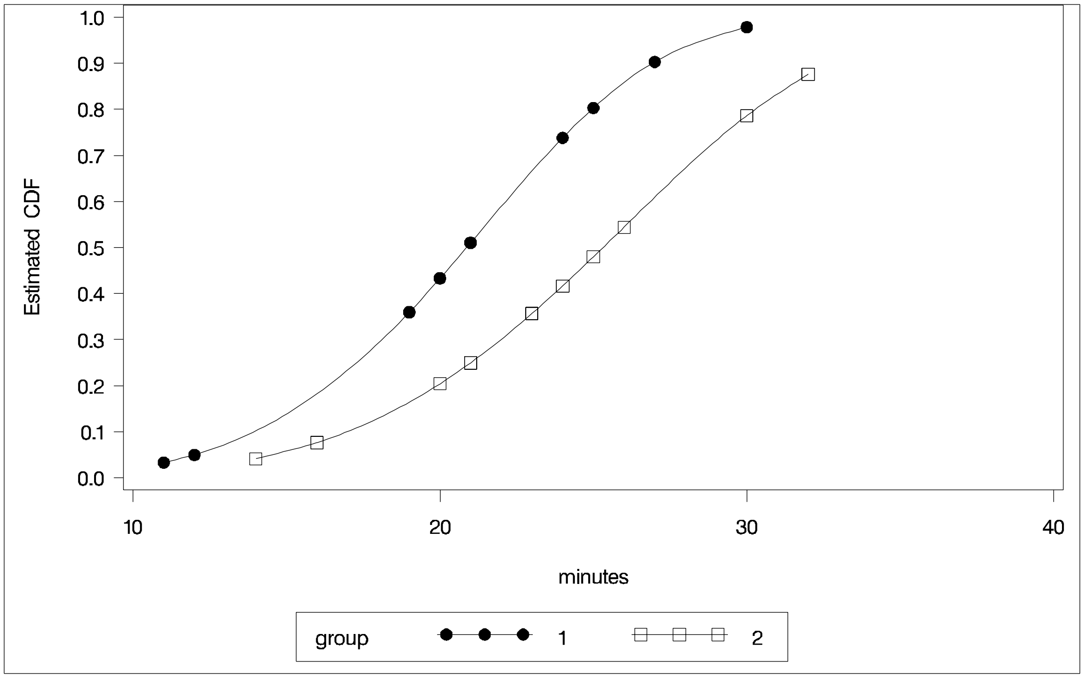

Figure 36.4. Plot of the Estimated Cumulative Distribution Function

Syntax

The following statements are available in PROC LIFEREG.

PROC LIFEREG < *options* > **;**
  MODEL *response=independents* < **/** *options* > **;**
  BY *variables* **;**
  CLASS *variables* **;**
  OUTPUT < **OUT=***SAS-data-set* >
    *keyword***=***name* < . . . *keyword***=***name* >
    < *options* > **;**
  WEIGHT *variable* **;**

The PROC LIFEREG statement invokes the procedure. The MODEL statement is required and specifies the variables used in the regression part of the model as well as the distribution used for the error, or random, component of the model. Only main effects can be specified in the MODEL statements. Interaction terms involving CLASS variables, allowed in the GLM procedure, are not available in PROC LIFEREG. Initial values can be specified in the MODEL statement. 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 containing predicted values and residuals.

PROC LIFEREG Statement

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

The PROC LIFEREG statement invokes the procedure. 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.

NOPRINT

suppresses the display of the output. Note that this option temporarily disables the Output Delivery System (ODS). For more information, see Chapter 15, "Using the Output Delivery System."

ORDER=DATA | FORMATTED | FREQ | INTERNAL

specifies the sorting order for the levels of the classification variables (specified in the CLASS statement). This ordering determines which parameters in the model correspond to each level in the data. The following table illustrates how PROC LIFEREG interprets values of the ORDER= option.

Value of ORDER=	Levels Sorted By
DATA	order of appearance in the input data set
FORMATTED	formatted value
FREQ	descending frequency count; levels with the most observations come first in the order
INTERNAL	unformatted value

By default, ORDER=FORMATTED. For FORMATTED and INTERNAL, the sort order is machine dependent. For more information on sorting order, refer to the chapter titled "The SORT Procedure" in the *SAS Procedures Guide*.

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 1784 for a detailed description of the contents of the OUTEST= data set. This data set is not created if class variables are used.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC LIFEREG to obtain separate analyses on observations in groups 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 your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING 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 using the DATASETS procedure.

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide*.

CLASS Statement

CLASS *variables* **;**

Variables that are classification variables rather than quantitative numeric variables must be listed in the CLASS statement. For each explanatory variable listed in the CLASS statement, indicator variables are generated for the levels assumed by the CLASS variable. If you use a CLASS statement, you cannot output parameter estimates to the OUTEST= data set (you can output them to a data set via ODS). If the CLASS statement is used, it must appear before any of the MODEL statements.

MODEL Statement

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

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

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

Multiple MODEL statements can be used with one invocation of the LIFEREG procedure. The optional *label* is used to label the model estimates in the output SAS data set.

The first MODEL syntax allows 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 variables following the equal sign are the covariates in the model. No higher order effects, such as interactions, are allowed in the covariables list; only variable names are allowed to appear in this list. However, a class variable can be used as a main effect, 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 are

```
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.

The following options can appear in the MODEL statement.

Task	Option
Model specification	
specify distribution type for failure time	DISTRIBUTION=
request no log transformation of response	NOLOG
initial estimate for intercept term	INTERCEPT=
hold intercept term fixed	NOINT
initial estimates for regression parameters	INITIAL=
initialize scale parameter	SCALE=
hold scale parameter fixed	NOSCALE
initialize first shape parameter	SHAPE1=
hold first shape parameter fixed	NOSHAPE1
Model fitting	
set convergence criterion	CONVERGE=
set maximum iterations	MAXITER=
set tolerance for testing singularity	SINGULAR=
Output	
display estimated correlation matrix	CORRB
display estimated covariance matrix	COVB
display iteration history, final gradient, and second derivative matrix	ITPRINT

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=0.001.

CONVG=*number*

sets the relative Hessian convergence criterion. The value of *number* 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 *number*, where **g** is the gradient vector, **H** is the Hessian matrix for the model parameters, and f is the log-likelihood function. If tc is greater than *number*, a warning that the relative Hessian convergence criterion has been exceeded is printed. 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 1780 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, 1982, p. 240). The two parameter gamma distribution is not available in PROC LIFEREG.
LLOGISTIC	a loglogistic 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 distribution options above 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 gamma
GAMMA	Yes	Generalized 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 1778 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, 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.

MAXITER=*value*

sets the maximum allowable number of iterations during the model estimation. By default, MAXITER=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 LIFEREG models the log of the response variable for the GAMMA, LLOGISTIC, LOGNORMAL, and WEIBULL 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 SHAPE= value is specified, SHAPE1 is fixed at a value that depends on the DISTRIBUTION type.

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 "Supported Distributions" section on page 1780 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 number 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 to the OUTPUT statement. These new variables contain fitted values and estimated quantiles. If you want to create a permanent SAS data set, you must specify a two-level name (refer to *SAS Language Reference: Concepts* for more information on permanent SAS data sets). Each OUTPUT statement applies to the preceding MODEL statement. See Example 36.1 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 DATAn 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 specifies an indicator variable to signal censoring. The variable takes on the value 1 if the observation is censored; otherwise, it is 0.

CDF specifies a variable to contain the estimates of the cumulative distribution function evaluated at the observed response. See the "Predicted Values" section on page 1783 for more information.

CONTROL specifies a variable in the input data set to control the estimation of quantiles. See Example 36.1 for an illustration. If the specified variable has the value of 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.

PREDICTED | P 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 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. The QUANTILES option can be specified as follows.

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`**

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.

STD_ERR | STD specifies a variable to contain the estimates of the standard errors of the estimated quantiles or $\mathbf{x}'\mathbf{b}$. If the response used in the MODEL statement is a binomial response, then these are the standard errors of $\mathbf{x}'\mathbf{b}$. 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 36.1 for such a transformation.

XBETA specifies a variable to contain the computed value of $\mathbf{x}'\mathbf{b}$, where $\mathbf{x}$ is the covariate vector and $\mathbf{b}$ is the vector of parameter estimates.

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

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.

Main Effects

Unlike the GLM procedure, only main effect terms are allowed in the model specification. For numeric variables, this 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 using ordinary least squares (OLS) 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 may be excessively influenced by the extreme observations, causing an overflow or convergence problems. See Example 36.3 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 INITIAL=, SCALE=, and SHAPE= options to set initial values of the intercept, scale, and shape 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 nonclass variables included in the model before any class 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.

A composite chi-square test statistic is computed for each class variable, testing whether there is any effect from any of the levels of the variable. This statistic is computed as a quadratic form in the appropriate parameter estimates using the corresponding submatrix of the asymptotic covariance matrix estimate. 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.

Model Specifications

LIFEREG procedure

Suppose there are n observations from the model $\mathbf{y} = \mathbf{X}\beta + \sigma\epsilon$, where $\mathbf{X}$ is an $n \times k$ matrix of covariate values (including the intercept), $\mathbf{y}$ is a vector of responses, and ϵ is a vector of errors with survival distribution 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 ϵ_i is a component of the error vector. Then, if all the responses are observed, the log likelihood, L, can be written as

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

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

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

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

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

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

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 ith observation. In this case, $P_i = 1 - F(-\mathbf{x}_i'\beta)$. For the symmetric distributions, logistic and normal, this is the same as $F(\mathbf{x}_i'\beta)$. Additional information on 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 **I**, which is the information matrix of second derivatives of L with respect to the parameters evaluated at the final parameter estimates. If **I** is not positive definite, a positive definite submatrix of **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 each distribution, the baseline survival distribution function (S) and the probability density function(f) are listed for the additive random disturbance. These distributions apply when the log of the response is modeled (this is the default analysis). The corresponding survival distribution function (G) and its density function (g) are given for the untransformed baseline distribution. For example, for the WEIBULL distribution, $S(w)$ and $f(w)$ are the baseline survival distribution function and the probability density function for the extreme value distribution (the log of the response) while $G(t)$ and $g(t)$ are the survival distribution function and probability distribution 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 σ is not, in general, the standard deviation of the baseline distribution.

Additionally, it is worth mentioning that, 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. μ = Intercept and σ = 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

(with $\mu = 0$, $\sigma = 1$)

$$S(w) = \begin{cases} \frac{\Gamma\left(\delta^{-2}, \delta^{-2}\exp(\delta w)\right)}{\Gamma(\delta^{-2})} & \text{if } \delta > 0 \\ 1 - \frac{\Gamma\left(\delta^{-2}, \delta^{-2}\exp(\delta w)\right)}{\Gamma(\delta^{-2})} & \text{if } \delta < 0 \end{cases}$$

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

$$G(t) = \begin{cases} \frac{\Gamma\left(\delta^{-2}, \delta^{-2}t^{\delta}\right)}{\Gamma(\delta^{-2})} & \text{if } \delta > 0 \\ 1 - \frac{\Gamma\left(\delta^{-2}, \delta^{-2}t^{\delta}\right)}{\Gamma(\delta^{-2})} & \text{if } \delta < 0 \end{cases}$$

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

where $\Gamma(a)$ denotes the complete gamma function, $\Gamma(a, z)$ denotes the incomplete gamma function, and δ is a free shape parameter. The δ parameter is referred to as Shape by PROC LIFEREG. Refer to Lawless, 1982, p.240 and Klein and Moeschberger, 1997, p.386 for a description of the generalized gamma distribution.

Loglogistic

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

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

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

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

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 Φ 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

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

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

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, $\mathbf{x}$, the expected value of the log response is not always $\mathbf{x}'\beta$.

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 pth quantile of the log response, y_p, is given by

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

where w_p is the pth 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 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

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

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

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \hat{w}_p \\ \hat{\sigma}\frac{\partial w_p}{\partial \delta} \end{bmatrix}$$

where δ 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 $\exp(\hat{y}_p)$STD. It may 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 36.1 for a 90% confidence interval of the response constructed by exponentiating a confidence interval for the log response.

The variable, CDF, is computed as

$$\mathrm{CDF}_i = F(w_i)$$

where the residual

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

and F is the baseline cumulative distribution function.

OUTEST= Data Set

The OUTEST= data set contains parameter estimates and the log likelihood for the specified models. A set of observations is created for each MODEL statement specified. You can specify a label in the MODEL statement to distinguish between the estimates for different MODEL statements. 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 is not created if there are any CLASS variables in any models. If created, this 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 of length 8 containing the label of the MODEL statement, if present. Otherwise, the variable's value is blank.

NAME a character variable of length 8 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 of length 8 containing the type of the observation, either PARMS for parameter estimates or COV for covariance estimates

DIST a character variable of length 8 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.

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.

Displayed Output

For each model, PROC LIFEREG displays

- the name of the Data Set
- the name of the Dependent Variable
- the name of the Censoring Variable
- the Censoring Value(s) that indicate a censored observation
- the number of Noncensored and Censored Values
- the final estimate of the maximized log likelihood
- the iteration history and the Last Evaluation of the Gradient and Hessian if the ITPRINT option is specified (not shown)

For each explanatory variable in the model, the LIFEREG procedure displays

- the name of the Variable
- the degrees of freedom (DF) associated with the variable in the model
- the Estimate of the parameter
- the standard error (Std Err) estimate from the observed information matrix
- an approximate chi-square statistic for testing that the parameter is zero (the class variables also have an overall chi-square test statistic computed that precedes the individual level parameters)
- the probability of a larger chi-square value (Pr>Chi)
- the Label of the variable or, if the variable is a class level, the Value of the class variable

If there are constrained parameters in the model, such as the scale or intercept, then PROC LIFEREG displays a Lagrange multiplier test for the constraint.

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 in the following table. For more information on ODS, see Chapter 15, "Using the Output Delivery System."

Table 36.1. ODS Tables Produced in PROC LIFEREG

ODS Table Name	Description	Statement	Option
ClassLevels	Class variable levels	CLASS	default*
ConvergenceStatus	Convergence status	MODEL	default
CorrB	Parameter estimate correlation matrix	MODEL	CORRB
CovB	Parameter estimate covariance matrix	MODEL	COVB
IterHistory	Iteration history	MODEL	ITPRINT
LagrangeStatistics	Lagrange statistics	MODEL	NOINT | NOSCALE
LastGrad	Last Evaluation of the Gradient	MODEL	ITPRINT
LastHess	Last Evaluation of the Hessian	MODEL	ITPRINT
ParameterEstimates	Parameter estimates	MODEL	default
ModelInfo	Model information	MODEL	default

* Depends on data.

Examples

Example 36.1. 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 requested 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 the 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 36.1.1 through Output 36.1.5.

```
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;

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

proc print data=models;
   id _model_;
   title 'fitted models';
run;

data out1;
   set out;
   ltime=log(predtime);
   stde=std/predtime;
   upper=exp(ltime+1.64*stde);
   lower=exp(ltime-1.64*stde);
proc print;
   id temp;
   title 'quantile estimates and confidence limits';
run;
```

Output 36.1.1. 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

Output 36.1.2. Motorette Failure: Model A

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
Missing Values	2
Name of Distribution	WEIBULL
Log Likelihood	-22.95148315

Analysis of Parameter Estimates

Variable	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq	Label
Intercept	1	-11.89122	1.96551	36.6019	<.0001	Intercept
z	1	9.03834	0.90599	99.5239	<.0001	
Scale	1	0.36128	0.07950			Extreme value scale

Output 36.1.3. Motorette Failure: Model B

```
                              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
                    Missing Values                          2
                    Name of Distribution              LNORMAL
                    Log Likelihood               -24.47381031


                          Analysis of Parameter Estimates

                                  Standard
         Variable   DF   Estimate      Error Chi-Square Pr > ChiSq Label

         Intercept   1  -10.47056    2.77192    14.2685     0.0002 Intercept
         z           1    8.32208    1.28412    42.0001     <.0001
         Scale       1    0.60403    0.11073                       Normal scale
```

Output 36.1.4. Motorette Failure: Fitted Models

```
                                          fitted models

_MODEL_   _NAME_      _TYPE_   _DIST_     _STATUS_     _LNLIKE_   Intercept       time        z       _SCALE_   _SHAPE1_

   A      time        PARMS    WEIBULL   0 Converged   -22.9515    -11.8912    -1.0000    9.03834    0.36128        .
   A      Intercept   COV      WEIBULL   0 Converged   -22.9515      3.8632   -11.8912   -1.77878    0.03448        .
   A      z           COV      WEIBULL   0 Converged   -22.9515     -1.7788     9.0383    0.82082   -0.01488        .
   A      Scale       COV      WEIBULL   0 Converged   -22.9515      0.0345     0.3613   -0.01488    0.00632        .
   B      time        PARMS    LNORMAL   0 Converged   -24.4738    -10.4706    -1.0000    8.32208    0.60403        .
   B      Intercept   COV      LNORMAL   0 Converged   -24.4738      7.6835   -10.4706   -3.55566    0.03267        .
   B      z           COV      LNORMAL   0 Converged   -24.4738     -3.5557     8.3221    1.64897   -0.01285        .
   B      Scale       COV      LNORMAL   0 Converged   -24.4738      0.0327     0.6040   -0.01285    0.01226        .
```

Output 36.1.5. Motorette Failure: Quantile Estimates and Confidence Limits

```
                                quantile estimates and confidence limits

   temp   time   censor   control      z       _PROB_   PREDTIME      STD       ltime      stde       upper       lower

    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 36.2. 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. Refer to 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

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

where $f_{\mathrm{Y}}(y)$ 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:

$$\mathrm{Y}_i^* = \mathbf{x}_i'\beta + \epsilon_i$$

where ϵ_i is a normal error term with zero mean and standard deviation σ.

Define the censored random variable Y_i as

$$\begin{aligned} \mathrm{Y}_i &= 0 \quad \text{if} \quad \mathrm{Y}_i^* \leq 0 \\ \mathrm{Y}_i &= \mathrm{Y}_i^* \quad \text{if} \quad \mathrm{Y}_i^* > 0 \end{aligned}$$

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 $\mathbf{x}_i'\beta$ and you can compute values of the mean for different settings of $\mathbf{x}_i$ by specifying XBETA=*variable-name* in an OUTPUT statement. Estimates of $\mathbf{x}_i'\beta$ 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

$$E(\mathrm{Y}_i) = \Phi\left(\frac{\mathbf{x}_i'\beta}{\sigma}\right)(\mathbf{x}_i'\beta + \sigma\lambda_i)$$

where

$$\lambda_i = \frac{\phi(\mathbf{x}_i'\beta/\sigma)}{\Phi(\mathbf{x}_i'\beta/\sigma)}$$

ϕ and Φ represent the normal probability density and cumulative distribution functions.

The following table shows a subset of the Mroz (1987) data set. In this 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 data above. 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 using Yrs_Ed and Yrs_Exp as covariates. You will need the estimated standard deviation of the normal distribution to compute the predicted values of the censored distribution from the formulas above. The data set OUTEST contains the standard deviation estimate in a variable named _SCALE_. You also need estimates of $\mathbf{x}_i'\beta$. 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 36.2.1 shows the results of the model fit. These tables show parameter estimates for the uncensored, or latent variable, distribution.

Output 36.2.1. 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
Name of Distribution	NORMAL
Log Likelihood	-74.9369977

Analysis of Parameter Estimates

Variable	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq	Label
Intercept	1	-5598.6	2850.2	3.8583	0.0495	Intercept
Yrs_Ed	1	373.14771	191.88717	3.7815	0.0518	
Yrs_Exp	1	63.33711	38.36317	2.7258	0.0987	
Scale	1	1582.9	442.67318			Normal scale

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 $\mathbf{x}_i'\beta$.

```
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;

proc print data=predict noobs label;
   var hours lower yrs: xbeta predict;
run;
```

Output 36.2.2 shows the original variables, the predicted means of the uncensored distribution, and the predicted means of the censored distribution.

Output 36.2.2. 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 36.3. 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 data set and request a Weibull regression model be fit to the data.

```
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
;
run;

proc print;
run;

title 'OLS (default) initial values';
proc lifereg data=raw;
   model x*censor(1) = c1 / distribution = weibull itprint;
run;
```

Output 36.3.1 shows the data set contents.

Output 36.3.1. 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

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

```
WARNING: Convergence not attained in 50 iterations.
WARNING: The procedure is continuing but the validity of the model
         fit is questionable.
```

The first line (iter=0) of the iteration history table, in Output 36.3.2, shows the default initial ordinary least squares (OLS) estimates of the parameters.

Output 36.3.2. Initial Least Squares

OLS (default) initial values

Iter	Ridge	Loglike	Intercept	c1	Scale
0	0	-22.891088	3.2324769714	0.0020664542	0.3995754195

The log logistic distribution is more robust to large values of the response than the Weibull, 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 36.3.3

Output 36.3.3. Estimates from the Log Logistic Distribution

```
                              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
                   Name of Distribution            LLOGISTC
                   Log Likelihood              12.093136846


                        Analysis of Parameter Estimates

                             Standard
   Variable   DF   Estimate      Error Chi-Square Pr > ChiSq Label

   Intercept   1    2.89828    0.03179  8309.4488     <.0001 Intercept
   c1          1    0.15921    0.01327   143.8537     <.0001
   Scale       1    0.04979    0.01218                       Logistic scale
```

The following statements re-fit the Weibull model 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 36.3.4 shows that the convergence problem has been solved by specifying different initial values.

Output 36.3.4. Final Estimates from the Weibull Distribution

```
                              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
                   Name of Distribution             WEIBULL
                   Log Likelihood              11.232023272


 Algorithm converged.


                        Analysis of Parameter Estimates

                          Standard
Variable   DF   Estimate      Error Chi-Square Pr > ChiSq Label

Intercept   1    2.96986    0.03264  8278.8602     <.0001 Intercept
c1          1    0.14346    0.01652    75.4316     <.0001
Scale       1    0.08437    0.01887                       Extreme value scale
```

References

Allison, P.D. (1995) *Survival Analysis Using the SAS System: A Practical Guide*, Cary, NC: SAS Institute.

Cox, D.R. (1972), “Regression Models and Life Tables (with discussion),” *Journal of the Royal Statistical Society, Series B*, 34, 187–220.

Cox, D.R. and Oakes, D. (1984), *Analysis of Survival Data*, London: Chapman and Hall.

Elandt-Johnson, R.C. and Johnson, N.L. (1980), *Survival Models and Data Analysis*, New York: John Wiley & Sons, Inc.

Green, W.H. (1993) *Econometric Analysis, 2nd Edition*, New York: Cambridge University Press.

Gross, A.J. and Clark, V.A. (1975), *Survival Distributions: Reliability Applications in the Biomedical Sciences*, New York: John Wiley & Sons, Inc.

Kalbfleisch, J.D. and Prentice, R.L. (1980), *The Statistical Analysis of Failure Time Data*, New York: John Wiley & Sons, Inc.

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

Lawless, J.E. (1982), *Statistical Models and Methods for Lifetime Data*, New York: John Wiley & Sons, Inc.

Lee, E.T. (1980), *Statistical Methods for Survival Data Analysis*, Belmont, CA: Lifetime Learning Publications.

Maddala, G.S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, New York: Cambridge University Press.

Mroz, T.A. (1987) “The Sensitivity of an Empirical Model of Married Women’s Work to Economic and Statistical Assumptions,” *Econometrica* 55, 765–799.

Rao, C.R. (1973), *Linear Statistical Inference and Its Applications*, New York: John Wiley & Sons, Inc.

Tobin, J. (1958), “Estimation of Relationships for Limited Dependent Variables,” *Econometrica*, 26, 24–36.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

SAS/STAT® User's Guide, Version 8

ISBN 1–58025–494–2

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999

The Institute is a private company devoted to the support and further development of its software and related services.