Chapter 49

The PHREG Procedure

Chapter Table of Contents

Chapter 49
The PHREG Procedure

Overview

The analysis of survival data requires special techniques because the data are almost always incomplete, and familiar parametric assumptions may be unjustifiable. Investigators follow subjects until they reach a prespecified endpoint (for example, death). However, subjects sometimes withdraw from a study, or the study is completed before the endpoint is reached. In these cases, the survival times (also known as failure times) are *censored*; subjects survived to a certain time beyond which their status is unknown. The noncensored survival times are referred to as *event* times. Methods for survival analysis must account for both censored and noncensored data.

There are many types of models that have been used for survival data. Two of the more popular types of models are the accelerated failure time model (Kalbfleisch and Prentice 1980) and the Cox proportional hazards model (Cox 1972). Each has its own assumptions on the underlying distribution of the survival times. Two closely related functions often used to describe the distribution of survival times are the survivor function and the hazard function (see the section "Failure Time Distribution" on page 2593 for definitions).

The accelerated failure time model assumes a parametric form for the effects of the explanatory variables and usually assumes a parametric form for the underlying survivor function. Cox's proportional hazards model also assumes a parametric form for the effects of the explanatory variables, but it allows an unspecified form for the underlying survivor function.

The PHREG procedure performs regression analysis of survival data based on the Cox proportional hazards model. Cox's semiparametric model is widely used in the analysis of survival data to explain the effect of explanatory variables on survival times.

The survival time of each member of a population is assumed to follow its own hazard function, $h_i(t)$, expressed as

$$h_i(t) = h(t;\mathbf{z}_i) = h_0(t)\exp(\mathbf{z}_i'\boldsymbol{\beta})$$

where $h_0(t)$ is an arbitrary and unspecified baseline hazard function, $\mathbf{z}_i$ is the vector of measured explanatory variables for the *i*th individual, and $\boldsymbol{\beta}$ is the vector of unknown regression parameters associated with the explanatory variables. The vector $\boldsymbol{\beta}$ is assumed to be the same for all individuals.

The survivor function can be expressed as

$$S(t;\mathbf{z}_i) = [S_0(t)]^{\exp(\mathbf{z}_i'\boldsymbol{\beta})}$$

where $S_0(t) = \exp(-\int_0^t h_0(u)du)$ is the baseline survivor function.

To estimate $\boldsymbol{\beta}$, Cox (1972, 1975) introduced the partial likelihood function, which eliminates the unknown baseline hazard $h_0(t)$ and accounts for censored survival times.

The partial likelihood of Cox also allows time-dependent explanatory variables. An explanatory variable is time-dependent if its value for any given individual can change over time. Time-dependent variables have many useful applications in survival analysis. You can use a time-dependent variable to model the effect of subjects changing treatment groups. Or you can include time-dependent variables such as blood pressure or blood chemistry measures that vary with time during the course of a study. You can also use time-dependent variables to test the validity of the proportional hazards model.

An alternative way to fit models with time-dependent explanatory variables is to use the counting process style of input. The counting process formulation allows PROC PHREG to fit a superset of the Cox model, known as the multiplicative hazards model. This extension also includes multiple events per subject, time-dependent strata, and left truncation of failure times. The theory of these models is based on the counting process pioneered by Andersen and Gill (1982), and the model is often referred to as the Andersen-Gill Model.

The population under study may consist of a number of subpopulations, each of which has its own baseline hazard function. PROC PHREG performs a stratified analysis to adjust for such subpopulation differences. Under the stratified model, the hazard function for the *j*th individual in the *i*th stratum is expressed as

$$h_{ij}(t) = h_{i0}(t)\exp(\mathbf{z}_{ij}'\boldsymbol{\beta})$$

where $h_{i0}(t)$ is the baseline hazard function for the *i*th stratum, and $\mathbf{z}_{ij}$ is the vector of explanatory variables for the *j*th individual. The regression coefficients are assumed to be the same for all individuals across all strata.

Ties in the failure times may arise when the time scale is genuinely discrete or when survival times generated from the continuous-time model are grouped into coarser units. The PHREG procedure includes four methods of handling ties. The *discrete* logistic model is available for discrete time-scale data. The other three methods apply to continuous time-scale data. The *exact* method computes the exact conditional probability under the model that the set of observed tied event times occurs before all the censored times with the same value or before larger values. *Breslow* and *Efron* methods provide approximations to the exact method.

Variable selection is a typical exploratory exercise in multiple regression when the investigator is interested in identifying important prognostic factors from a large number of candidate variables. The PHREG procedure provides four model selection methods: forward selection, backward elimination, stepwise selection, and best subset selection. The best subset selection method is based on the likelihood score statistic. This method identifies a specified number of best models containing one, two, three variables and so on, up to the single model containing all of the explanatory variables.

The PHREG procedure also enables you to

- include an offset variable in the model
- test linear hypotheses about the regression parameters
- perform conditional logistic regression analysis for matched case-control studies
- create a SAS data set containing survivor function estimates, residuals, and regression diagnostics
- create a SAS data set containing survival distribution estimates and confidence interval for the survivor function at each event time for a given realization of the explanatory variables

The remaining sections of this chapter contain information on how to use PROC PHREG, information on the underlying statistical methodology, and some sample applications of the procedure. The "Getting Started" section on page 2573 introduces PROC PHREG with two examples. The "Syntax" section on page 2577 describes the syntax of the procedure. The "Details" section on page 2593 summarizes the statistical techniques employed in PROC PHREG. The "Examples" section on page 2608 includes eight additional examples of useful applications. Experienced SAS/STAT software users may decide to proceed to the "Syntax" section, while other users may choose to read both the "Getting Started" and "Examples" sections before proceeding to "Syntax" and "Details."

Getting Started

PROC PHREG syntax is similar to that of the other regression procedures in the SAS System. For simple uses, only the PROC PHREG and MODEL statements are required.

Consider the following data from Kalbfleisch and Prentice (1980). Two groups of rats received different pretreatment regimes and then were exposed to a carcinogen. Investigators recorded the survival times of the rats from exposure to mortality from vaginal cancer. Four rats died of other causes, so their survival times are censored. Interest lies in whether the survival curves differ between the two groups.

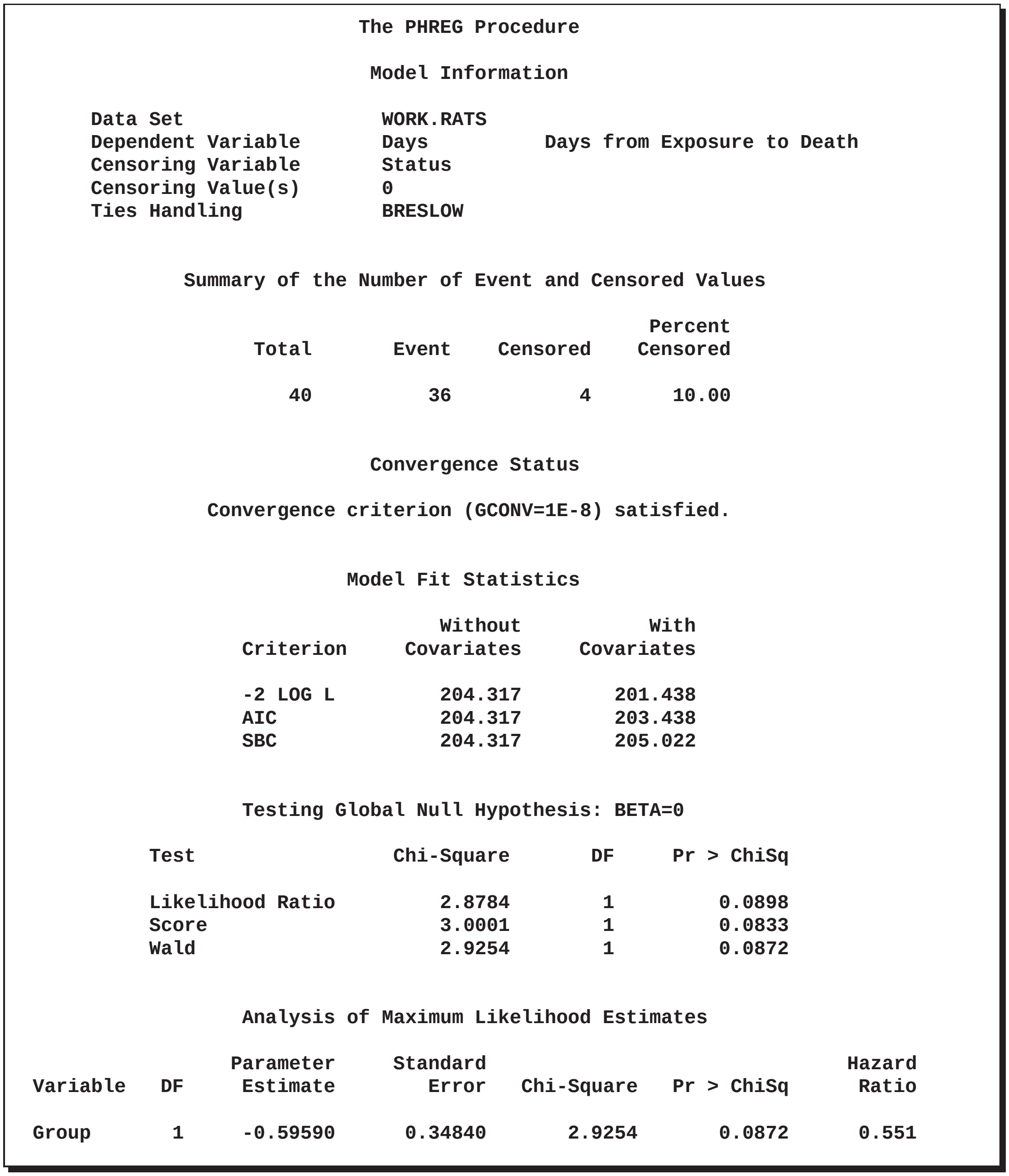
The data set Rats contains the variable Days (the survival time in days), the variable Status (the censoring indicator variable: 0 if censored and 1 if not censored), and the variable Group (the pretreatment group indicator).

```
data Rats;
   label Days  ='Days from Exposure to Death';
   input Days Status Group @@;
   datalines;
143 1 0   164 1 0   188 1 0   188 1 0
190 1 0   192 1 0   206 1 0   209 1 0
213 1 0   216 1 0   220 1 0   227 1 0
230 1 0   234 1 0   246 1 0   265 1 0
304 1 0   216 0 0   244 0 0   142 1 1
156 1 1   163 1 1   198 1 1   205 1 1
232 1 1   232 1 1   233 1 1   233 1 1
233 1 1   233 1 1   239 1 1   240 1 1
261 1 1   280 1 1   280 1 1   296 1 1
296 1 1   323 1 1   204 0 1   344 0 1
;
run;
```

In the MODEL statement, the response variable, Days, is crossed with the censoring variable, Status, with the value that indicates censoring enclosed in parentheses (0). The values of Days are considered censored if the value of Status is 0; otherwise, they are considered event times.

```
proc phreg data=Rats;
   model Days*Status(0)=Group;
run;
```

Results of the PROC PHREG analysis appear in Figure 49.1. Since Group takes only two values, the null hypothesis for no difference between the two groups is identical to the null hypothesis that the regression coefficient for Group is 0. All three tests in the "Testing Global Null Hypothesis: BETA=0" table (see the section "Testing the Global Null Hypothesis" on page 2597) suggest that the survival curves for the two pretreatment groups may not be the same. In this model, the hazards ratio (or risk ratio) for Group, defined as the exponentiation of the regression coefficient for Group, is the ratio of the hazard functions between the two groups. The estimate is 0.551, implying that the hazard function for Group=1 is smaller than that for Group=0. In other words, rats in Group=1 lived longer than those in Group=0.



```
                            The PHREG Procedure

                             Model Information

     Data Set                 WORK.RATS
     Dependent Variable       Days          Days from Exposure to Death
     Censoring Variable       Status
     Censoring Value(s)       0
     Ties Handling            BRESLOW


             Summary of the Number of Event and Censored Values

                                                     Percent
                   Total       Event    Censored    Censored

                      40          36           4       10.00


                             Convergence Status

               Convergence criterion (GCONV=1E-8) satisfied.


                           Model Fit Statistics

                                   Without           With
                  Criterion     Covariates     Covariates

                  -2 LOG L         204.317        201.438
                  AIC              204.317        203.438
                  SBC              204.317        205.022


                  Testing Global Null Hypothesis: BETA=0

          Test                 Chi-Square       DF     Pr > ChiSq

          Likelihood Ratio         2.8784        1         0.0898
          Score                    3.0001        1         0.0833
          Wald                     2.9254        1         0.0872


                  Analysis of Maximum Likelihood Estimates

                 Parameter     Standard                               Hazard
Variable   DF     Estimate        Error   Chi-Square   Pr > ChiSq      Ratio

Group       1     -0.59590      0.34840       2.9254       0.0872      0.551
```

Figure 49.1. Comparison of Two Survival Curves

In this example, the comparison of two survival curves is put in the form of a proportional hazards model. This approach is essentially the same as the log-rank (Mantel-Haenszel) test. In fact, if there are no ties in the survival times, the likelihood score test in the Cox regression analysis is identical to the log-rank test. The advantage of the Cox regression approach is the ability to adjust for the other variables by including them in the model. For example, the present model could be expanded by including a variable that contains the initial body weights of the rats.

Next, consider a simple test of the validity of the proportional hazards assumption. The proportional hazards model for comparing the two pretreatment groups is given by the following:

$$h(t) = \begin{cases} h_0(t) & \text{if GROUP} = 0 \\ h_0(t)e^{\beta_1} & \text{if GROUP} = 1 \end{cases}$$

The ratio of hazards is e^{β_1}, which does not depend on time. If the hazard ratio changes with time, the proportional hazards model assumption is invalid. Simple forms of departure from the proportional hazards model can be investigated with the following time-dependent explanatory variable $x = x(t)$:

$$x(t) = \begin{cases} 0 & \text{if GROUP} = 0 \\ \log(t) - 5.4 & \text{if GROUP} = 1 \end{cases}$$

Here, $\log(t)$ is used instead of *t* to avoid numerical instability in the computation. The constant, 5.4, is the average of the logs of the survival times and is included to improve interpretability. The hazard ratio in the two groups then becomes $e^{\beta_1 - 5.4\beta_2}t^{\beta_2}$, where β_2 is the regression parameter for the time-dependent variable *x*. The term e^{β_1} represents the hazard ratio at the geometric mean of the survival times. A nonzero value of β_2 would imply an increasing $(\beta_2 > 0)$ or decreasing $(\beta_2 < 0)$ trend in the hazard ratio with time.

The MODEL statement in this analysis also includes the time-dependent explanatory variable X, which is defined within the procedure by the programming statement that follows the MODEL statement. At each event time, subjects in the risk set (those alive just before the event time) have their X values changed accordingly.

```
proc phreg data=Rats;
   model Days*Status(0)=Group X;
   X=Group*(log(Days) - 5.4);
run;
```

```
                              The PHREG Procedure

                    Analysis of Maximum Likelihood Estimates

                  Parameter     Standard                                 Hazard
Variable    DF     Estimate        Error   Chi-Square   Pr > ChiSq        Ratio

Group        1     -0.59976      0.34837       2.9639       0.0851        0.549
X            1     -0.22952      1.82489       0.0158       0.8999        0.795
```

Figure 49.2. A Simple Test of Trend in the Hazard Ratio

The analysis of the parameter estimates is displayed in Figure 49.2. The Wald chi-squared statistic for testing the null hypothesis that $\beta_2 = 0$ is 0.0158. The statistic is not statistically significant when compared to a chi-squared distribution with one degree of freedom ($p = 0.8999$). Thus, you can conclude that there is no evidence of an increasing or decreasing trend over time in the hazard ratio. See the “Examples” section beginning on page 2608 for additional illustrations of PROC PHREG usage.

Syntax

The following statements are available in PROC PHREG.

```
PROC PHREG < options > ;
    MODEL response < *censor(list) > = variables < /options > ;
    < programming statements >
    STRATA variable < (list) > < ...variable < (list) >>< /option > ;
    < label: > TEST equation1 < ,..., equationk >< /option > ;
    FREQ variable ;
    ID variables ;
    OUTPUT < OUT=SAS-data-set >
          < keyword=name... keyword=name >< /options > ;
    BASELINE < OUT=SAS-data-set >
          < COVARIATES=SAS-data-set >
          < keyword=name... keyword=name >< /options > ;
    BY variables ;
```

The PROC PHREG statement invokes the procedure. All other statements except the MODEL statement are optional. Items within < > are optional, and there is no required order for the statements following the PROC PHREG statement. The MODEL statement specifies the variables that define the survival time, the censoring variable, and the explanatory variables. The STRATA statement specifies a variable or set of variables defining the strata for the analysis. The TEST statement contains equations that define linear hypotheses concerning the model parameters. The ID statement specifies the variables with values that are used to label the observations in the OUTPUT data set. The OUTPUT and BASELINE statements create data sets containing the survival estimates. DATA step programming statements can be included to create time-dependent explanatory variables.

PROC PHREG Statement

```
PROC PHREG < options > ;
```

You can specify the following options in the PROC PHREG statement.

COVOUT

adds the estimated covariance matrix of the parameter estimates to the OUTEST= data set. The COVOUT option has no effect unless the OUTEST= option is specified.

DATA=*SAS-data-set*

names the SAS data set containing the data to be analyzed. If you omit the DATA= option, the procedure uses the most recently created SAS data set.

MULTIPASS

requests that, for each Newton-Raphson iteration, PROC PHREG recompiles the risk sets corresponding to the event times for the (start,stop) style of response and recomputes the values of the time-dependent variables defined by the programming statements for each observation in the risk sets. If the MULTIPASS option is not specified, PROC PHREG computes all risk sets and all the variable values and saves them into a utility file. The MULTIPASS option decreases required disk space at the expense of increased execution time; however, for very large data, it may actually save time since it is time consuming to write and read large utility files. This option has an effect only when the (start,stop) style of response is used or when there are time-dependent explanatory variables.

NOPRINT

suppresses all displayed output. Note that this option temporarily disables the Output Delivery System (ODS); see Chapter 15, “Using the Output Delivery System,” for more information.

NOSUMMARY

suppresses the display of the event and censored observation frequencies.

OUTEST=*SAS-data-set*

creates an output SAS data set that contains estimates of the regression coefficients. If you use the COVOUT option, the data set also contains the estimated covariance matrix of the parameter estimates. The data set includes

- any BY variables specified
- _TIES_, a character variable of length 8 with four possible values: BRESLOW, DISCRETE, EFRON, and EXACT. These are the four values of the TIES= option in the MODEL statement.
- _TYPE_, a character variable of length 8 with two possible values: PARMS for parameter estimates or COV for covariance estimates
- _STATUS_, a character variable indicating whether the estimates have converged
- _NAME_, a character variable containing the name of the TIME variable for the row of parameter estimates and the name of each explanatory variable to label the rows of covariance estimates
- one variable for each explanatory variable in the MODEL statement. In a forward, backward, or stepwise regression analysis, if an explanatory variable is not included in the final model, the corresponding parameter estimate and covariances are set to missing.
- _LNLIKE_, a numeric variable containing the last computed value of the log likelihood

SIMPLE
displays simple descriptive statistics (mean, standard deviation, minimum, and maximum) for each explanatory variable in the MODEL statement.

BASELINE Statement

BASELINE < **OUT=** *SAS-data-set* >< **COVARIATES=** *SAS-data-set* >
< *keyword=name ... keyword=name* >< */options* > **;**

The BASELINE statement creates a new SAS data set that contains the survivor function estimates at the event times of each stratum for every pattern of explanatory variable values ($\mathbf{x}$) given in the COVARIATES= data set. By default, the data set also contains the survivor function estimates corresponding to the means of the explanatory variables ($\mathbf{x} = \bar{\mathbf{z}}$) for each stratum. If you want only these estimates, you can omit the COVARIATES= option. No BASELINE data set is created if the counting process style of input is used or if the model contains a time-dependent variable.

The following list explains specifications in the BASELINE statement.

OUT=*SAS-data-set*
names the output BASELINE data set. If you omit the OUT= option, the data set is created and given a default name using the DATA*n* convention.

COVARIATES=*SAS-data-set*
names the SAS data set containing the set of explanatory variable values for which the survivor functions are estimated. There must be a corresponding variable in the COVARIATES= data set for each explanatory variable in the final model.

keyword=name
specifies the statistics included in the BASELINE data set and assigns names to the new variables that contain the statistics. 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 and the corresponding statistics are

LOGLOGS	log of the negative log of SURVIVAL
LOGSURV	log of SURVIVAL
LOWER | L	lower confidence limit for the survivor function
STDERR	standard error of the survivor function estimate
STDXBETA	standard error of the estimated linear predictor, $\sqrt{\mathbf{x}'\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})\mathbf{x}}$
SURVIVAL	survivor function estimate $\hat{S}(t) = [\hat{S}_0(t)]^{\exp(\mathbf{x}'\hat{\boldsymbol{\beta}})}$
UPPER | U	upper confidence limit for the survivor function
XBETA	estimate of the linear predictor, $\mathbf{x}'\hat{\boldsymbol{\beta}}$

The following options can appear in the BASELINE statement after a slash (/).

ALPHA=*value*

specifies the significance level of the confidence interval for the survivor function. The value must be between 0 and 1. The default is 0.05, which results in a 95% confidence interval.

CLTYPE=*method*

specifies the method used to compute the confidence limits for $S(t, \mathbf{z})$, the survivor function for a subject with a fixed covariate vector $\mathbf{z}$ at event time t. The CLTYPE= option can take the following values:

LOG specifies that the confidence limits for log($S(t, \mathbf{z})$) are to be computed using the normal theory approximation. The confidence limits for $S(t, \mathbf{z})$ are obtained by back-transforming the confidence limits for log($S(t, \mathbf{z})$). The default is CLTYPE=LOG.

LOGLOG specifies that the confidence limits for the log(-log($S(t, \mathbf{z})$)) are to be computed using normal theory approximation. The confidence limits for $S(t, \mathbf{z})$ are obtained by back-transforming the confidence limits for log(-log($S(t, \mathbf{z})$)).

NORMAL specifies that the confidence limits for $S(t, \mathbf{z})$ are to be computed directly using normal theory approximation.

METHOD=*method*

specifies the method used to compute the survivor function estimates. The two available methods are

CH | EMP specifies that the empirical cumulative hazard function estimate of the survivor function is to be computed; that is, the survivor function is estimated by exponentiating the negative empirical cumulative hazard function.

PL specifies that the product-limit estimate of the survivor function is to be computed. The default is METHOD=PL.

NOMEAN

excludes the survivor function estimates corresponding to the sample means of the explanatory variables.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC PHREG 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. The *variables* are one or more variables in the input data set.

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 PHREG 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 (in base SAS software).

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

FREQ Statement

FREQ *variable* < */option* > **;**

The *variable* in the FREQ statement identifies a variable (in the input data set) containing the frequency of occurrence of each observation. PROC PHREG treats each observation as if it appears *n* times, where *n* is the value of the FREQ variable for the observation. If not an integer, the frequency value is truncated to an integer. If the frequency value is missing, the observation is not used in the estimation of the regression parameters.

The following option can be specified in the FREQ statement after a slash (/):

NOTRUNCATE
NOTRUNC
specifies that frequency values are not truncated to integers.

ID Statement

ID *variables* **;**

The ID statement specifies additional variables to be placed in the OUT= data set created by the OUTPUT statement. Only variables in the input data set can be included in the ID statement.

MODEL Statement

MODEL *response* $<$ **censor (list)* $>$ *= variables* $<$ */options* $>$ **;**
MODEL *(t1, t2)* $<$ **censor(list)* $>$ *= variables* $<$ */options* $>$ **;**

The MODEL statement identifies the variables to be used as the failure time variables, the optional censoring variable, and the explanatory variables. Two forms of MODEL syntax can be specified; the first form allows one response variable, while the second form allows two variables for the counting process style of input (see the section "Counting Process Style of Input" on page 2595 for more information).

In the first MODEL statement, preceding the equal sign, is the name of the failure time variable. This can optionally be followed by an asterisk, the name of the censoring variable, and a list of censoring values (separated by blanks or commas if there is more than one) enclosed in parentheses. If the censoring variable takes on one of these values, the corresponding failure time is considered to be censored. The variables following the equal sign are the explanatory variables (sometimes called independent variables or covariates) for the model.

Instead of a single failure time variable, the second MODEL statement identifies a pair of failure time variables. Their names are enclosed in parentheses, and they signify the endpoints of a semi-closed interval $(t1, t2]$ during which the subject is at risk. If the censoring variable takes on one of the censoring values, the time $t2$ is considered to be censored.

The censoring variable and the explanatory variables must be numeric. The failure time variables must contain nonnegative values. Any observation with a negative failure time is excluded from the analysis, as is any observation with a missing value for any of the variables listed in the MODEL statement.

You can specify the following options in the MODEL statement.

Ties-Handling Option

TIES=*method*

specifies how to handle ties in the failure time. The TIES= option can take the following values:

BRESLOW uses the approximate likelihood of Breslow (1974). This is the default value.

DISCRETE replaces the proportional hazards model by the discrete logistic model

$$\frac{h(t;\mathbf{z})}{1-h(t;\mathbf{z})} = \frac{h_0(t)}{1-h_0(t)} \exp(\mathbf{z}'\boldsymbol{\beta})$$

where $h_0(t)$ and $h(t;\mathbf{z})$ are discrete hazard functions.

EFRON uses the approximate likelihood of Efron (1977).

EXACT computes the exact conditional probability under the proportional hazards assumption that all tied event times occur before censored times of the same value or before larger values. This is equivalent to summing all terms of the marginal likelihood for $\boldsymbol{\beta}$ that are consistent with the observed data (Kalbfleisch and Prentice 1980; DeLong, Guirguis, and So 1994).

The EXACT method may take a considerable amount of computer resources. If ties are not extensive, the EFRON and BRESLOW methods provide satisfactory approximations to the EXACT method for the continuous time-scale model. In general, Efron's approximation gives results that are much closer to the EXACT method results than Breslow's approximation does. If the time scale is genuinely discrete, you should use the DISCRETE method. The DISCRETE method is also required in the analysis of case-control studies when there is more than one case in a matched set. If there are no ties, all four methods result in the same likelihood and yield identical estimates. The default, TIES=BRESLOW, is the most efficient method when there are no ties.

Model-Specification Options

ENTRYTIME=*variable*
ENTRY=*variable*

specifies the name of the variable that represents the left truncation time. This option has no effect when the counting process style of input is specified. See the section "Left Truncation of Failure Times" on page 2604 for more information.

NOFIT

performs the global score test, which tests the joint significance of all the explanatory variables in the MODEL statement. No parameters are estimated. If the NOFIT option is specified along with other MODEL statement options, NOFIT takes precedence, and all other options are ignored except the TIES= option.

OFFSET=*name*

specifies the name of an offset variable, which is an explanatory variable with a regression coefficient fixed as one. This option can be used to incorporate risk weights for the likelihood function.

SELECTION=*method*

specifies the method used to select the model. The *method*s available are

BACKWARD | B requests backward elimination.

FORWARD | F requests forward selection.

NONE | N fits the complete model specified in the MODEL statement. This is the default value.

SCORE	requests best subset selection. It identifies a specified number of models with the highest score chi-squared statistic for all possible model sizes ranging from one explanatory variable to the total number of explanatory variables listed in the MODEL statement.
STEPWISE | S	requests stepwise selection.

For more information, see the section "Variable Selection Methods" on page 2604.

Model-Building Options

The following options enable you to provide additional specifications for the BACKWARD, FORWARD, SCORE, and STEPWISE model selection methods. They have no effect when SELECTION=NONE. Only the INCLUDE=, START=, STOP=, and BEST= options work with the SCORE method.

BEST=*n*

is used exclusively with the SCORE model selection method. The BEST=*n* option specifies that *n* models with the highest score chi-squared statistics are to be displayed for each model size. If the option is omitted and there are no more than ten explanatory variables, then all possible models are listed for each model size. If the option is omitted and there are more than ten explanatory variables, then the number of models selected for each model size is, at most, equal to the number of explanatory variables listed in the MODEL statement. See Example 49.2 on page 2616 for an illustration of the SCORE selection method and the BEST= option.

DETAILS

produces a detailed display at each step of the model-building process. It produces an "Analysis of Variables Not in the Model" table before displaying the variable selected for entry for FORWARD or STEPWISE selection. For each model fitted, it produces the "Analysis of Maximum Likelihood Estimates" table. See Example 49.1 on page 2608 for a discussion of these tables.

INCLUDE=*n*

includes the first *n* explanatory variables listed in the MODEL statement in every model. The value for *n* ranges from 1 to *s*, where *s* is the number of explanatory variables in the MODEL statement. The default value of *n* is 0.

MAXSTEP=*n*

specifies the maximum number of times the explanatory variables can move in and out of the model before the STEPWISE model-building process ends. The default value for *n* is twice the number of explanatory variables in the MODEL statement. The option has no effect for other model selection methods.

SEQUENTIAL

forces variables to be added to the model in the order specified in the MODEL statement or to be eliminated from the model in the reverse order specified in the MODEL statement.

SLENTRY=*value*
SLE=*value*

specifies the significance level (a value between 0 and 1) for entering an explanatory variable into the model in the FORWARD or STEPWISE method. For all variables not in the model, the one with the smallest *p*-value is entered if the *p*-value is less than or equal to the specified significance level. The default value is 0.05.

SLSTAY=*value*
SLS=*value*

specifies the significance level (a value between 0 and 1) for removing an explanatory variable from the model in the BACKWARD or STEPWISE method. For all variables in the model, the one with the largest *p*-value is removed if the *p*-value exceeds the specified significance level. The default value is 0.05.

START=*n*

begins the FORWARD, BACKWARD, or STEPWISE model selection process with the first *n* explanatory variables listed in the MODEL statement. The value for *n* ranges from 0 to *s*, where *s* is the total number of explanatory variables in the MODEL statement. The default value of *n* is *s* for the BACKWARD method and 0 for the FORWARD and STEPWISE methods. Note that START=*n* specifies only that the first *n* explanatory variables appear in the first model, while INCLUDE=*n* specifies that the first *n* explanatory variables be included in every model. For the SCORE method, START=*n* specifies that the smallest models contain *n* explanatory variables, where *n* ranges from 1 to *s*. The default value of *n* is 1.

STOP=*n*

specifies the maximum (FORWARD method) or minimum (BACKWARD method) number of explanatory variables to be included in the final model. The value for *n* ranges from 0 to *s*, where *s* is the number of explanatory variables in the MODEL statement. The default value of *n* is 0 for the BACKWARD method and *s* for the FORWARD method. For the SCORE method, STOP=*n* specifies that the largest models contain *n* explanatory variables, where *n* ranges from 1 to *s*. The default value of *n* is *s*. The STOP= option has no effect for the STEPWISE method.

STOPRES
SR

specifies that the addition and deletion of variables are to be based on the result of the likelihood score test for testing the joint significance of variables not in the model. This score chi-squared statistic is referred to as the residual chi-square. In the FORWARD method, the STOPRES option enters the explanatory variables into the model one at a time until the residual chi-square becomes insignificant (that is, until the *p*-value of the residual chi-square exceeds the SLENTRY= value). In the BACKWARD method, the STOPRES option removes variables from the model one at a time until the residual chi-square becomes significant (that is, until the *p*-value of the residual chi-square becomes less than the SLSTAY= value). The STOPRES option has no effect for the STEPWISE method.

Optimization Options

Four convergence criteria are allowed: ABSFCONV=, FCONV=, GCONV=, and XCONV=. If you specify more than one convergence criterion, the optimization is terminated as soon as one of the criteria is satisfied. If none of the criteria is specified, the default is GCONV=1E−8.

ABSFCONV=*value*

specifies the absolute function convergence criterion. Termination requires a small change in the objective function (log partial likelihood function) in subsequent iterations,

$$|l_k - l_{k-1}| < value$$

where l_k is the value of the objective function at iteration k.

CONVERGELIKE=*value*

is the same as specifying the ABSFCONV= option.

CONVERGEPARM=*value*

is the same as specifying the XCONV= option.

FCONV=*value*

specifies the relative function convergence criterion. Termination requires a small relative change in the objective function (log partial likelihood function) in subsequent iterations,

$$\frac{|l_k - l_{k-1}|}{|l_{k-1}| + 1\mathrm{E}-6} < value$$

where l_k is the value of the objective function at iteration k.

GCONV=*value*

specifies the relative gradient convergence criterion. Termination requires that the normalized prediction function reduction is small,

$$\frac{\mathbf{g}_k \mathbf{H}_k^{-1} \mathbf{g}_k}{|l_k| + 1\mathrm{E}-6} < value$$

where l_k is the log partial likelihood, $\mathbf{g}_k$ is the gradient vector (first partial derivatives of the log partial likelihood), and $\mathbf{H}_k$ is the negative Hessian matrix (second partial derivatives of the log partial likelihood), all at iteration k.

MAXITER=*n*

specifies the maximum number of iterations allowed. The default value for *n* is 25. If convergence is not attained in *n* iterations, the displayed output and all data sets created by PROC PHREG contain results that are based on the last maximum likelihood iteration.

RIDGING=ABSOLUTE | RELATIVE | NONE

specifies the technique to improve the log-likelihood when its value is worse than that of the previous step. For RIDGING=ABSOLUTE, the diagonal elements of the negative (expected) Hessian are inflated by adding the ridge value. For RIDGING=RELATIVE, the diagonal elements are inflated by the factor equal to 1 plus the ridge value. For RIDGING=NONE, the crude line-search method of taking half a step is used instead of ridging.

SINGULAR=*value*

specifies the singularity criterion for determining linear dependencies in the set of explanatory variables. The default value is 10^{-12}.

XCONV=*value*

specifies the relative parameter convergence criterion. Termination requires a small relative parameter change in subsequent iterations,

$$\max_i |\delta_k^{(i)}| < value$$

where

$$\delta_k^{(i)} = \begin{cases} \theta_k^{(i)} - \theta_{k-1}^{(i)} & |\theta_{k-1}^{(i)}| < .01 \\ \frac{\theta_k^{(i)} - \theta_{k-1}^{(i)}}{\theta_{k-1}^{(i)}} & \text{otherwise} \end{cases}$$

where $\theta_k^{(i)}$ is the estimate of the ith parameter at iteration k.

Display Options

ALPHA=*value*

sets the significance level used for the confidence limits for the hazards ratios. The value must be between 0 and 1. The default value is 0.05, which results in the calculation of a 95% confidence interval. This option has no effect unless the RISKLIMITS option is specified.

CORRB

displays the estimated correlation matrix of the parameter estimates.

COVB

displays the estimated covariance matrix of the parameter estimates.

ITPRINT

displays the iteration history, including the last evaluation of the gradient vector.

RISKLIMITS
RL

displays, for each explanatory variable, the $100(1-\alpha)\%$ confidence limits for the hazards ratio (e^{β_i}). The value for α is determined by the ALPHA= option.

OUTPUT Statement

OUTPUT < **OUT=** *SAS-data-set* >
< *keyword=name ... keyword=name* >< */options* > **;**

The OUTPUT statement creates a new SAS data set containing statistics calculated for each observation. These can include the estimated linear predictor ($\mathbf{z}_j'\widehat{\boldsymbol{\beta}}$) and its standard error, survival distribution estimates, residuals, and influence statistics. In addition, this data set includes the time variable, the explanatory variables listed in the MODEL statement, the censoring variable (if specified), and the BY, STRATA, FREQ, and ID variables (if specified).

For observations with missing values in the time variable or any explanatory variables, the output statistics are set to missing. However, for observations with missing values only in the censoring variable or the FREQ variable, survival estimates are still computed. Therefore, by adding observations with missing values in the FREQ variable or the censoring variable, you can compute the survivor function estimates for new observations or for settings of explanatory variables not present in the data without affecting the model fit.

No OUTPUT data set is created if the model contains a time-dependent variable defined by means of programming statements. Survival distribution estimates are set to missing for the counting process style of input.

The following list explains specifications in the OUTPUT statement.

OUT=*SAS-data-set*

names the output data set. If you omit the OUT= option, the OUTPUT data set is created and given a default name using the DATA*n* convention.

keyword=name

specifies the statistics included in the OUTPUT data set and names the new variables that contain the statistics. Specify a keyword for each desired statistic (see the following list of keywords), an equal sign, and either a variable or a list of variables to contain the statistic. The keywords that accept a list of variables are DFBETA, RESSCH, RESSCO, and WTRESSCH. For these keywords, you can specify as many names in *name* as the number of explanatory variables specified in the MODEL statement. If you specify *k* names and *k* is less than the total number of explanatory variables, only the changes for the first *k* parameter estimates are output. The keywords and the corresponding statistics are as follows:

DFBETA

approximate changes in the parameter estimates $(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(j)})$ when the jth observation is omitted. These variables are a weighted transform of the score residual variables and are useful in assessing local influence and in computing robust variance estimates.

LD approximate likelihood displacement when the observation is left out. This diagnostic can be used to assess the impact of each observation on the overall fit of the model.

LMAX relative influence of observations on the overall fit of the model. This diagnostic is useful in assessing the sensitivity of the fit of the model to each observation.

LOGLOGS log of the negative log of SURVIVAL

LOGSURV log of SURVIVAL

NUM_LEFT number of subjects at risk at the observation time τ_j (or at the right endpoint of the at risk interval when a counting process MODEL specification is used)

RESDEV deviance residual $\widehat{D}_j$. This is a transform of the martingale residual to achieve a more symmetric distribution.

RESMART martingale residual $\widehat{M}_j$. The residual at the observation time τ_j can be interpreted as the difference over $[0, \tau_j]$ in the observed number of events minus the expected number of events given by the model.

RESSCH Schoenfeld residuals. These residuals are useful in assessing the proportional hazards assumption.

RESSCO score residuals. These residuals are a decomposition of the first partial derivative of the log likelihood. They can be used to assess the leverage exerted by each subject in the parameter estimation. They are also useful in constructing robust sandwich variance estimators.

STDXBETA standard error of the estimated linear predictor, $\sqrt{\mathbf{z}_j' \widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}})\mathbf{z}_j}$

SURVIVAL survivor function estimate $\widehat{S}_j = [\widehat{S}_0(\tau_j)]^{\exp(\mathbf{z}_j'\widehat{\boldsymbol{\beta}})}$, where τ_j is the observation time

WTRESSCH weighted Schoenfeld residuals. These residuals are useful in investigating the nature of nonproportionality if the proportional hazard assumption does not hold.

XBETA estimate of the linear predictor, $\mathbf{z}_j'\widehat{\boldsymbol{\beta}}$

The following options can appear in the OUTPUT statement after a slash (/).

ORDER=*sort_order*

specifies the order of the observations in the OUTPUT data set. Available values for *sort_order* are

DATA requests that the output observations be sorted the same as the input data set.

SORTED requests that the output observations be sorted by strata and descending order of the time variable within each stratum.

The default is ORDER=DATA.

METHOD=*method*

specifies the method used to compute the survivor function estimates. The two available methods are

CH | EMP	specifies that the empirical cumulative hazard function estimate of the survivor function is to be computed; that is, the survivor function is estimated by exponentiating the negative empirical cumulative hazard function.
PL	specifies that the product-limit estimate of the survivor function is to be computed. The default is METHOD=PL.

Programming Statements

Programming statements are used to create or modify the values of the explanatory variables in the MODEL statement. They are especially useful in fitting models with time-dependent explanatory variables. Programming statements can also be used to create explanatory variables that are not time dependent. For example, you can create indicator variables from a categorical variable and incorporate them into the model. PROC PHREG programming statements cannot be used to create or modify the values of the response variable, the censoring variable, the frequency variable, or the strata variables.

The following DATA step statements are available in PROC PHREG:

```
ABORT
ARRAY
assignment statements
CALL
DO
iterative DO
DO UNTIL
DO WHILE
END
GOTO
IF-THEN/ELSE
LINK-RETURN
PUT
SELECT
SUM statement
```

By default, the PUT statement in PROC PHREG writes to the Output window instead of the Log window. If you want the results of the PUT statements to go to the Log window, add the following statement before the PUT statements:

```
FILE LOG;
```

DATA step functions are also available. Use these programming statements the same way you use them in the DATA step. For detailed information, refer to *SAS Language Reference: Dictionary*.

Consider the following example of using programming statements in PROC PHREG. Suppose blood pressure is measured at multiple times during the course of a study investigating the effect of blood pressure on some survival time. By treating the blood pressure as a time-dependent explanatory variable, you are able to use the value of the most recent blood pressure at each specific point of time in the modeling process rather than using the initial blood pressure or the final blood pressure. The values of the following variables are recorded for each patient, if they are available. Otherwise, the variables contain missing values.

Time	survival time
Censor	censoring indicator (with 0 as the censoring value)
BP0	blood pressure on entry to the study
T1	time 1
BP1	blood pressure at T1
T2	time 2
BP2	blood pressure at T2

The following programming statements create a variable BP. At each time T, the value of BP is the blood pressure reading for that time, if available. Otherwise, it is the last blood pressure reading.

```
proc phreg;
   model Time*Censor(0)=BP;
   BP = BP0;
   if Time>=T1 and T1^=. then BP=BP1;
   if Time>=T2 and T2^=. then BP=BP2;
run;
```

For other illustrations of using programming statements, see the “Getting Started” section on page 2573 and Example 49.4 on page 2622.

STRATA Statement

STRATA *variable* < *(list)* >< ... *variable* < *(list)* >>< */option* > **;**

The proportional hazards assumption may not be realistic for all data. If so, it may still be reasonable to perform a stratified analysis. The STRATA statement names the variables that determine the stratification. Strata are formed according to the nonmissing values of the STRATA variables unless the MISSING option is specified. In the STRATA statement, *variable* is a variable with values that are used to determine the strata levels, and *list* is an optional list of values for a numeric variable. Multiple variables can appear in the STRATA statement.

The values for *variable* can be formatted or unformatted. If the variable is a character variable, or if the variable is numeric and no list appears, then the strata are defined by the unique values of the variable. If the variable is numeric and is followed by a list, then the levels for that variable correspond to the intervals defined by the list. The corresponding strata are formed by the combination of levels and unique values. The list can include numeric values separated by commas or blanks, *value* to *value* by *value* range specifications, or combinations of these.

For example, the specification

```
strata age (5, 10 to 40 by 10) sex ;
```

indicates that the levels for age are to be less than 5, 5 to 10, 10 to 20, 20 to 30, 30 to 40, and greater than 40. (Note that observations with exactly the cutpoint value fall into the interval above the cutpoint.) Thus, with the sex variable, this STRATA statement specifies 12 strata altogether.

The following option can be specified in the STRATA statement after a slash (/).

MISSING

allows missing values ('.' for numeric variables and blanks for character variables) as valid STRATA variable values. Otherwise, observations with missing STRATA variable values are deleted from the analysis.

TEST Statement

< *label:* > **TEST** *equation1* < , . . . , *equationk* >< */option* > **;**

The TEST statement tests linear hypotheses about the regression coefficients. PROC PHREG performs a Wald test for the joint hypothesis specified in a single TEST statement. Each equation specifies a linear hypothesis; multiple equations (rows of the joint hypothesis) are separated by commas. The label is used to identify the resulting output, and it should always be included. You can submit multiple TEST statements.

The form of an equation is as follows:

term < ±*term* **. . .** > < = < ±*term* < ±*term* **. . .** >>>

here *term* is a variable or a constant or a constant times a variable. The variable is any explanatory variable in the MODEL statement. When no equal sign appears, the expression is set to 0. The following code illustrates possible uses of the TEST statement:

```
proc phreg;
   model time= a1 a2 a3 a4;
   Test1: Test a1, a2;
   Test2: Test a1=0,a2=0;
   Test3: Test a1=a2=a3;
   Test4: Test a1=a2,a2=a3;
run;
```

Note that the first and second TEST statements are equivalent, as are the third and fourth TEST statements.

The following option can be specified in the TEST statement after a slash (/).

PRINT

displays intermediate calculations. This includes $\mathbf{L}\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}})\mathbf{L}'$ bordered by $(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{c})$, and $[\mathbf{L}\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}})\mathbf{L}']^{-1}$ bordered by $[\mathbf{L}\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}})\mathbf{L}']^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{c})$, where $\mathbf{L}$ is a matrix of linear coefficients and $\mathbf{c}$ is a vector of constants. See the section "Testing Linear Hypotheses about Regression Coefficients" on page 2598.

Details

Failure Time Distribution

Let T be a nonnegative random variable representing the failure time of an individual from a homogeneous population. The survival distribution function (also known as the survivor function) of T is written as

$$S(t) = \Pr(T \geq t)$$

A mathematically equivalent way of specifying the distribution of T is through its hazard function. The hazard function $h(t)$ specifies the instantaneous failure rate at t. If T is a continuous random variable, $h(t)$ is expressed as

$$h(t) = \lim_{\Delta t \to 0^+} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$$

where $f(t)$ is the probability density function of T. If T is discrete with masses at $x_1 < x_2 < \ldots$, then $h(t)$ is given by

$$h(t) = \sum_j h_j \delta(t - x_j)$$

where

$$\delta(u) = \begin{cases} 0 & \text{if } u < 0 \\ 1 & \text{otherwise} \end{cases}$$

$$h_j = \Pr(T = x_j \mid T \geq x_j) = \frac{\Pr(T = x_j)}{S(x_j)}$$

for $j = 1, 2, \ldots$

Partial Likelihood Function for the Cox Model

Let $\mathbf{z}_l$ denote the vector of (possibly time-dependent) explanatory variables for the lth individual. Let $t_1 < t_2 < \ldots < t_k$ denote the k distinct, ordered event times. Let d_i denote the multiplicity of failures at t_i; that is, d_i is the size of the set $\mathcal{D}_i$ of individuals that fail at t_i. Let $\mathbf{s}_i$ be the sum of the vectors $\mathbf{z}_l$ over the individuals who fail at t_i; that is, $\mathbf{s}_i = \sum_{l \in \mathcal{D}_i} \mathbf{z}_l$. Using this notation, the likelihood functions used in PROC PHREG to estimate $\boldsymbol{\beta}$ are described in the following sections.

Continuous Time Scale

Let $\mathcal{R}_i$ denote the risk set just before the ith ordered event time t_i . Let $\mathcal{R}_i^*$ denote the set of individuals whose event or censored times exceed t_i or whose censored times are equal to t_i.

The exact likelihood is

$$L_1(\boldsymbol{\beta}) = \prod_{i=1}^{k} \left\{ \int_0^\infty \prod_{j=1}^{d_i} \left[1 - \exp\left(-\frac{\exp(\mathbf{z}_j'\boldsymbol{\beta})}{\sum_{l \in \mathcal{R}_i^*} \exp(\mathbf{z}_l'\boldsymbol{\beta})} t \right) \right] \exp(-t)dt \right\}$$

The Breslow likelihood is

$$L_2(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\mathbf{s}_i'\boldsymbol{\beta})}{\left[\sum_{l \in \mathcal{R}_i} \exp(\mathbf{z}_l'\boldsymbol{\beta}) \right]^{d_i}}$$

The Efron likelihood is

$$L_3(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\mathbf{s}_i'\boldsymbol{\beta})}{\prod_{j=1}^{d_i} \left[\sum_{l \in \mathcal{R}_i} \exp(\mathbf{z}_l'\boldsymbol{\beta}) - \frac{j-1}{d_i} \sum_{l \in \mathcal{D}_i} \exp(\mathbf{z}_l'\boldsymbol{\beta}) \right]}$$

Discrete Time Scale

Let $\mathcal{Q}_i$ denote the set of all subsets of d_i individuals from the risk set $\mathcal{R}_i$. For each $\mathbf{q} \in \mathcal{Q}_i$, $\mathbf{q}$ is a d_i-tuple $(q_1, q_2, \ldots, q_{d_i})$ of individuals who might have failed at t_i. Let $\mathbf{s}_\mathbf{q}^* = \sum_{l=1}^{d_i} z_{q_l}$.

The discrete logistic likelihood is

$$L_4(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\mathbf{s}_i'\boldsymbol{\beta})}{\sum_{\mathbf{q} \in \mathcal{Q}_i} \exp(\mathbf{s}_\mathbf{q}^{*\prime}\boldsymbol{\beta})}$$

When there are no ties on the event times (that is, $d_i \equiv 1$), all four likelihood functions $L_1(\boldsymbol{\beta})$, $L_2(\boldsymbol{\beta})$, $L_3(\boldsymbol{\beta})$, and $L_4(\boldsymbol{\beta})$ reduce to the same expression. In a stratified analysis, the partial likelihood is the product of the partial likelihood functions for the individual strata.

Counting Process Style of Input

In the counting process formulation, data for each subject are identified by a triple $\{N, Y, \mathbf{Z}\}$ of counting, censoring, and covariate processes. Here, $N(t)$ indicates the number of events that the subject experiences over the time interval $(0, t]$; $Y(t)$ indicates whether the subject is at risk at time t (one if at risk and zero otherwise); and $\mathbf{Z}(t)$ is a vector of explanatory variables for the subject at time t. The sample path of N is a step function with jumps of size +1 at the event times, and $N(0) = 0$. Unless $\mathbf{Z}(t)$ changes continuously with time, the data for each subject can be represented by multiple observations, each identifying a semiclosed time interval $(t1, t2]$, the values of the explanatory variables over that interval, and the event status at $t2$. The subject remains at risk during the interval $(t1, t2]$, and an event may occur at $t2$. Values of the explanatory variables for the subject remain unchanged in the interval. This style of data input was originated by Terry M. Therneau (1994).

For example, a patient has a tumor recurrence at weeks 3, 10, and 15 and is followed to week 23. The explanatory variables are Trt (treatment), Z1 (initial tumor number), and Z2 (initial tumor size), and, for this patient, the values of Trt, Z1, and Z2 are (1,1,3). The data for this patient are represented by the following four observations:

T1	T2	Event	Trt	Z1	Z2
0	3	1	1	1	3
3	10	1	1	1	3
10	15	1	1	1	3
15	23	0	1	1	3

Here (T1,T2] contains the at-risk intervals. The variable Event is a censoring variable indicating whether a recurrence has occurred at T2; a value of 1 indicates a tumor recurrence, and a value of 0 indicates nonrecurrence. The PHREG procedure fits the multiplicative hazards model, which is specified as follows:

```
proc phreg;
   model (T1,T2) * Event(0) = Trt Z1 Z2;
run;
```

Another useful application of the counting process formulation is delayed entry of subjects into the risk set. For example, in studying the mortality of workers exposed to a carcinogen, the survival time is chosen to be the worker's age at death by malignant neoplasm. Any worker joining the workplace at a later age than a given event failure time is not included in the corresponding risk set. The variables of a worker consist of Entry (age at which the worker entered the workplace), Age (age at death or age censored), Status (an indicator of whether the observation time is censored, with the value 0 identifying a censored time), and X1 and X2 (explanatory variables thought to be related to survival). The specification for such an application is as follows.

```
proc phreg;
   model (Entry, Age) * Status(0) = X1 X2;
run;
```

Alternatively, you can use a time-dependent variable to control the risk set, as illustrated in the following specification:

```
proc phreg;
   model Age * Status(0) = X1 X2;
   if Age < Entry then X1= .;
run;
```

Here, X1 becomes a time-dependent variable. At a given death time t, the value of X1 is reevaluated for each subject with Age $\geq t$; subjects with Entry $> t$ are given a missing value in X1 and are subsequently removed from the risk set. Computationally, this approach is not as efficient as the one that uses the counting process formulation.

The Multiplicative Hazards Model

Consider a set of n subjects such that the counting process $N_i \equiv \{N_i(t), t \geq 0\}$ for the ith subject represents the number of observed events experienced over time t. The sample paths of the process N_i are step functions with jumps of size $+1$, with $N_i(0) = 0$. Let $\boldsymbol{\beta}$ denote the vector of unknown regression coefficients. The multiplicative hazards function $\Lambda(t, \mathbf{Z}_i(t))$ for N_i is given by

$$Y_i(t)d\Lambda(t, \mathbf{Z}_i(t)) = Y_i(t)\exp(\boldsymbol{\beta}'\mathbf{Z}_i(t))d\Lambda_0(t)$$

where

- $Y_i(t)$ indicates whether the ith subject is at risk at time t (specifically, $Y_i(t) = 1$ if at risk and $Y_i(t) = 0$ otherwise)
- $\mathbf{Z}_i(t)$ is the vector of explanatory variables for the ith subject at time t
- $\Lambda_0(t)$ is an unspecified baseline hazard function

Refer to Fleming and Harrington (1991) and Andersen and others (1992). The Cox model is a special case of this multiplicative hazards model, where $Y_i(t) = 1$ until the first event or censoring, and $Y_i(t) = 0$ thereafter.

The partial likelihood for n independent triplets $(N_i, Y_i, \mathbf{Z}_i), i = 1, \ldots, n$, has the form

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n}\prod_{t\geq 0}\left\{\frac{Y_i(t)\exp(\boldsymbol{\beta}'\mathbf{Z}_i(t))}{\sum_{j=1}^{n} Y_j(t)\exp(\boldsymbol{\beta}'\mathbf{Z}_j(t))}\right\}^{\Delta N_i(t)}$$

where $\Delta N_i(t) = 1$ if $N_i(t) - N_i(t-) = 1$, and $\Delta N_i(t) = 0$ otherwise.

Newton-Raphson Method

Let $L(\boldsymbol{\beta})$ be one of the likelihood functions described in the previous subsections. Let $l(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta})$. Finding $\boldsymbol{\beta}$ such that $L(\boldsymbol{\beta})$ is maximized is equivalent to finding the solution $\widehat{\boldsymbol{\beta}}$ to the likelihood equations

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$$

With $\widehat{\boldsymbol{\beta}}^0 = \mathbf{0}$ as the initial solution, the iterative scheme is expressed as

$$\widehat{\boldsymbol{\beta}}^{j+1} = \widehat{\boldsymbol{\beta}}^j - \left[\frac{\partial^2 l(\widehat{\boldsymbol{\beta}}^j)}{\partial \boldsymbol{\beta}^2}\right]^{-1} \frac{\partial l(\widehat{\boldsymbol{\beta}}^j)}{\partial \boldsymbol{\beta}}$$

The term after the minus sign is the Newton-Raphson step. If the likelihood function evaluated at $\widehat{\boldsymbol{\beta}}^{j+1}$ is less than that evaluated at $\widehat{\boldsymbol{\beta}}^j$, then $\widehat{\boldsymbol{\beta}}^{j+1}$ is recomputed using half the step size. The iterative scheme continues until convergence is obtained, that is, until $\widehat{\boldsymbol{\beta}}_{m+1}$ is sufficiently close to $\widehat{\boldsymbol{\beta}}_m$. Then the maximum likelihood estimate of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{m+1}$.

The estimated covariance matrix of $\widehat{\boldsymbol{\beta}}$ is given by

$$\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}}) = -\left[\frac{\partial^2 l(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^2}\right]^{-1}$$

Testing the Global Null Hypothesis

The following three likelihood statistics can be used to test the global null hypothesis $H_0\!:\boldsymbol{\beta} = \mathbf{0}$. Under mild assumptions, each statistic has an asymptotic chi-squared distribution with *p* degrees of freedom given the null hypothesis. The value *p* is the dimension of $\boldsymbol{\beta}$.

Likelihood ratio test:

$$\chi^2_{LR} = 2\left[l(\widehat{\boldsymbol{\beta}}) - l(\mathbf{0})\right]$$

Wald's test:

$$\chi^2_W = \widehat{\boldsymbol{\beta}}'\left[\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}})\right]^{-1}\widehat{\boldsymbol{\beta}}$$

Score test:

$$\chi^2_S = \left[\frac{\partial l(\mathbf{0})}{\partial \boldsymbol{\beta}}\right]' \left[-\frac{\partial^2 l(\mathbf{0})}{\partial \boldsymbol{\beta}^2}\right]^{-1} \left[\frac{\partial l(\mathbf{0})}{\partial \boldsymbol{\beta}}\right]$$

Hazards Ratio Estimates and Confidence Limits

Let β_i and $\widehat{\beta}_i$ denote the *i*th component of $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$, respectively. The hazards ratio (also known as risk ratio) for the explanatory variable with regression coefficient β_i is defined as $\exp(\beta_i)$. The hazards ratio is estimated by $\exp(\widehat{\beta}_i)$. The $100(1-\alpha)\%$ confidence limits for the hazards ratio are calculated as

$$\exp\left(\widehat{\beta}_i \pm z_{\alpha/2}\sqrt{\widehat{\mathbf{V}}_{ii}(\widehat{\boldsymbol{\beta}})}\right)$$

where $\widehat{\mathbf{V}}_{ii}(\widehat{\boldsymbol{\beta}})$ is the *i*th diagonal element of $\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}})$, and $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile point of the standard normal distribution.

The hazards ratio is the ratio of the hazards functions that correspond to a change of one unit of the given variable and conditional on fixed values of all other variables.

Testing Linear Hypotheses about Regression Coefficients

Linear hypotheses for $\boldsymbol{\beta}$ are expressed in matrix form as

$$H_0 \colon \mathbf{L}\boldsymbol{\beta} = \mathbf{c}$$

where **L** is a matrix of coefficients for the linear hypotheses, and **c** is a vector of constants. The Wald chi-squared statistic for testing H_0 is computed as

$$\chi^2_W = \left(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{c}\right)' \left[\mathbf{L}\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}})\mathbf{L}'\right]^{-1} \left(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{c}\right)$$

Under H_0, χ^2_W has an asymptotic chi-squared distribution with *r* degrees of freedom, where *r* is the rank of **L**.

Residuals

The cumulative baseline hazard function Λ_0 is estimated by

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(s)}{\sum_{j=1}^{n} Y_j(s)\exp(\widehat{\boldsymbol{\beta}}'\mathbf{Z}_j(s))}$$

Although this formula is for the BRESLOW=TIES option, the same formula is used for other TIES= options. The discrepancies between results obtained by using an appropriate formula for a nondefault TIES= option and those obtained by the given formula are minimal.

The martingale residual at t is defined as

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s)\exp(\widehat{\boldsymbol{\beta}}'\mathbf{Z}_i(s))d\hat{\Lambda}_0(s)$$

Here $\hat{M}_i(t)$ estimates the difference over $(0, t]$ between the observed number of events for the ith subject and a conditional expected number of events. The quantity $\hat{M}_i \equiv \hat{M}_i(\infty)$ is referred to as the martingale residual for the ith subject. When the counting process MODEL specification is used, the RESMART= variable contains the component $(\hat{M}_i(t_2) - \hat{M}_i(t_1))$ instead of the martingale residual at t_2. The martingale residual for a subject can be obtained by summing up these component residuals within the subject. For the Cox model with no time-dependent explanatory variables, the martingale residual for the ith subject with observation time t_i and event status δ_i, where

$$\delta_i = \begin{cases} 0 & \text{if } t_i \text{ is a censored time} \\ 1 & \text{if } t_i \text{ is an event time} \end{cases}$$

is

$$\hat{M}_i = \delta_i - \hat{\Lambda}_0(t_i)\exp(\hat{\boldsymbol{\beta}}'\mathbf{z}_i)$$

The deviance residuals d_i are a transform of the martingale residuals:

$$d_i = sign(\hat{M}_i)\sqrt{2\left[-\hat{M}_i - N_i(\infty)\log\left(\frac{N_i(\infty) - \hat{M}_i}{N_i(\infty)}\right)\right]}$$

The square root shrinks large negative martingale residuals, while the logarithmic transformation expands martingale residuals that are close to unity. As such, the deviance residuals are more symmetrically distributed about zero than the martingale residuals. For the Cox model, the deviance residual reduces to the form

$$d_i = sign(\hat{M}_i)\sqrt{2[-\hat{M}_i - \delta_i\log(\delta_i - \hat{M}_i)]}$$

When the counting process MODEL specification is used, values of the RESDEV= variable are set to missing because the deviance residuals can be calculated on a per subject basis only.

The Schoenfeld residual vector is calculated on a per event time basis as

$$\mathbf{U}_i(t) = \mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)$$

where t is an event time, and $\bar{\mathbf{Z}}(t)$ is a weighted average of the covariates over the risk set at time t and is given by

$$\bar{\mathbf{Z}}(t) = \frac{\sum_{l=1}^{n} Y_l(t)\mathbf{Z}_l(t)\exp(\hat{\boldsymbol{\beta}}'\mathbf{Z}_l(t))}{\sum_{l=1}^{n} Y_l(t)\exp(\hat{\boldsymbol{\beta}}'\mathbf{Z}_l(t))}$$

Under the proportional hazards assumption, the Schoenfeld residuals have the sample path of a random walk; therefore, they are useful in assessing time trend or lack of proportionality. Harrell (1986) proposed a z-transform of the Pearson correlation between these residuals and the rank order of the failure time as a test statistic for nonproportional hazards. Therneau, Grambsch, and Fleming (1990) considered a Kolmogorov-type test using the cumulative sum of the residuals.

The score process for the ith subject at time t is

$$\mathbf{L}_i(t) = \int_0^t [\mathbf{Z}_i(s) - \bar{\mathbf{Z}}(s)] d\hat{M}_i(s)$$

The vector $\mathbf{L}_i \equiv \mathbf{L}_i(\infty)$ is the score residual for the ith subject. When the counting process MODEL specification is used, the RESSCO= variables contain the components of $(\mathbf{L}_i(t2) - \mathbf{L}_i(t1))$ instead of the score process at $t2$. The score residual for a subject can be obtained by summing up these component residuals within the subject.

The score residuals are a decomposition of the first partial derivative of the log likelihood. They are useful in assessing the influence of each subject on individual parameter estimates. They also play an important role in the computation of the robust sandwich variance estimators of Lin and Wei (1989) and Wei, Lin, and Weissfeld (1989).

Diagnostics Based on Weighted Residuals

The vector of weighted Schoenfeld residuals, $\mathbf{r}_i$, is computed as

$$\mathbf{r}_i = n_e \widehat{\mathbf{V}} \mathbf{U}_i(t_i)$$

where n_e is the total number of events, $\widehat{\mathbf{V}} = \widehat{\mathbf{V}}(\hat{\boldsymbol{\beta}})$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$, and $\mathbf{U}_i(t_i)$ is the vector of Schoenfeld residuals at the event time t_i. The components of $\mathbf{r}_i$ are output to the WTRESSCH= variables.

The weighted Schoenfeld residuals are useful in assessing the proportional hazards assumption. The idea is that most of the common alternatives to the proportional hazards can be cast in terms of a time-varying coefficient model

$$\lambda(t, \mathbf{Z}) = \lambda_0(t) \exp(\beta_1(t) Z_1 + \beta_2(t) Z_2 + \ldots)$$

where $\lambda(t, \mathbf{Z})$ and $\lambda_0(t)$ are hazards rates. Let $\hat{\beta}_j$ and r_{ij} be the jth component of $\hat{\boldsymbol{\beta}}$ and $\mathbf{r}_i$, respectively. Grambsch and Therneau (1993) suggest using a smoothed plot of $(\hat{\beta}_j + r_{ij})$ versus t_i to discover the functional form of the time-varying coefficient $\beta_j(t)$. A zero slope indicates that the coefficient is not varying with time.

The weighted score residuals are used more often than their unscaled counterparts in assessing local influence. Let $\hat{\boldsymbol{\beta}}_{(i)}$ be the estimate of $\boldsymbol{\beta}$ when the ith subject is left out, and let $\boldsymbol{\Delta}_i = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$. The jth component of $\boldsymbol{\Delta}_i$ can be used to assess any untoward effect of the ith subject on $\hat{\beta}_j$. The exact computation of $\boldsymbol{\Delta}_i$ involves refitting the model each time a subject is omitted. Cain and Lange (1984) derived the following approximation of $\boldsymbol{\Delta}_i$ as weighted score residuals:

$$\widehat{\boldsymbol{\Delta}}_i = \widehat{\mathbf{V}} \mathbf{L}_i$$

Here, $\widehat{\mathbf{V}} = \widehat{\mathbf{V}}(\hat{\boldsymbol{\beta}})$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$, and $\mathbf{L}_i$ is the vector of the score residuals for the ith subject. Values of $\widehat{\boldsymbol{\Delta}}_i$ are output to the DFBETA= variables. Again, when the counting process MODEL specification is used, the DFBETA= variables contain the component $(\widehat{\mathbf{V}}\mathbf{L}_i(t2) - \widehat{\mathbf{V}}\mathbf{L}_i(t1))$. The vector $\widehat{\boldsymbol{\Delta}}_i$ for a subject can be obtained by summing these components within the subject.

Note that these DFBETA statistics are a transform of the score residuals. In computing the robust sandwich variance estimators of Lin and Wei (1989) and Wei, Lin, and Weissfeld (1989), it is more convenient to use the DFBETA statistics than the score residuals (see Example 49.8 on page 2642).

Influence of Observations on Overall Fit of the Model

The LD statistic approximates the likelihood displacement, which is the amount by which minus twice the log likelihood $(-2\log\mathcal{L}(\hat{\boldsymbol{\beta}}))$, under a fitted model, changes when each subject in turn is left out. When the ith subject is omitted, the likelihood displacement is

$$2\log\mathcal{L}(\hat{\boldsymbol{\beta}}) - 2\log\mathcal{L}(\hat{\boldsymbol{\beta}}_{(i)})$$

where $\hat{\boldsymbol{\beta}}_{(i)}$ is the vector of parameter estimates obtained by fitting the model without the ith subject. Instead of refitting the model without the ith subject, Pettitt and Bin Daud (1989) propose that the likelihood displacement for the ith subject be approximated by

$$LD_i = \mathbf{L}_i'\widehat{\mathbf{V}}\mathbf{L}_i$$

This approximation is output to the LD= variable.

The LMAX statistic is another global influence statistic. This statistic is based on the symmetric matrix

$$\mathbf{B} = \mathbf{L}\widehat{\mathbf{V}}\mathbf{L}'$$

where $\mathbf{L}$ is the matrix with rows that are the score residual vectors $\mathbf{L}_i$. The elements of the eigenvector associated with the largest eigenvalue of the matrix $\mathbf{B}$, standardized to unit length, give a measure of the sensitivity of the fit of the model to each observation in the data. The influence of the ith subject on the global fit of the model is proportional to the magnitude of ζ_i, where ζ_i is the ith element of the vector $\boldsymbol{\zeta}$ that satisfies

$$\mathbf{B}\boldsymbol{\zeta} = \lambda_{\max}\boldsymbol{\zeta} \quad \text{and} \quad \boldsymbol{\zeta}'\boldsymbol{\zeta} = 1$$

with $\lambda_{\max}$ being the largest eigenvalue of $\mathbf{B}$. The sign of ζ_i is irrelevant, and its absolute value is output to the LMAX= variable.

When the counting process MODEL specification is used, the LD= and LMAX= variables are set to missing, because these two global influence statistics can be calculated on a per subject basis only.

Survival Distribution Estimates for the Cox Model

Two estimators of the survivor function are available: one is the product-limit estimate and the other is based on the empirical cumulative hazard function.

Product-Limit Estimates

Let $\mathcal{C}_i$ denote the set of individuals censored in the half-open interval $[t_i, t_{i+1})$, where $t_0 = 0$ and $t_{k+1} = \infty$. Let γ_l denote the censoring times in $[t_i, t_{i+1})$; l ranges over $\mathcal{C}_i$. The likelihood function for all individuals is given by

$$\mathcal{L} = \prod_{i=0}^{k} \left\{ \prod_{l \in \mathcal{D}_i} \left([S_0(t_i)]^{\exp(\mathbf{z}_l'\boldsymbol{\beta})} - [S_0(t_i+0)]^{\exp(\mathbf{z}_l'\boldsymbol{\beta})} \right) \prod_{l \in \mathcal{C}_i} [S_0(\gamma_l+0)]^{\exp(\mathbf{z}_l'\boldsymbol{\beta})} \right\}$$

where $\mathcal{D}_0$ is empty. The likelihood $\mathcal{L}$ is maximized by taking $S_0(t) = S_0(t_i+0)$ for $t_i < t \le t_{i+1}$ and allowing the probability mass to fall only on the observed event times $t_1, \ldots, t_k$. By considering a discrete model with hazard contribution $1-\alpha_i$ at t_i, you take $S_0(t_i) = S_0(t_{i-1}+0) = \prod_{j=0}^{i-1} \alpha_j$, where $\alpha_0 = 1$. Substitution into the likelihood function produces

$$\mathcal{L} = \prod_{i=0}^{k} \left\{ \prod_{j \in \mathcal{D}_i} \left(1 - \alpha_i^{\exp(\mathbf{z}_j'\boldsymbol{\beta})} \right) \prod_{l \in \mathcal{R}_i - \mathcal{D}_i} \alpha_i^{\exp(\mathbf{z}_l'\boldsymbol{\beta})} \right\}$$

If you replace $\boldsymbol{\beta}$ with $\widehat{\boldsymbol{\beta}}$ estimated from the partial likelihood function and then maximize with respect to α_1 , . . . ,α_k , the maximum likelihood estimate $\widehat{\alpha}_i$ of α_i becomes a solution of

$$\sum_{j \in \mathcal{D}_i} \frac{\exp(\mathbf{z}_j'\widehat{\boldsymbol{\beta}})}{1 - \widehat{\alpha}_i^{\exp(\mathbf{z}_j'\widehat{\boldsymbol{\beta}})}} = \sum_{l \in \mathcal{R}_i} \exp(\mathbf{z}_l'\widehat{\boldsymbol{\beta}})$$

When only a single failure occurs at t_i, $\widehat{\alpha}_i$ can be found explicitly. Otherwise, an iterative solution is obtained by the Newton method.

The estimated baseline cumulative hazard function is

$$\widehat{H}_0(t) = -\log(\widehat{S}_0(t))$$

where $\widehat{S}_0(t)$ is the estimated baseline survivor function given by

$$\widehat{S}_0(t) = \widehat{S}_0(t_{i-1}+0) = \prod_{j=0}^{i-1} \widehat{\alpha}_j, \qquad t_{i-1} < t \le t_i$$

For details, refer to Kalbfleisch and Prentice (1980). For a given realization of the explanatory variables $\boldsymbol{\xi}$, the product-limit estimate of the survival function at $\mathbf{Z} = \boldsymbol{\xi}$ is

$$\hat{S}(t, \boldsymbol{\xi}) = [\widehat{S}_0(t)]^{\exp(\boldsymbol{\beta}'\boldsymbol{\xi})}$$

Empirical Cumulative Hazards Function Estimates

Let $\boldsymbol{\xi}$ be a given realization of the explanatory variables. The empirical cumulative hazard function estimate at $\mathbf{Z} = \boldsymbol{\xi}$ is

$$\widehat{\Lambda}(t,\boldsymbol{\xi}) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(s)}{\sum_{j=1}^{n} Y_j(s) \exp(\widehat{\boldsymbol{\beta}}'(\mathbf{z}_j - \boldsymbol{\xi}))}$$

The variance estimator of $\widehat{\Lambda}(t,\boldsymbol{\xi})$ is given by the following (Tsiatis 1981):

$$\begin{aligned} & \widehat{var}\{n^{\frac{1}{2}}(\widehat{\Lambda}(t,\boldsymbol{\xi}) - \Lambda(t,\boldsymbol{\xi}))\} \\ = \quad & n\left\{\sum_{i=1}^{n} \int_0^t \frac{dN_i(s)}{[\sum_{j=1}^{n} Y_j(s) \exp(\widehat{\boldsymbol{\beta}}'(\mathbf{z}_j - \boldsymbol{\xi}))]^2} + \mathbf{H}'(t,\boldsymbol{\xi})\widehat{\mathbf{V}}(\hat{\boldsymbol{\beta}})\mathbf{H}(t,\boldsymbol{\xi})\right\} \end{aligned}$$

where $\widehat{\mathbf{V}}(\hat{\boldsymbol{\beta}})$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ and

$$\mathbf{H}(t,\boldsymbol{\xi}) = \sum_{i=1}^{n} \int_0^t \frac{\sum_{l=1}^{n} Y_l(s)(\mathbf{Z}_l - \boldsymbol{\xi}) \exp(\widehat{\boldsymbol{\beta}}'(\mathbf{z}_l - \boldsymbol{\xi}))}{[\sum_{j=1}^{n} Y_j(s) \exp(\widehat{\boldsymbol{\beta}}'(\mathbf{z}_j - \boldsymbol{\xi}))]^2} dN_i(s)$$

The empirical cumulative hazard function (CH) estimate of the survivor function for $\mathbf{Z} = \boldsymbol{\xi}$ is

$$\tilde{S}(t,\boldsymbol{\xi}) = \exp(-\widehat{\Lambda}(t,\boldsymbol{\xi}))$$

Confidence Intervals for the Survivor Function

Let $\hat{S}(t,\boldsymbol{\xi})$ and $\tilde{S}(t,\boldsymbol{\xi})$ correspond to the product-limit (PL) and empirical cumulative hazard function (CH) estimates of the survivor function for $\mathbf{Z} = \boldsymbol{\xi}$, respectively. Both the standard error of log($\hat{S}(t,\boldsymbol{\xi})$) and the standard error of log($\tilde{S}(t,\boldsymbol{\xi})$) are approximated by $\tilde{\sigma}_0(t,\boldsymbol{\xi})$, which is the square root of the variance estimate of $\widehat{\Lambda}(t,\boldsymbol{\xi})$; refer to Kalbfleish and Prentice (1980, p. 116). By the delta method, the standard errors of $\hat{S}(t,\boldsymbol{\xi})$ and $\tilde{S}(t,\boldsymbol{\xi})$ are given by

$$\hat{\sigma}_1(t,\boldsymbol{\xi}) \doteq \hat{S}(t,\boldsymbol{\xi})\tilde{\sigma}_0(t,\boldsymbol{\xi}) \qquad \text{and} \qquad \tilde{\sigma}_1(t,\boldsymbol{\xi}) \doteq \tilde{S}(t,\boldsymbol{\xi})\tilde{\sigma}_0(t,\boldsymbol{\xi})$$

respectively. The standard errors of log[-log($\hat{S}(t,\boldsymbol{\xi})$)] and log[-log($\tilde{S}(t,\boldsymbol{\xi})$)] are given by

$$\hat{\sigma}_2(t,\boldsymbol{\xi}) \doteq \frac{-\tilde{\sigma}_0(t,\boldsymbol{\xi})}{\log(\hat{S}(t,\boldsymbol{\xi}))} \qquad \text{and} \qquad \tilde{\sigma}_2(t,\boldsymbol{\xi}) \doteq \frac{\tilde{\sigma}_0(t,\boldsymbol{\xi})}{\widehat{\Lambda}(t,\boldsymbol{\xi})}$$

respectively.

Let $z_{\alpha/2}$ be the upper $100(1-\frac{\alpha}{2})$ percentile point of the standard normal distribution. A $100(1-\alpha)\%$ confidence interval for the survivor function $S(t,\boldsymbol{\xi})$ is given in the following table.

Method	CLTYPE	Confidence Limits
LOG	PL	$\exp[\log(\hat{S}(t,\boldsymbol{\xi})) \pm z_{\frac{\alpha}{2}}\tilde{\sigma}_0(t,\boldsymbol{\xi})]$
LOG	CH	$\exp[\log(\tilde{S}(t,\boldsymbol{\xi})) \pm z_{\frac{\alpha}{2}}\tilde{\sigma}_0(t,\boldsymbol{\xi})]$
LOGLOG	PL	$\exp\{-\exp[\log(-\log(\hat{S}(t,\boldsymbol{\xi}))) \pm z_{\frac{\alpha}{2}}\hat{\sigma}_2(t,\boldsymbol{\xi})]\}$
LOGLOG	CH	$\exp\{-\exp[\log(-\log(\tilde{S}(t,\boldsymbol{\xi}))) \pm z_{\frac{\alpha}{2}}\tilde{\sigma}_2(t,\boldsymbol{\xi})]\}$
NORMAL	PL	$\hat{S}(t,\boldsymbol{\xi}) \pm z_{\frac{\alpha}{2}}\hat{\sigma}_1(t,\boldsymbol{\xi})$
NORMAL	CH	$\tilde{S}(t,\boldsymbol{\xi}) \pm z_{\frac{\alpha}{2}}\tilde{\sigma}_1(t,\boldsymbol{\xi})$

Left Truncation of Failure Times

Left truncation arises when individuals come under observation only some known time after the natural time origin of the phenomenon under study. The risk set just prior to an event time does not include individuals whose left truncation times exceed the given event time. Thus, any contribution to the likelihood must be conditional on the truncation limit having been exceeded.

Although left truncation can be accommodated in PROC PHREG through the counting process style of input, such specification does not allow survival estimates to be output. Using the ENTRY= option in PROC PHREG for left truncation does not suppress computing the survival estimates. Consider the following specifications of PROC PHREG:

```
proc phreg data=one;
   model t2*dead(0)=x1-x10/entry=t1;
   baseline out=out1 survival=s;
   title 'The ENTRY= option is Specified';
run;

proc phreg data=one;
   model (t1,t2)*dead(0)=x1-x10;
   baseline out=out2 survival=s;
   title 'Counting Process Style of Input';
run;
```

Both specifications yield the same model estimates; however, the baseline data set out2 is empty, since survivor function estimates are not computed when you use the counting process style of input.

Variable Selection Methods

Five variable selection methods are available. The simplest method (and the default) is SELECTION=NONE, for which PROC PHREG fits the complete model as specified in the MODEL statement. The other four methods are FORWARD for forward selection, BACKWARD for backward elimination, STEPWISE for stepwise selection, and SCORE for best subsets selection. These methods are specified with the SELECTION= option in the MODEL statement. Intercept parameters are forced to stay in the model unless the NOINT option is specified.

When SELECTION=FORWARD, PROC PHREG first estimates parameters for variables forced into the model. These variables are the intercepts and the first n explanatory variables in the MODEL statement, where n is the number specified by the START= or INCLUDE= option in the MODEL statement (n is zero by default). Next, the procedure computes the adjusted chi-square statistics for each variable not in the model and examines the largest of these statistics. If it is significant at the SLSENTRY= level, the corresponding variable is added to the model. Once a variable is entered in the model, it is never removed from the model. The process is repeated until none of the remaining variables meet the specified level for entry or until the STOP= value is reached.

When SELECTION=BACKWARD, parameters for the complete model as specified in the MODEL statement are estimated unless the START= option is specified. In that case, only the parameters for the intercepts and the first n explanatory variables in the MODEL statement are estimated, where n is the number specified by the START= option. Results of the Wald test for individual parameters are examined. The least significant variable that does not meet the SLSSTAY= level for staying in the model is removed. Once a variable is removed from the model, it remains excluded. The process is repeated until no other variable in the model meets the specified level for removal or until the STOP= value is reached.

The SELECTION=STEPWISE option is similar to the SELECTION=FORWARD option except that variables already in the model do not necessarily remain. Variables are entered into and removed from the model in such a way that each forward selection step can be followed by one or more backward elimination steps. The stepwise selection process terminates if no further variable can be added to the model or if the variable just entered into the model is the only variable removed in the subsequent backward elimination.

For SELECTION=SCORE, PROC PHREG uses the branch and bound algorithm of Furnival and Wilson (1974) to find a specified number of models with the highest likelihood score (chi-square) statistic for all possible model sizes, from 1, 2, 3 variables, and so on, up to the single model containing all of the explanatory variables. The number of models displayed for each model size is controlled by the BEST= option. You can use the START= option to impose a minimum model size, and you can use the STOP= option to impose a maximum model size. For instance, with BEST=3, START=2, and STOP=5, the SCORE selection method displays the best three models (that is, the three models with the highest score chi-squares) containing 2, 3, 4, and 5 variables.

The SEQUENTIAL and STOPRES options can alter the default criteria for adding variables to or removing variables from the model when they are used with the FORWARD, BACKWARD, or STEPWISE selection methods.

Computational Resources

Let n be the number of observations in a BY group. Let p be the number of explanatory variables. The minimum working space (in bytes) needed to process the BY group is

$$\max\{12n, 24p^2 + 160p\}$$

Extra memory is needed for certain TIES= options. Let k be the maximum multiplicity of tied times. The TIES=DISCRETE option requires extra memory (in bytes) of

$$4k(p^2 + 4p)$$

The TIES=EXACT option requires extra memory (in bytes) of

$$24(k^2 + 5k)$$

If sufficient space is available, the input data set is also kept in memory. Otherwise, the input data is reread from the utility file for each evaluation of the likelihood function and its derivatives, with the resulting execution time substantially increased.

Displayed Output

The displayed output of the PHREG procedure contains the following:

- the two-level name of the input data set
- the name and label of the failure-time variable
- the name and label of the censoring variable
- the censoring values
- the name and label of the offset variable
- the name and label of the frequency variable
- the method of handling ties in the failure time
- the “Summary of the Number of Event and Censored Values” table, which displays, for each stratum, the breakdown of the number of events and censored values. This table is not produced if the NOSUMMARY option is specified.
- the “Simple Statistics for Explanatory Variables” table, which displays, for each stratum, the mean, standard deviation, and minimum and maximum for each explanatory variable in the MODEL statement (if you specify the SIMPLE option in the PROC PHREG statement)
- the “Iteration History” table, which displays the iteration number, step size, log likelihood, and parameter estimates at each iteration (if you specify the ITPRINT option in the MODEL statement). The last evaluation of the gradient vector is also displayed.

- the "Model Fit Statistics" table, which gives the values of -2 log likelihood for fitting a model with no explanatory variable and for fitting a model with all the explanatory variables. The AIC and SBC are also given in this table.
- the "Testing Global Null Hypothesis: BETA=0" table, which displays results of the likelihood ratio test, the score test, and the Wald test
- the "Analysis of Maximum Likelihood Estimates" table, which contains the following:
  - the maximum likelihood estimate of the parameter
  - the estimated standard error of the parameter estimate, computed as the square root of the corresponding diagonal element of the estimated covariance matrix
  - the Wald Chi-Square statistic, computed as the square of the parameter estimate divided by its standard error estimate
  - the degrees of freedom of the Wald chi-square statistic. It has a value of 1 unless the corresponding parameter is redundant or infinite, in which case the value is 0.
  - the *p*-value of the Wald chi-squared statistic with respect to a chi-squared distribution with one degree of freedom
  - the hazards ratio computed by exponentiating the parameter estimate
  - the confidence limits for the hazards ratio (if you specified the option RISKLIMITS)
- the "Regression Models Selected by Score Criterion" table, which gives the number of explanatory variables in each model, the score chi-squared statistic, and the names of the variables included in the model (if you specify SELECTION=SCORE in the MODEL statement)
- the "Analysis of Variables Not in the Model" table, which gives the Score chi-squared statistic for testing the significance of each variable not in the model after adjusting for the variables already in the model, and the *p*-value of the chi-squared statistic with respect to a chi-squared distribution with one degree of freedom (if you specify SELECTION=FORWARD, SELECTION=BACKWARD, or SELECTION=STEPWISE in the MODEL statement). This table is produced before the variable selected for entry for SELECTION=FORWARD or SELECTION=STEPWISE is displayed.
- a summary of the model-building process, which gives the step number, the explanatory variables entered or removed at each step, the chi-squared statistic, and the corresponding *p*-value on which the entry or removal is based (if you specify SELECTION=FORWARD, SELECTION=BACKWARD, or SELECTION=STEPWISE in the MODEL statement)
- the estimated covariance matrix of the parameter estimates (if you use the COVB option in the MODEL statement)
- the estimated correlation matrix of the parameter estimates (if you use the CORRB option in the MODEL statement)
- the "Linear Hypothesis Testing" table, which gives the results of the Wald test for each TEST statement (if specified)

ODS Table Names

PROC PHREG 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 49.1. ODS Tables Produced in PROC PHREG

ODS Table Name	Description	Statement	Option
BestSubsets	Best subset selection	MODEL	SELECTION=SCORE
CensoredSummary	Summary of event and censored observations	MODEL	default
ConvergenceStatus	Convergence status	MODEL	default
CorrB	Estimated correlation matrix of parameter estimators	MODEL	CORRB
CovB	Estimated covariance matrix of parameter estimators	MODEL	COVB
FitStatistics	Model fit statistics	MODEL	default
GlobalScore	Global chi-square test	MODEL	NOFIT
GlobalTests	Tests of the global null hypothesis	MODEL	default
IterHistory	Iteration history	MODEL	ITPRINT
LastGradient	Last evaluation of gradient	MODEL	ITPRINT
ModelBuildingSummary	Summary of model building	MODEL	SELECTION=B/F/S
ModelInfo	Model information	PROC	default
ParameterEstimates	Maximum likelihood estimates of model parameters	MODEL	default
ResidualChiSq	Residual chi-square	MODEL	SELECTION=F/B
SimpleStatistics	Summary statistics for explanatory variables	PROC	SIMPLE
TestPrint1	**L**[cov(**b**)]**L**’ and **Lb-c**	TEST	PRINT
TestPrint2	Ginv(**L**[cov(**b**)]**L**’) and Ginv(**L**[cov(**b**)]**L**’)(**Lb-c**)	TEST	PRINT
VariablesNotInModel	Analysis of variables not in the model	MODEL	SELECTION=F/S

Examples

Example 49.1. Stepwise Regression

Krall, Uthoff, and Harley (1975) analyzed data from a study on multiple myeloma in which researchers treated 65 patients with alkylating agents. Of those patients, 48 died during the study and 17 survived. In the data set Myeloma, the variable Time represents the survival time in months from diagnosis. The variable VStatus consists of two values, 0 and 1, indicating whether the patient was alive or dead, respectively, at the end of the study. If the value of VStatus is 0, the corresponding value of Time is censored. The variables thought to be related to survival are LogBUN (log(BUN) at diagnosis), HGB (hemoglobin at diagnosis), Platelet (platelets at diagnosis: 0=abnormal, 1=normal), Age (age at diagnosis in years), LogWBC (log(WBC) at diagnosis), Frac (fractures at diagnosis: 0=none, 1=present), LogPBM (log percentage of plasma cells in bone marrow), Protein (proteinuria at diagnosis), and SCalc (serum calcium at diagnosis). Interest lies in identifying important prognostic factors from these nine explanatory variables.

```
data Myeloma;
   input Time VStatus LogBUN HGB Platelet Age LogWBC Frac
         LogPBM Protein SCalc;
   label Time='Survival Time'
         VStatus='0=Alive 1=Dead';
   datalines;
 1.25  1  2.2175   9.4  1  67  3.6628  1  1.9542  12  10
 1.25  1  1.9395  12.0  1  38  3.9868  1  1.9542  20  18
 2.00  1  1.5185   9.8  1  81  3.8751  1  2.0000   2  15
 2.00  1  1.7482  11.3  0  75  3.8062  1  1.2553   0  12
 2.00  1  1.3010   5.1  0  57  3.7243  1  2.0000   3   9
 3.00  1  1.5441   6.7  1  46  4.4757  0  1.9345  12  10
 5.00  1  2.2355  10.1  1  50  4.9542  1  1.6628   4   9
 5.00  1  1.6812   6.5  1  74  3.7324  0  1.7324   5   9
 6.00  1  1.3617   9.0  1  77  3.5441  0  1.4624   1   8
 6.00  1  2.1139  10.2  0  70  3.5441  1  1.3617   1   8
 6.00  1  1.1139   9.7  1  60  3.5185  1  1.3979   0  10
 6.00  1  1.4150  10.4  1  67  3.9294  1  1.6902   0   8
 7.00  1  1.9777   9.5  1  48  3.3617  1  1.5682   5  10
 7.00  1  1.0414   5.1  0  61  3.7324  1  2.0000   1  10
 7.00  1  1.1761  11.4  1  53  3.7243  1  1.5185   1  13
 9.00  1  1.7243   8.2  1  55  3.7993  1  1.7404   0  12
11.00  1  1.1139  14.0  1  61  3.8808  1  1.2788   0  10
11.00  1  1.2304  12.0  1  43  3.7709  1  1.1761   1   9
11.00  1  1.3010  13.2  1  65  3.7993  1  1.8195   1  10
11.00  1  1.5682   7.5  1  70  3.8865  0  1.6721   0  12
11.00  1  1.0792   9.6  1  51  3.5051  1  1.9031   0   9
13.00  1  0.7782   5.5  0  60  3.5798  1  1.3979   2  10
14.00  1  1.3979  14.6  1  66  3.7243  1  1.2553   2  10
15.00  1  1.6021  10.6  1  70  3.6902  1  1.4314   0  11
16.00  1  1.3424   9.0  1  48  3.9345  1  2.0000   0  10
16.00  1  1.3222   8.8  1  62  3.6990  1  0.6990  17  10
17.00  1  1.2304  10.0  1  53  3.8808  1  1.4472   4   9
17.00  1  1.5911  11.2  1  68  3.4314  0  1.6128   1  10
18.00  1  1.4472   7.5  1  65  3.5682  0  0.9031   7   8
19.00  1  1.0792  14.4  1  51  3.9191  1  2.0000   6  15
19.00  1  1.2553   7.5  0  60  3.7924  1  1.9294   5   9
24.00  1  1.3010  14.6  1  56  4.0899  1  0.4771   0   9
25.00  1  1.0000  12.4  1  67  3.8195  1  1.6435   0  10
26.00  1  1.2304  11.2  1  49  3.6021  1  2.0000  27  11
32.00  1  1.3222  10.6  1  46  3.6990  1  1.6335   1   9
35.00  1  1.1139   7.0  0  48  3.6532  1  1.1761   4  10
37.00  1  1.6021  11.0  1  63  3.9542  0  1.2041   7   9
41.00  1  1.0000  10.2  1  69  3.4771  1  1.4771   6  10
41.00  1  1.1461   5.0  1  70  3.5185  1  1.3424   0   9
51.00  1  1.5682   7.7  0  74  3.4150  1  1.0414   4  13
```

```
   52.00  1  1.0000  10.1  1  60  3.8573  1  1.6532   4  10
   54.00  1  1.2553   9.0  1  49  3.7243  1  1.6990   2  10
   58.00  1  1.2041  12.1  1  42  3.6990  1  1.5798  22  10
   66.00  1  1.4472   6.6  1  59  3.7853  1  1.8195   0   9
   67.00  1  1.3222  12.8  1  52  3.6435  1  1.0414   1  10
   88.00  1  1.1761  10.6  1  47  3.5563  0  1.7559  21   9
   89.00  1  1.3222  14.0  1  63  3.6532  1  1.6232   1   9
   92.00  1  1.4314  11.0  1  58  4.0755  1  1.4150   4  11
    4.00  0  1.9542  10.2  1  59  4.0453  0  0.7782  12  10
    4.00  0  1.9243  10.0  1  49  3.9590  0  1.6232   0  13
    7.00  0  1.1139  12.4  1  48  3.7993  1  1.8573   0  10
    7.00  0  1.5315  10.2  1  81  3.5911  0  1.8808   0  11
    8.00  0  1.0792   9.9  1  57  3.8325  1  1.6532   0   8
   12.00  0  1.1461  11.6  1  46  3.6435  0  1.1461   0   7
   11.00  0  1.6128  14.0  1  60  3.7324  1  1.8451   3   9
   12.00  0  1.3979   8.8  1  66  3.8388  1  1.3617   0   9
   13.00  0  1.6628   4.9  0  71  3.6435  0  1.7924   0   9
   16.00  0  1.1461  13.0  1  55  3.8573  0  0.9031   0   9
   19.00  0  1.3222  13.0  1  59  3.7709  1  2.0000   1  10
   19.00  0  1.3222  10.8  1  69  3.8808  1  1.5185   0  10
   28.00  0  1.2304   7.3  1  82  3.7482  1  1.6721   0   9
   41.00  0  1.7559  12.8  1  72  3.7243  1  1.4472   1   9
   53.00  0  1.1139  12.0  1  66  3.6128  1  2.0000   1  11
   57.00  0  1.2553  12.5  1  66  3.9685  0  1.9542   0  11
   77.00  0  1.0792  14.0  1  60  3.6812  0  0.9542   0  12
   ;
```

The stepwise selection process consists of a series of alternating step-up and step-down phases. The former adds variables to the model, while the latter removes variables from the model.

Stepwise regression analysis is requested by specifying the SELECTION=STEPWISE option in the MODEL statement. The option SLENTRY=0.25 specifies that a variable has to be significant at the 0.25 level before it can be entered into the model, while the option SLSTAY=0.15 specifies that a variable in the model has to be significant at the 0.15 level for it to remain in the model. The DETAILS option requests detailed results for the variable selection process.

```
   proc phreg data=Myeloma;
      model Time*VStatus(0)=LogBUN HGB Platelet Age LogWBC
                            Frac LogPBM Protein SCalc
                            / selection=stepwise slentry=0.25
                              slstay=0.15 details;
   run;
```

Results of the stepwise regression analysis are displayed in Output 49.1.1 through Output 49.1.7.

Output 49.1.1. Individual Score Test Results for all Variables

```
                              The PHREG Procedure

                               Model Information

            Data Set                 WORK.MYELOMA
            Dependent Variable       Time             Survival Time
            Censoring Variable       VStatus          0=Alive 1=Dead
            Censoring Value(s)       0
            Ties Handling            BRESLOW


               Summary of the Number of Event and Censored Values

                                                       Percent
                     Total       Event    Censored    Censored

                        65          48          17       26.15


                     Analysis of Variables Not in the Model

                                       Score
                      Variable    Chi-Square    Pr > ChiSq

                      LogBUN          8.5164        0.0035
                      HGB             5.0664        0.0244
                      Platelet        3.1816        0.0745
                      Age             0.0183        0.8924
                      LogWBC          0.5658        0.4519
                      Frac            0.9151        0.3388
                      LogPBM          0.5846        0.4445
                      Protein         0.1466        0.7018
                      SCalc           1.1109        0.2919


                            Residual Chi-Square Test

                       Chi-Square       DF     Pr > ChiSq

                          18.4550        9         0.0302
```

Output 49.1.2. First Model in the Stepwise Selection Process

```
                              The PHREG Procedure

Step  1. Variable LogBUN is entered.  The model contains the following
         explanatory variables:

         LogBUN


                               Convergence Status

                 Convergence criterion (GCONV=1E-8) satisfied.


                             Model Fit Statistics

                                     Without           With
                    Criterion     Covariates     Covariates

                    -2 LOG L         309.716        301.959
                    AIC              309.716        303.959
                    SBC              309.716        305.830


                    Testing Global Null Hypothesis: BETA=0

            Test                 Chi-Square       DF     Pr > ChiSq

            Likelihood Ratio         7.7572        1         0.0053
            Score                    8.5164        1         0.0035
            Wald                     8.3392        1         0.0039


                    Analysis of Maximum Likelihood Estimates

                   Parameter     Standard                               Hazard
  Variable    DF     Estimate        Error   Chi-Square   Pr > ChiSq      Ratio

  LogBUN      1      1.74595      0.60460       8.3392       0.0039      5.731
```

Individual score tests are used to determine which of the nine explanatory variables is first selected into the model. In this case, the score test for each variable is the global score test for the model containing that variable as the only explanatory variable. The chi-squared statistic is compared to a chi-squared distribution with one degree of freedom. Output 49.1.1 displays the chi-squared statistics and the corresponding p-values. The variable LogBUN has the largest chi-squared value (8.5164), and it is significant ($p = 0.0035$) at the SLENTRY=0.25 level. The variable LogBUN is thus entered into the model. Output 49.1.2 displays the model results. Since the Wald chi-squared statistic is significant ($p = 0.0039$) at the SLSTAY=0.15 level, LogBUN stays in the model.

Output 49.1.3. Score Tests Adjusted for the Variable LogBUN

```
                      Analysis of Variables Not in the Model

                                      Score
                       Variable    Chi-Square    Pr > ChiSq

                       HGB             4.3468        0.0371
                       Platelet        2.0183        0.1554
                       Age             0.7159        0.3975
                       LogWBC          0.0704        0.7908
                       Frac            1.0354        0.3089
                       LogPBM          1.0334        0.3094
                       Protein         0.5214        0.4703
                       SCalc           1.4150        0.2342


                             Residual Chi-Square Test

                        Chi-Square       DF     Pr > ChiSq

                            9.3164        8         0.3163
```

Output 49.1.4. Second Model in the Stepwise Selection Process

```
Step  2. Variable HGB is entered.  The model contains the following explanatory
         variables:

         LogBUN  HGB


                               Convergence Status

                     Convergence criterion (GCONV=1E-8) satisfied.


                             Model Fit Statistics

                                  Without           With
                     Criterion     Covariates     Covariates

                     -2 LOG L         309.716        297.767
                     AIC              309.716        301.767
                     SBC              309.716        305.509


                     Testing Global Null Hypothesis: BETA=0

            Test                 Chi-Square       DF     Pr > ChiSq

            Likelihood Ratio        11.9493        2         0.0025
            Score                   12.7252        2         0.0017
            Wald                    12.1900        2         0.0023


                     Analysis of Maximum Likelihood Estimates

                        Parameter     Standard                               Hazard
  Variable    DF     Estimate        Error   Chi-Square   Pr > ChiSq      Ratio

  LogBUN       1      1.67440      0.61209       7.4833       0.0062      5.336
  HGB          1     -0.11899      0.05751       4.2811       0.0385      0.888
```

The next step consists of selecting another variable to add to the model. Output 49.1.3 displays the chi-squared statistics and p-values of individual score tests (adjusted for LogBUN) for the remaining eight variables. The score chi-square for a given variable is the value of the likelihood score test for testing the significance of the variable in the presence of LogBUN. The variable HGB is selected because it has the highest chi-squared value (4.3468), and it is significant ($p = 0.0371$) at the SLENTRY=0.25 level. Output 49.1.4 displays the fitted model containing both LogBUN and HGB. Based on the Wald statistics, neither LogBUN nor HGB is removed from the model.

Output 49.1.5. Third Model in the Stepwise Regression

```
Step  3. Variable SCalc is entered.  The model contains the following
         explanatory variables:

         LogBUN  HGB  SCalc


                              Convergence Status

                 Convergence criterion (GCONV=1E-8) satisfied.


                            Model Fit Statistics

                                 Without           With
                    Criterion     Covariates     Covariates

                    -2 LOG L         309.716        296.078
                    AIC              309.716        302.078
                    SBC              309.716        307.692


                    Testing Global Null Hypothesis: BETA=0

            Test                 Chi-Square       DF     Pr > ChiSq

            Likelihood Ratio        13.6377        3         0.0034
            Score                   15.3053        3         0.0016
            Wald                    14.4542        3         0.0023


                    Analysis of Maximum Likelihood Estimates

                   Parameter     Standard                                Hazard
  Variable   DF     Estimate        Error   Chi-Square   Pr > ChiSq      Ratio

  LogBUN      1      1.63593      0.62359       6.8822       0.0087      5.134
  HGB         1     -0.12643      0.05868       4.6419       0.0312      0.881
  SCalc       1      0.13286      0.09868       1.8127       0.1782      1.142
```

Output 49.1.5 shows Step 3 of the selection process, in which the variable SCalc is added, resulting in the model with LogBUN, HGB, and SCalc as the explanatory variables. Note that SCalc has the smallest Wald chi-squared statistic, and it is not significant ($p = 0.1782$) at the SLSTAY=0.15 level. The variable SCalc is then removed from the model in a step-down phase in Step 4 (Output 49.1.6). The removal of SCalc brings the stepwise selection process to a stop in order to avoid repeatedly entering and removing the same variable.

The procedure also displays a summary table of the steps in the stepwise selection process, as shown in Output 49.1.7.

The stepwise selection process results in a model with two explanatory variables, LogBUN and HGB.

Output 49.1.6. Final Model in the Stepwise Regression

```
Step  4. Variable SCalc is removed.  The model contains the following
         explanatory variables:

         LogBUN  HGB


                              Convergence Status

                Convergence criterion (GCONV=1E-8) satisfied.


                            Model Fit Statistics

                                    Without           With
                   Criterion     Covariates     Covariates

                   -2 LOG L         309.716        297.767
                   AIC              309.716        301.767
                   SBC              309.716        305.509


                   Testing Global Null Hypothesis: BETA=0

           Test                 Chi-Square       DF     Pr > ChiSq

           Likelihood Ratio        11.9493        2         0.0025
           Score                   12.7252        2         0.0017
           Wald                    12.1900        2         0.0023


                   Analysis of Maximum Likelihood Estimates

                  Parameter     Standard                                 Hazard
  Variable   DF    Estimate        Error    Chi-Square   Pr > ChiSq       Ratio

  LogBUN      1     1.67440      0.61209        7.4833       0.0062       5.336
  HGB         1    -0.11899      0.05751        4.2811       0.0385       0.888


NOTE: Model building terminates because the variable to be entered is the
      variable that was removed in the last step.
```

Output 49.1.7. Model Selection Summary

```
                          Summary of Stepwise Selection

            Variable          Number         Score          Wald
Step    Entered     Removed       In     Chi-Square    Chi-Square    Pr > ChiSq

   1    LogBUN                     1         8.5164         .            0.0035
   2    HGB                        2         4.3468         .            0.0371
   3    SCalc                      3         1.8225         .            0.1770
   4                SCalc          2          .            1.8127        0.1782
```

Example 49.2. Best Subset Selection

An alternative to stepwise selection of variables is best subset selection. The procedure uses the branch and bound algorithm of Furnival and Wilson (1974) to find a specified number of best models containing one, two, three variables and so on, up to the single model containing all of the explanatory variables. The criterion used to determine "best" is based on the global score chi-squared statistic. For two models A and B, each having the same number of explanatory variables, model A is considered to be better than model B if the global score chi-squared statistic for A exceeds that for B.

Best subset selection analysis is requested by specifying the SELECTION=SCORE option in the MODEL statement. The BEST=3 option requests the procedure to identify only the three best models for each size. In other words, PROC PHREG will list the three models having the highest score statistics of all the models possible for a given number of covariates.

```
proc phreg data=Myeloma;
   model Time*VStatus(0)=LogBUN HGB Platelet Age LogWBC
                         Frac LogPBM Protein SCalc
                         / selection=score best=3;
run;
```

Output 49.2.1 displays the results of this analysis. The number of explanatory variables in the model is given in the first column, and the names of the variables are listed on the right. The models are listed in descending order of their score chi-squared values within each model size. For example, among all models containing two explanatory variables, the model that contains the variables LogBUN and HGB has the largest score value (12.7252), the model that contains the variables LogBUN and Platelet has the second largest score value (11.1842), and the model that contains the variables LogBUN and SCalc has the third largest score value (9.9962).

Output 49.2.1. Best Variable Combinations

```
                              The PHREG Procedure

                 Regression Models Selected by Score Criterion

Number of       Score
Variables  Chi-Square  Variables Included in Model

        1      8.5164  LogBUN
        1      5.0664  HGB
        1      3.1816  Platelet
-------------------------------------------------------------------------------
        2     12.7252  LogBUN HGB
        2     11.1842  LogBUN Platelet
        2      9.9962  LogBUN SCalc
-------------------------------------------------------------------------------
        3     15.3053  LogBUN HGB SCalc
        3     13.9911  LogBUN HGB Age
        3     13.5788  LogBUN HGB Frac
-------------------------------------------------------------------------------
        4     16.9873  LogBUN HGB Age SCalc
        4     16.0457  LogBUN HGB Frac SCalc
        4     15.7619  LogBUN HGB LogPBM SCalc
-------------------------------------------------------------------------------
        5     17.6291  LogBUN HGB Age Frac SCalc
        5     17.3519  LogBUN HGB Age LogPBM SCalc
        5     17.1922  LogBUN HGB Age LogWBC SCalc
-------------------------------------------------------------------------------
        6     17.9120  LogBUN HGB Age Frac LogPBM SCalc
        6     17.7947  LogBUN HGB Age LogWBC Frac SCalc
        6     17.7744  LogBUN HGB Platelet Age Frac SCalc
-------------------------------------------------------------------------------
        7     18.1517  LogBUN HGB Platelet Age Frac LogPBM SCalc
        7     18.0568  LogBUN HGB Age LogWBC Frac LogPBM SCalc
        7     18.0223  LogBUN HGB Platelet Age LogWBC Frac SCalc
-------------------------------------------------------------------------------
        8     18.3925  LogBUN HGB Platelet Age LogWBC Frac LogPBM SCalc
        8     18.1636  LogBUN HGB Platelet Age Frac LogPBM Protein SCalc
        8     18.1309  LogBUN HGB Platelet Age LogWBC Frac Protein SCalc
-------------------------------------------------------------------------------
        9     18.4550  LogBUN HGB Platelet Age LogWBC Frac LogPBM Protein SCalc

--------------------------------------------------------------------------------
```

Example 49.3. Conditional Logistic Regression for m:n Matching

Conditional logistic regression is used to investigate the relationship between an outcome and a set of prognostic factors in matched case-control studies. The outcome is whether the subject is a case or a control. If there is only one case and one control, the matching is 1:1. The *m:n* matching refers to the situation in which there is a varying number of cases and controls in the matched sets. You can perform conditional logistic regression with the PHREG procedure by using the discrete logistic model and forming a stratum for each matched set. In addition, you need to create dummy survival times so that all the cases in a matched set have the same event time value, and the corresponding controls are censored at later times.

Consider the following set of low infant birth-weight data extracted from Appendix 1 of Hosmer and Lemeshow (1989). These data represent 189 women, of whom 59 had low birth-weight babies and 130 had normal weight babies. Under investigation are the following risk factors: weight in pounds at the last menstrual period (LWT), presence of hypertension (HT), smoking status during pregnancy (Smoke), and presence of uterine irritability (UI). For HT, Smoke, and UI, a value of 1 indicates a “yes” and a value of 0 indicates a “no.” The woman’s age (Age) is used as the matching variable. The SAS data set LBW contains a subset of the data corresponding to women between the ages of 16 and 32.

```
data LBW;
   input id Age Low LWT Smoke HT UI @@;
   Time=2-Low;
   datalines;
 25  16   1   130   0  0  0    143  16   0   110   0  0  0
166  16   0   112   0  0  0    167  16   0   135   1  0  0
189  16   0   135   1  0  0    206  16   0   170   0  0  0
216  16   0    95   0  0  0     37  17   1   130   1  0  1
 45  17   1   110   1  0  0     68  17   1   120   1  0  0
 71  17   1   120   0  0  0     83  17   1   142   0  1  0
 93  17   0   103   0  0  0    113  17   0   122   1  0  0
116  17   0   113   0  0  0    117  17   0   113   0  0  0
147  17   0   119   0  0  0    148  17   0   119   0  0  0
180  17   0   120   1  0  0     49  18   1   148   0  0  0
 50  18   1   110   1  0  0     89  18   0   107   1  0  1
100  18   0   100   1  0  0    101  18   0   100   1  0  0
132  18   0    90   1  0  1    133  18   0    90   1  0  1
168  18   0   229   0  0  0    205  18   0   120   1  0  0
208  18   0   120   0  0  0     23  19   1    91   1  0  1
 33  19   1   102   0  0  0     34  19   1   112   1  0  1
 85  19   0   182   0  0  1     96  19   0    95   0  0  0
 97  19   0   150   0  0  0    124  19   0   138   1  0  0
129  19   0   189   0  0  0    135  19   0   132   0  0  0
142  19   0   115   0  0  0    181  19   0   105   0  0  0
187  19   0   235   1  1  0    192  19   0   147   1  0  0
193  19   0   147   1  0  0    197  19   0   184   1  1  0
224  19   0   120   1  0  0     27  20   1   150   1  0  0
 31  20   1   125   0  0  1     40  20   1   120   1  0  0
 44  20   1    80   1  0  1     47  20   1   109   0  0  0
 51  20   1   121   1  0  1     60  20   1   122   1  0  0
 76  20   1   105   0  0  0     87  20   0   105   1  0  0
104  20   0   120   0  0  1    146  20   0   103   0  0  0
155  20   0   169   0  0  1    160  20   0   141   0  0  1
172  20   0   121   1  0  0    177  20   0   127   0  0  0
201  20   0   120   0  0  0    211  20   0   170   1  0  0
217  20   0   158   0  0  0     20  21   1   165   1  1  0
 28  21   1   200   0  0  1     30  21   1   103   0  0  0
 52  21   1   100   0  0  0     84  21   1   130   1  1  0
 88  21   0   108   1  0  1     91  21   0   124   0  0  0
128  21   0   185   1  0  0    131  21   0   160   0  0  0
144  21   0   110   1  0  1    186  21   0   134   0  0  0
219  21   0   115   0  0  0     42  22   1   130   1  0  1
 67  22   1   130   1  0  0     92  22   0   118   0  0  0
 98  22   0    95   0  1  0    137  22   0    85   1  0  0
138  22   0   120   0  1  0    140  22   0   130   1  0  0
161  22   0   158   0  0  0    162  22   0   112   1  0  0
174  22   0   131   0  0  0    184  22   0   125   0  0  0
204  22   0   169   0  0  0    220  22   0   129   0  0  0
 17  23   1    97   0  0  1     59  23   1   187   1  0  0
 63  23   1   120   0  0  0    69  23   1   110   1  0  0
 82  23   1    94   1  0  0   130  23   0   130   0  0  0
139  23   0   128   0  0  0   149  23   0   119   0  0  0
164  23   0   115   1  0  0   173  23   0   190   0  0  0
179  23   0   123   0  0  0   182  23   0   130   0  0  0
200  23   0   110   0  0  0    18  24   1   128   0  0  0
 19  24   1   132   0  1  0    29  24   1   155   1  0  0
 36  24   1   138   0  0  0    61  24   1   105   1  0  0
118  24   0    90   1  0  0   136  24   0   115   0  0  0
150  24   0   110   0  0  0   156  24   0   115   0  0  0
185  24   0   133   0  0  0   196  24   0   110   0  0  0
199  24   0   110   0  0  0   225  24   0   116   0  0  0
 13  25   1   105   0  1  0    15  25   1    85   0  0  1
 24  25   1   115   0  0  0    26  25   1    92   1  0  0
 32  25   1    89   0  0  0    46  25   1   105   0  0  0
103  25   0   118   1  0  0   111  25   0   120   0  0  1
120  25   0   155   0  0  0   121  25   0   125   0  0  0
169  25   0   140   0  0  0   188  25   0    95   1  0  1
202  25   0   241   0  1  0   215  25   0   120   0  0  0
221  25   0   130   0  0  0    35  26   1   117   1  0  0
 54  26   1    96   0  0  0    75  26   1   154   0  1  0
 77  26   1   190   1  0  0    95  26   0   113   1  0  0
115  26   0   168   1  0  0   154  26   0   133   1  0  0
218  26   0   160   0  0  0    16  27   1   150   0  0  0
 43  27   1   130   0  0  1   125  27   0   124   1  0  0
  4  28   1   120   1  0  1    79  28   1    95   1  0  0
105  28   0   120   1  0  0   109  28   0   120   0  0  0
112  28   0   167   0  0  0   151  28   0   140   0  0  0
159  28   0   250   1  0  0   212  28   0   134   0  0  0
214  28   0   130   0  0  0    10  29   1   130   0  0  1
 94  29   0   123   1  0  0   114  29   0   150   0  0  0
123  29   0   140   1  0  0   190  29   0   135   0  0  0
191  29   0   154   0  0  0   209  29   0   130   1  0  0
 65  30   1   142   1  0  0    99  30   0   107   0  0  1
141  30   0    95   1  0  0   145  30   0   153   0  0  0
176  30   0   110   0  0  0   195  30   0   137   0  0  0
203  30   0   112   0  0  0    56  31   1   102   1  0  0
107  31   0   100   0  0  1   126  31   0   215   1  0  0
163  31   0   150   1  0  0   222  31   0   120   0  0  0
 22  32   1   105   1  0  0   106  32   0   121   0  0  0
134  32   0   132   0  0  0   170  32   0   134   1  0  0
175  32   0   170   0  0  0   207  32   0   186   0  0  0
;
```

The variable Low is used to determine whether the subject is a case (Low=1, low birth-weight baby) or a control (Low=0, normal weight baby). The dummy time variable Time takes the value 1 for cases and 2 for controls.

The following SAS statements produce a conditional logistic regression analysis of the data. The variable Time is the response, and Low is the censoring variable. Note that the data set is created so that all the cases have the same event time, and the controls have later censored times. The matching variable Age is used in the STRATA statement so each unique age value defines a stratum. The variables LWT, Smoke, HT, and UI are specified as explanatory variables. The TIES=DISCRETE option requests the discrete logistic model.

```
proc phreg data=LBW;
   model Time*Low(0)= LWT Smoke HT UI / ties=discrete;
   strata Age;
run;
```

The procedure displays a summary of the number of event and censored observations for each stratum. These are the number of cases and controls for each matched set shown in Output 49.3.1. Results of the conditional logistic regression analysis are shown in Output 49.3.2. Based on the Wald test for individual variables, the variables LWT, Smoke, and HT are statistically significant while UI is marginal.

The hazards ratios, computed by exponentiating the parameter estimates, are useful in interpreting the results of the analysis. If the hazards ratio of a prognostic factor is larger than 1, an increment in the factor increases the hazard rate. If the hazards ratio is less than 1, an increment in the factor decreases the hazard rate. Results indicate that women were more likely to have low birth-weight babies if they were underweight in the last menstrual cycle, were hypertensive, smoked during pregnancy, or suffered uterine irritability.

For matched case-control studies with one case per matched set (1:*n* matching), the likelihood function for the conditional logistic regression reduces to that of the Cox model for the continuous time scale. For this situation, you can use the default TIES=BRESLOW.

Output 49.3.1. Summary of Number of Case and Controls

```
                              The PHREG Procedure

                               Model Information

                       Data Set                 WORK.LBW
                       Dependent Variable       Time
                       Censoring Variable       Low
                       Censoring Value(s)       0
                       Ties Handling            DISCRETE


               Summary of the Number of Event and Censored Values

                                                                  Percent
      Stratum    Age             Total        Event     Censored     Censored

            1    16                  7            1            6        85.71
            2    17                 12            5            7        58.33
            3    18                 10            2            8        80.00
            4    19                 16            3           13        81.25
            5    20                 18            8           10        55.56
            6    21                 12            5            7        58.33
            7    22                 13            2           11        84.62
            8    23                 13            5            8        61.54
            9    24                 13            5            8        61.54
           10    25                 15            6            9        60.00
           11    26                  8            4            4        50.00
           12    27                  3            2            1        33.33
           13    28                  9            2            7        77.78
           14    29                  7            1            6        85.71
           15    30                  7            1            6        85.71
           16    31                  5            1            4        80.00
           17    32                  6            1            5        83.33
      -------------------------------------------------------------------
        Total                      174           54          120        68.97
```

Output 49.3.2. Conditional Logistic Regression Analysis for the Low Birth-Weight Study

```
                              The PHREG Procedure

                               Convergence Status

                 Convergence criterion (GCONV=1E-8) satisfied.


                             Model Fit Statistics

                                     Without           With
                    Criterion     Covariates     Covariates

                    -2 LOG L         159.069        141.108
                    AIC              159.069        149.108
                    SBC              159.069        157.064


                    Testing Global Null Hypothesis: BETA=0

            Test                 Chi-Square       DF     Pr > ChiSq

            Likelihood Ratio        17.9613        4         0.0013
            Score                   17.3152        4         0.0017
            Wald                    15.5577        4         0.0037


                    Analysis of Maximum Likelihood Estimates

                   Parameter     Standard                                 Hazard
Variable   DF       Estimate        Error   Chi-Square   Pr > ChiSq        Ratio

LWT         1       -0.01498      0.00706       4.5001       0.0339        0.985
Smoke       1        0.80805      0.36797       4.8221       0.0281        2.244
HT          1        1.75143      0.73932       5.6120       0.0178        5.763
UI          1        0.88341      0.48032       3.3827       0.0659        2.419
```

Example 49.4. Model Using Time-Dependent Explanatory Variables

Time-dependent variables can be used to model the effects of subjects transferring from one treatment group to another. One example of the need for such strategies is the Stanford heart transplant program. Patients are accepted if physicians judge them suitable for heart transplant. Then, when a donor becomes available, physicians choose transplant recipients according to various medical criteria. A patient's status can be changed during the study from waiting for a transplant to transplant recipient. Transplant status can be defined by the time-dependent covariate function $z = z(t)$ as

$$z(t) = \begin{cases} 0 & \text{if the patient has not received the transplant at time } t \\ 1 & \text{if the patient has received the transplant at time } t \end{cases}$$

The Stanford heart transplant data that appear in Crowley and Hu (1977) consist of 103 patients, 69 of whom received transplants. The data are saved in a SAS data set called Heart. For each patient in the program, there is a birth date (Bir_Date), a date of acceptance (Acc_Date), and a date last seen (Ter_Date). The survival time

(Time) in days is defined as Time = Ter_Date − Acc_Date. The survival time is said to be uncensored (Status=1) or censored (Status=0), depending on whether Ter_Date is the date of death or the closing date of the study. The age in years at acceptance into the program is Acc_Age = (Acc_Date − Bir_Date) / 365. Previous open-heart surgery for each patient is indicated by the variable PrevSurg. For each transplant recipient, there is a date of transplant (Xpl_Date) and three measures (NMismatch, Antigen, Mismatch) of tissue-type mismatching. The waiting period (WaitTime) in days for a transplant recipient is calculated as WaitTime = Xpl_Date − Acc_Date, and the age in years at transplant is Xpl_Age = (Xpl_Date − Bir_Date) / 365. For those who do not receive heart transplants, the WaitTime, Xpl_Age, NMismatch, Antigen, and Mismatch variables contain missing values.

The input data contains dates that have a two-digit year representation. The SAS option YEARCUTOFF=1900 is specified to ensure that a two-digit year xx is year 19xx.

The code is as follows:

```
options yearcutoff=1900;

data Heart;
   input ID
         @5  Bir_Date mmddyy8.
         @14 Acc_Date mmddyy8.
         @23 Xpl_Date mmddyy8.
         @32 Ter_Date mmddyy8.
         @41 Status 1.
         @43 PrevSurg 1.
         @45 NMismatch 1.
         @47 Antigen 1.
         @49 Mismatch 4.
         @54 Reject 1.
         @56 NotTyped $1.;
   label Bir_Date =’Date of birth’
         Acc_Date =’Date of acceptance’
         Xpl_Date =’Date of transplant’
         Ter_Date =’Date last seen’
         Status   =  ’Dead=1 Alive=0’
         PrevSurg =’Previous surgery’
         NMismatch= ’No of mismatches’
         Antigen  = ’HLA-A2 antigen’
         Mismatch =’Mismatch score’
         NotTyped = ’y=not tissue-typed’;
   Time= Ter_Date - Acc_Date;
   Acc_Age=int( (Acc_Date - Bir_Date)/365 );
   if ( Xpl_Date ne .) then do;
      WaitTime= Xpl_Date - Acc_Date;
      Xpl_Age= int( (Xpl_Date - Bir_Date)/365 );
   end;
   datalines;
  1 01 10 37 11 15 67          01 03 68 1 0
  2 03 02 16 01 02 68          01 07 68 1 0
  3 09 19 13 01 06 68 01 06 68 01 21 68 1 0 2 0 1.11 0
  4 12 23 27 03 28 68 05 02 68 05 05 68 1 0 3 0 1.66 0
  5 07 28 47 05 10 68          05 27 68 1 0
  6 11 18 13 06 13 68          06 15 68 1 0
```

```
 7 08 29 17 07 12 68 08 31 68 05 17 70 1 0 4 0 1.32 1
 8 03 27 23 08 01 68          09 09 68 1 0
 9 06 11 21 08 09 68          11 01 68 1 0
10 02 09 26 08 11 68 08 22 68 10 07 68 1 0 2 0 0.61 1
11 08 22 20 08 15 68 09 09 68 01 14 69 1 0 1 0 0.36 0
12 07 09 15 09 17 68          09 24 68 1 0
13 02 22 14 09 19 68 10 05 68 12 08 68 1 0 3 0 1.89 1
14 09 16 14 09 20 68 10 26 68 07 07 72 1 0 1 0 0.87 1
15 12 04 14 09 27 68          09 27 68 1 1
16 05 16 19 10 26 68 11 22 68 08 29 69 1 0 2 0 1.12 1
17 06 29 48 10 28 68          12 02 68 1 0
18 12 27 11 11 01 68 11 20 68 12 13 68 1 0 3 0 2.05 0
19 10 04 09 11 18 68          12 24 68 1 0
20 10 19 13 01 29 69 02 15 69 02 25 69 1 0 3 1 2.76 1
21 09 29 25 02 01 69 02 08 69 11 29 71 1 0 2 0 1.13 1
22 06 05 26 03 18 69 03 29 69 05 07 69 1 0 3 0 1.38 1
23 12 02 10 04 11 69 04 13 69 04 13 71 1 0 3 0 0.96 1
24 07 07 17 04 25 69 07 16 69 11 29 69 1 0 3 1 1.62 1
25 02 06 36 04 28 69 05 22 69 04 01 74 0 0 2 0 1.06 0
26 10 18 38 05 01 69          03 01 73 0 0
27 07 21 60 05 04 69          01 21 70 1 0
28 05 30 15 06 07 69 08 16 69 08 17 69 1 0 2 0 0.47 0
29 02 06 19 07 14 69          08 17 69 1 0
30 09 20 24 08 19 69 09 03 69 12 18 71 1 0 4 0 1.58 1
31 10 04 14 08 23 69          09 07 69 1 0
32 04 02 05 08 29 69 09 14 69 11 13 69 1 0 4 0 0.69 1
33 01 01 21 11 27 69 01 16 70 04 01 74 0 0 3 0 0.91 0
34 05 24 29 12 12 69 01 03 70 04 01 74 0 0 2 0 0.38 0
35 08 04 26 01 21 70          02 01 70 1 0
36 05 01 21 04 04 70 05 19 70 07 12 70 1 0 2 0 2.09 1
37 10 24 08 04 25 70 05 13 70 06 29 70 1 0 3 1 0.87 1
38 11 14 28 05 05 70 05 09 70 05 09 70 1 0 3 0 0.87 0
39 11 12 19 05 20 70 05 21 70 07 11 70 1 0             y
40 11 30 21 05 25 70 07 04 70 04 01 74 0 1 4 0 0.75 0
41 04 30 25 08 19 70 10 15 70 04 01 74 0 1 2 0 0.98 0
42 03 13 34 08 21 70          08 23 70 1 0
43 06 01 27 10 22 70          10 23 70 1 1
44 05 02 28 11 30 70          01 08 71 1 1
45 10 30 34 01 05 71 01 05 71 02 18 71 1 0 1 0 0.0  0
46 06 01 22 01 10 71 01 11 71 10 01 73 1 1 2 0 0.81 1
47 12 28 23 02 02 71 02 22 71 04 14 71 1 0 3 0 1.38 1
48 01 23 15 02 05 71          02 13 71 1 0
49 06 21 34 02 15 71 03 22 71 04 01 74 0 1 4 0 1.35 0
50 03 28 25 02 15 71 05 08 71 10 21 73 1 1             y
51 06 29 22 03 24 71 04 24 71 01 02 72 1 0 4 1 1.08 1
52 01 24 30 04 25 71          08 04 71 1 0
53 02 27 24 07 02 71 08 11 71 01 05 72 1 0             y
54 09 16 23 07 02 71          07 04 71 1 0
55 02 24 19 08 09 71 08 18 71 10 08 71 1 0 2 0 1.51 1
56 12 05 32 09 03 71 11 08 71 04 01 74 0 0 4 0 0.98 0
57 06 08 30 09 13 71          02 08 72 1 0
58 09 17 23 09 23 71 10 13 71 08 30 72 1 1 2 1 1.82 1
59 05 12 30 09 29 71 12 15 71 04 01 74 0 1 2 0 0.19 0
60 10 29 22 11 18 71 11 20 71 01 09 72 1 0 3 0 0.66 1
61 05 12 19 12 04 71          12 05 71 1 0
62 08 01 32 12 09 71          02 15 72 1 0
63 04 15 39 12 12 71 01 07 72 04 01 74 0 0 3 1 1.93 0
64 04 09 23 02 01 72 03 04 72 09 06 73 1 1 1 0 0.12 0
65 11 19 20 03 06 72 03 17 72 05 22 72 1 0 2 0 1.12 1
```

```
 66 01 02 19 03 20 72          04 20 72 1 0
 67 09 03 52 03 23 72 05 18 72 01 01 73 1 0 3 0 1.02 0
 68 01 10 27 04 07 72 04 09 72 06 13 72 1 0 3 1 1.68 1
 69 06 05 24 06 01 72 06 10 72 04 01 74 0 0 2 0 1.20 0
 70 06 17 19 06 17 72 06 21 72 07 16 72 1 0 3 1 1.68 1
 71 02 22 25 07 21 72 08 20 72 04 01 74 0 0 3 0 0.97 0
 72 11 22 45 08 14 72 08 17 72 04 01 74 0 0 3 1 1.46 0
 73 05 13 16 09 11 72 10 07 72 12 09 72 1 0 3 1 2.16 1
 74 07 20 43 09 18 72 09 22 72 10 04 72 1 0 1 0 0.61 0
 75 07 25 20 09 29 72          09 30 72 1 0
 76 09 03 20 10 04 72 11 18 72 04 01 74 0 1 3 1 1.70 0
 77 08 27 31 10 06 72          10 26 72 1 0
 78 02 20 24 11 03 72 05 31 73 04 01 74 0 0 3 0 0.81 0
 79 02 18 19 11 30 72 02 04 73 03 05 73 1 0 2 0 1.08 1
 80 06 27 26 12 06 72 12 31 72 04 01 74 0 1 3 0 1.41 0
 81 02 21 20 01 12 73 01 17 73 04 01 74 0 0 4 1 1.94 0
 82 09 19 42 11 01 71          01 01 73 0 0
 83 10 04 19 01 24 73 02 24 73 04 13 73 1 0 4 0 3.05 0
 84 05 13 30 01 30 73 03 07 73 12 29 73 1 0 4 0 0.60 1
 85 02 13 25 02 06 73          02 10 73 1 0
 86 03 30 24 03 01 73 03 08 73 04 01 74 0 0 3 1 1.44 0
 87 12 19 26 03 21 73 05 19 73 07 08 73 1 0 2 0 2.25 1
 88 11 16 18 03 28 73 04 27 73 04 01 74 0 0 3 0 0.68 0
 89 03 19 22 04 05 73 08 21 73 10 28 73 1 0 4 1 1.33 1
 90 03 25 21 04 06 73 09 12 73 10 08 73 1 1 3 1 0.82 0
 91 09 08 25 04 13 73          03 18 74 1 0
 92 05 03 28 04 27 73 03 02 74 04 01 74 0 0 1 0 0.16 0
 93 10 10 25 07 11 73 08 07 73 04 01 74 0 0 2 0 0.33 0
 94 11 11 29 09 14 73 09 17 73 02 25 74 1 1 3 0 1.20 1
 95 06 11 33 09 22 73 09 23 73 10 07 73 1 0           y
 96 02 09 47 10 04 73 10 16 73 04 01 74 0 0 2 0 0.46 0
 97 04 11 50 11 22 73 12 12 73 04 01 74 0 0 3 1 1.78 0
 98 04 28 45 12 14 73 03 19 74 04 01 74 0 0 4 1 0.77 0
 99 02 24 24 12 25 73          01 14 74 1 0
100 01 31 39 02 22 74 03 31 74 04 01 74 0 1 3 0 0.67 0
101 08 25 24 03 02 74          04 01 74 0 0
102 10 30 33 03 22 74          04 01 74 0 0
103 05 20 28 09 13 67          09 18 67 1 0
;
```

Crowley and Hu (1977) have presented a number of analyses to assess the effects of various explanatory variables on the survival of patients. This example fits two of the models that they have considered.

The first model consists of two explanatory variables—the transplant status and the age at acceptance. The transplant status (XStatus) is a time-dependent variable defined by the programming statements between the MODEL statement and the RUN statement. The XStatus variable takes the value 1 or 0 at time t (measured from the date of acceptance), depending on whether or not the patient has received a transplant at that time. Note that the value of XStatus changes for subjects in each risk set (subjects still alive just before each distinct event time); therefore, the variable cannot be created in the DATA step. The variable Acc_Age, which is not time-dependent, accounts for the possibility that pretransplant risks vary with age.

```
proc phreg data= Heart;
   model Time*Status(0)= XStatus Acc_Age;
   if (WaitTime = . or Time < WaitTime) then XStatus=0.;
   else  XStatus= 1.0;
run;
```

Output 49.4.1. Heart Transplant Study Analysis I

```
                            The PHREG Procedure

                             Model Information

           Data Set                 WORK.HEART
           Dependent Variable       Time
           Censoring Variable       Status         Dead=1 Alive=0
           Censoring Value(s)       0
           Ties Handling            BRESLOW


             Summary of the Number of Event and Censored Values

                                                     Percent
                   Total       Event    Censored    Censored

                     103          75          28       27.18


                             Convergence Status

               Convergence criterion (GCONV=1E-8) satisfied.


                           Model Fit Statistics

                                 Without           With
                  Criterion     Covariates     Covariates

                  -2 LOG L         596.649        591.312
                  AIC              596.649        595.312
                  SBC              596.649        599.947


                  Testing Global Null Hypothesis: BETA=0

          Test                 Chi-Square       DF     Pr > ChiSq

          Likelihood Ratio         5.3370        2         0.0694
          Score                    4.7900        2         0.0912
          Wald                     4.7812        2         0.0916
```

```
                            The PHREG Procedure

                  Analysis of Maximum Likelihood Estimates

                 Parameter     Standard                                 Hazard
Variable   DF     Estimate        Error   Chi-Square   Pr > ChiSq        Ratio

XStatus     1     -0.06046      0.30572       0.0391       0.8432        0.941
Acc_Age     1      0.03147      0.01445       4.7443       0.0294        1.032
```

Results of this analysis are shown in Output 49.4.1. Transplantation appears to be associated with a slight decrease in risk, although the effect is not significant ($p = 0.8432$). The age at acceptance as a pretransplant risk factor adds significantly to the model ($p = 0.0294$). The risk increases significantly with age at acceptance.

The second model consists of three explanatory variables—the transplant status, the transplant age, and the mismatch score. Four transplant recipients who were not typed have no Mismatch values; they are excluded from the analysis by the use of a WHERE clause. The transplant age (XAge) and the mismatch score (XScore) are also time-dependent and are defined in a fashion similar to that of XStatus. While the patient is waiting for a transplant, XAge and XScore have a value of 0. After the patient has migrated to the recipient population, XAge takes on the value of Xpl_Age (transplant age for the recipient), and XScore takes on the value of Mismatch (a measure of the degree of dissimilarity between donor and recipient).

```
proc phreg data= Heart;
   model Time*Status(0)= XStatus XAge XScore;
   where NotTyped ^= ’y’;
   if (WaitTime = . or Time < WaitTime) then do;
      XStatus=0.;
      XAge=0.;
      XScore= 0.;
   end;
   else do;
      XStatus= 1.0;
      XAge= Xpl_Age;
      XScore= Mismatch;
   end;
run;
```

Output 49.4.2. Heart Transplant Study Analysis II

```
                              The PHREG Procedure

                               Model Information

             Data Set                 WORK.HEART
             Dependent Variable       Time
             Censoring Variable       Status         Dead=1 Alive=0
             Censoring Value(s)       0
             Ties Handling            BRESLOW


               Summary of the Number of Event and Censored Values

                                                     Percent
                     Total       Event    Censored    Censored

                        99          71          28       28.28


                               Convergence Status

                 Convergence criterion (GCONV=1E-8) satisfied.


                             Model Fit Statistics

                                   Without           With
                    Criterion     Covariates     Covariates

                    -2 LOG L         561.646        551.911
                    AIC              561.646        557.911
                    SBC              561.646        564.699


                    Testing Global Null Hypothesis: BETA=0

            Test                 Chi-Square       DF     Pr > ChiSq

            Likelihood Ratio         9.7350        3         0.0210
            Score                    9.0127        3         0.0291
            Wald                     9.0156        3         0.0291


                    Analysis of Maximum Likelihood Estimates

                       Parameter     Standard                                 Hazard
  Variable    DF        Estimate        Error   Chi-Square   Pr > ChiSq       Ratio

  XStatus      1        -3.17799      1.18612       7.1787       0.0074       0.042
  XAge         1         0.05517      0.02259       5.9649       0.0146       1.057
  XScore       1         0.44424      0.28026       2.5125       0.1129       1.559
```

Results of the analysis are shown in Output 49.4.2. Note that only 99 patients are included in this analysis, instead of 103 patients as in the previous analysis, since four transplant recipients who were not typed are excluded. The variable XAge is statistically significant ($p = 0.0146$) with a hazards ratio exceeding 1. Therefore, patients who had a transplant at younger ages lived longer than those who received a transplant later in their lives. The variable XScore has only minimal effect on the survival ($p = 0.1129$).

Example 49.5. Time-Dependent Repeated Measurements

Repeated determinations may be made during the course of a study of variables thought to be related to survival. Consider an experiment to study the dosing effect of a tumor-promoting agent. Forty-five rodents initially exposed to a carcinogen were randomly assigned to three dose groups. After the first death of an animal, the rodents were examined every week for the number of papillomas. Investigators were interested in determining the effects of dose on the carcinoma incidence after adjusting for the number of papillomas.

The input data set TUMOR consists of the following 19 variables:

- ID (subject identification)
- Time (survival time of the subject)
- Dead (censoring status where 1=dead and 0=censored)
- Dose (dose of the tumor-promoting agent)
- P1–P15 (number of papillomas at the 15 times that animals died. These 15 death times are weeks 27, 34, 37, 41, 43, 45, 46, 47, 49, 50, 51, 53, 65, 67, and 71. For instance, subject 1 died at week 47; it had no papilloma at week 27, five papillomas at week 34, six at week 37, eight at week 41, and 10 at weeks 43, 45, 46, and 47. For an animal that died before week 71, the number of papillomas is missing for those times beyond its death.)

The following SAS statements create the data set TUMOR:

```
data Tumor;
   infile datalines missover;
   input ID Time Dead Dose P1-P15;
   label ID='Subject ID';
   datalines;
 1 47 1  1.0  0  5  6  8 10 10 10 10
 2 71 1  1.0  0  0  0  0  0  0  0  0  1  1  1  1 1 1 1
 3 81 0  1.0  0  1  1  1  1  1  1  1  1  1  1  1 1 1 1
 4 81 0  1.0  0  0  0  0  0  0  0  0  0  0  0  0 0 0 0
 5 81 0  1.0  0  0  0  0  0  0  0  0  0  0  0  0 0 0 0
 6 65 1  1.0  0  0  0  1  1  1  1  1  1  1  1  1 1
 7 71 0  1.0  0  0  0  0  0  0  0  0  0  0  0  0 0 0 0
 8 69 0  1.0  0  0  0  0  0  0  0  0  0  0  0  0 0 0
 9 67 1  1.0  0  0  1  1  2  2  2  2  3  3  3  3 3 3
10 81 0  1.0  0  0  0  0  0  0  0  0  0  0  0  0 0 0 0
11 37 1  1.0  9  9  9
12 81 0  1.0  0  0  0  0  0  0  0  0  0  0  0  0 0 0 0
13 77 0  1.0  0  0  0  0  1  1  1  1  1  1  1  1 1 1 1
14 81 0  1.0  0  0  0  0  0  0  0  0  0  0  0  0 0 0 0
15 81 0  1.0  0  0  0  0  0  0  0  0  0  0  0  0 0 0 0
16 54 0  2.5  0  1  1  1  2  2  2  2  2  2  2  2
17 53 0  2.5  0  0  0  0  0  0  0  0  0  0  0  0
18 38 0  2.5  5 13 14
19 54 0  2.5  2  6  6  6  6  6  6  6  6  6  6  6
```

```
20 51 1  2.5 15 15 15 16 16 17 17 17 17 17 17
21 47 1  2.5 13 20 20 20 20 20 20 20
22 27 1  2.5 22
23 41 1  2.5  6 13 13 13
24 49 1  2.5  0  3  3  3  3  3  3  3  3
25 53 0  2.5  0  0  1  1  1  1  1  1  1  1  1  1
26 50 1  2.5  0  0  2  3  4  6  6  6  6  6
27 37 1  2.5  3 15 15
28 49 1  2.5  2  3  3  3  3  4  4  4  4
29 46 1  2.5  4  6  7  9  9  9  9
30 48 0  2.5 15 26 26 26 26 26 26 26
31 54 0 10.0 12 14 15 15 15 15 15 15 15 15 15 15
32 37 1 10.0 12 16 17
33 53 1 10.0  3  6  6  6  6  6  6  6  6  6  6  6
34 45 1 10.0  4 12 15 20 20 20
35 53 0 10.0  6 10 13 13 13 15 15 15 15 15 15 20
36 49 1 10.0  0  2  2  2  2  2  2  2  2
37 39 0 10.0  7  8  8
38 27 1 10.0 17
39 49 1 10.0  0  6  9 14 14 14 14 14 14
40 43 1 10.0 14 18 20 20 20
41 28 0 10.0  8
42 34 1 10.0 11 18
43 45 1 10.0 10 12 16 16 16 16
44 37 1 10.0  0  1  1
45 43 1 10.0  9 19 19 19 19
;
```

The number of papillomas (NPap) for each animal in the study was measured repeatedly over time. One way of handling time-dependent repeated measurements in the PHREG procedure is to use programming statements to capture the appropriate covariate values of the subjects in each risk set. In this example, NPap is a time-dependent explanatory variable with values that are calculated by means of the programming statements shown in the following SAS statements:

```
proc phreg data=Tumor;
   model Time*Dead(0)=Dose NPap;
   array pp{*} P1-P14;
   array tt{*} t1-t15;
   t1 = 27;
   t2 = 34;
   t3 = 37;
   t4 = 41;
   t5 = 43;
   t6 = 45;
   t7 = 46;
   t8 = 47;
   t9 = 49;
   t10= 50;
   t11= 51;
   t12= 53;
   t13= 65;
```

```
      t14= 67;
      t15= 71;
      if Time <  tt[1]  then NPap=0;
      else if time >= tt[15] then NPap=P15;
      else do i=1 to dim(pp);
         if tt[i] <= Time < tt[i+1] then NPap= pp[i];
      end;
   run;
```

At each death time, the NPap value of each subject in the risk set is recalculated to reflect the actual number of papillomas at the given death time. For instance, subject one in the data set Tumor was in the risk sets at weeks 27 and 34; at week 27, the animal had no papilloma, while at week 34, it had five papillomas. Results of the analysis are shown in Output 49.5.1.

Output 49.5.1. Cox Regression Analysis on the Survival of Rodents

```
                            The PHREG Procedure

                             Model Information

                    Data Set                 WORK.TUMOR
                    Dependent Variable       Time
                    Censoring Variable       Dead
                    Censoring Value(s)       0
                    Ties Handling            BRESLOW


             Summary of the Number of Event and Censored Values

                                                     Percent
                   Total        Event    Censored    Censored

                      45           25          20       44.44


                             Convergence Status

               Convergence criterion (GCONV=1E-8) satisfied.


                           Model Fit Statistics

                                   Without           With
                  Criterion     Covariates     Covariates

                  -2 LOG L         166.793        143.269
                  AIC              166.793        147.269
                  SBC              166.793        149.707


                  Testing Global Null Hypothesis: BETA=0

          Test                 Chi-Square       DF     Pr > ChiSq

          Likelihood Ratio        23.5243        2         <.0001
          Score                   28.0498        2         <.0001
          Wald                    21.1646        2         <.0001


                  Analysis of Maximum Likelihood Estimates

                 Parameter     Standard                                 Hazard
Variable   DF     Estimate        Error   Chi-Square   Pr > ChiSq        Ratio

Dose        1      0.06885      0.05620       1.5010       0.2205        1.071
NPap        1      0.11714      0.02998      15.2705       <.0001        1.124
```

After the number of papillomas is adjusted for, the dose effect of the tumor-promoting agent is not statistically significant.

Another way to handle time-dependent repeated measurements in the PHREG procedure is to use the counting process style of input. Multiple records are created for each subject, one record for each distinct pattern of the time-dependent measurements. Each record contains a T1 value and a T2 value representing the time interval (T1,T2] during which the values of the explanatory variables remain unchanged. Each record also contains the censoring status at T2.

One advantage of using the counting process formulation is that you can easily obtain various residuals and influence statistics that are not available when programming statements are used to compute the values of the time-dependent variables. On the other hand, creating multiple records for the counting process formulation requires extra effort in data manipulation.

Consider a counting process style of input data set named Tumor1. It contains multiple observations for each subject in the data set Tumor. In addition to variables ID, Time, Dead, and Dose, four new variables are generated:

- T1 (left endpoint of the risk interval)
- T2 (right endpoint of the risk interval)
- NPap (number of papillomas in the time interval (T1,T2])
- Status (censoring status at T2)

For example, five observations are generated for the rodent that died at week 47 and that had no papilloma at week 27, five papillomas at week 34, six at week 37, eight at week 41, and 10 at weeks 43, 45, 46, and 47. The values of T1, T2, NPap, and Status for these five observations are (0,27,0,0), (27,34,5,0), (34,37,6,0), (37,41,8,0), and (41,47,10,1). Note that the variables ID, Time, and Dead are not needed for the estimation of the regression parameters, but they are useful for plotting the residuals.

The following SAS statements create the data set Tumor1:

```
data Tumor1(keep=ID Time Dead Dose T1 T2 NPap Status);
   array pp{*} P1-P14;
   array qq{*} P2-P15;
   array tt{1:15} _temporary_
      (27 34 37 41 43 45 46 47 49 50 51 53 65 67 71);
   set Tumor;
   T1 = 0;
   T2 = 0;
   Status = 0;
   if ( Time = tt[1] ) then do;
      T2 = tt[1];
      NPap = p1;
      Status = Dead;
      output;
   end;
   else do _i_=1 to dim(pp);
      if ( tt[_i_] = Time ) then do;
         T2= Time;
         NPap = pp[_i_] ;
         Status = Dead;
         output;
      end;
      else if (tt[_i_]  < Time ) then do;
         if (pp[_i_]  ^= qq[_i_] ) then do;
            if qq[_i_]  = . then T2= Time;
            else                 T2= tt[_i_] ;
         NPap= pp[_i_] ;
         Status= 0;
         output;
         T1 = T2;
      end;
   end;
end;
if ( Time >= tt[15] ) then do;
   T2 = Time;
   NPap = P15;
   Status = Dead;
   output;
end;
run;
```

In the following SAS statements, the counting process MODEL specification is used. The DFBETA statistics are output to a SAS data set named Out1. Note that Out1 contains multiple observations for each subject, that is, one observation for each risk interval (T1,T2].

```
proc phreg data=Tumor1;
   model (T1,T2)*Status(0)=Dose NPap;
   output out=Out1 resmart=mart dfbeta=db1-db2/order=data;
   id ID Time Dead;
run;
```

The output from PROC PHREG (not shown) is identical to Output 49.8.1 except for the "Summary of the Number of Event and Censored Values" table. The number of event observations remains unchanged between the two specifications of PROC PHREG, but the number of censored observations differs due to the splitting of each subject's data into multiple observations for the counting process style of input.

Next, the MEANS procedure sums up the component statistics for each subject and outputs the results to a SAS data set named Out2.

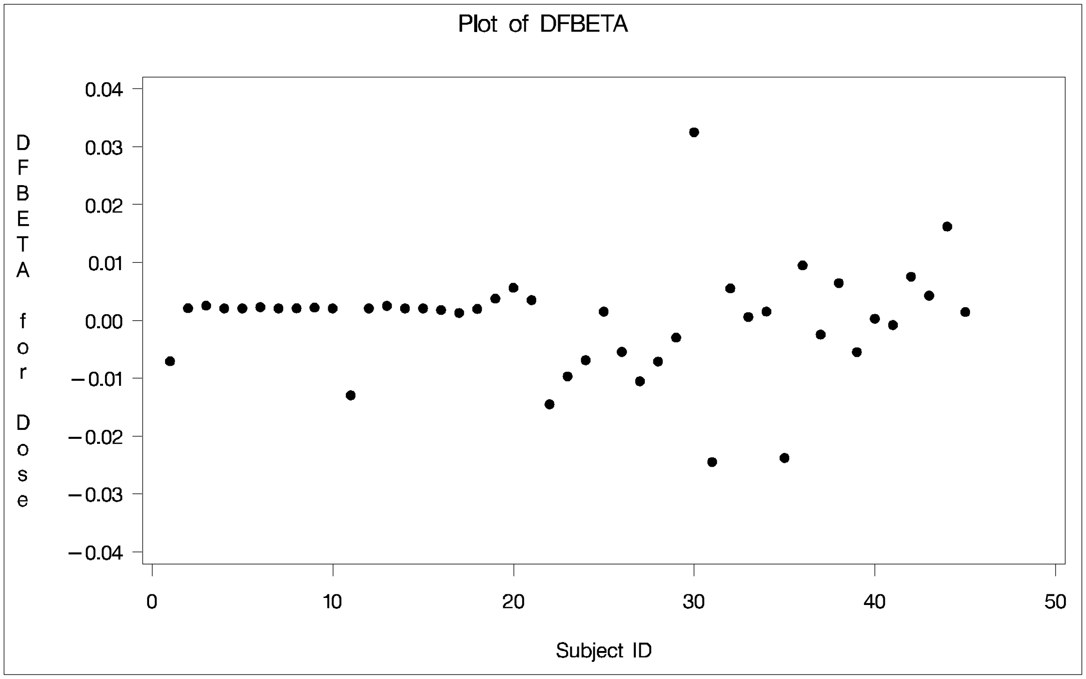
```
proc means data=Out1 noprint;
   by ID Time Dead;
   var mart db1-db2;
   output out=Out2 sum=mart db_dose db_npap;
run;
```

Finally, DFBETA statistics are plotted against subject ID for easy identification of influential points.

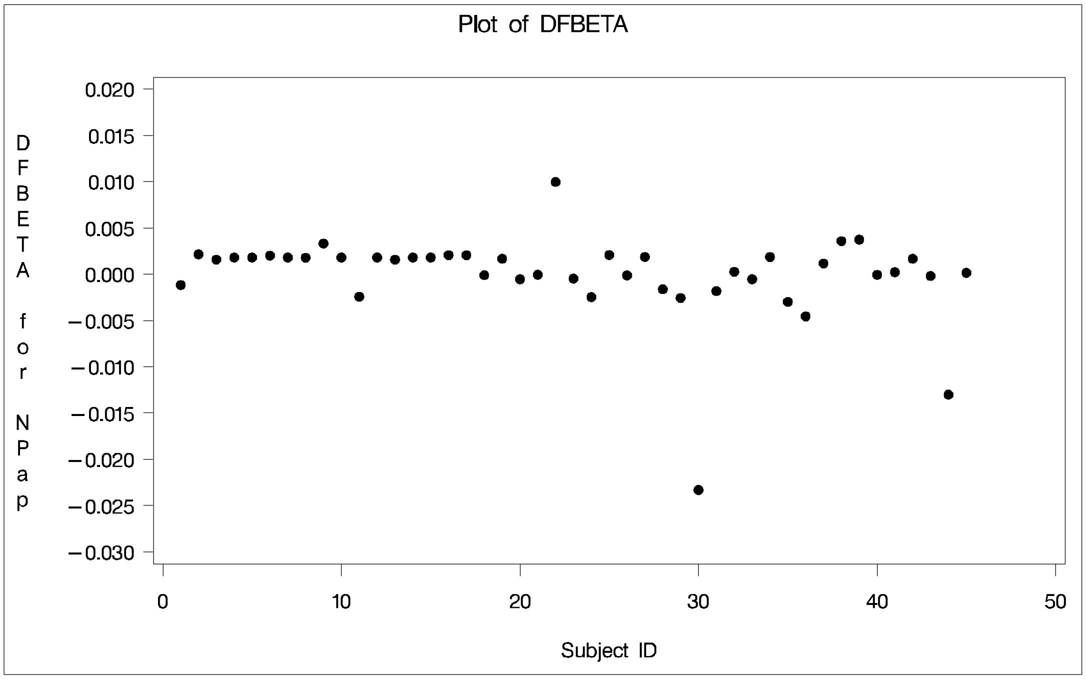
```
symbol1 v=dot h=0.8 c=blue;
axis1 label = (angle=-90 rotate=90 'DFBETA for Dose')
      minor = none
      order =  (-.04 to .04 by .01);
axis2 label = (angle=-90 rotate=90 'DFBETA for NPap')
      minor = none
      order =  (-.030 to .020 by .005);
title 'Plot of DFBETA';
proc gplot data=Out2;
   plot db_dose * ID / frame hminor=0 vaxis=axis1 cframe=ligr;
   plot db_npap * ID / frame hminor=0 vaxis=axis2 cframe=ligr;
run;
```

The plots of the DFBETA statistics are shown in Output 49.5.2 and Output 49.5.3. Subject 30 appears to have a large influence on both the Dose and NPap coefficients. Subjects 31 and 35 have considerable influences on the DOSE coefficient, while subjects 22 and 44 have rather large influences on the NPap coefficient.

Output 49.5.2. Plot of DFBETA Statistic for DOSE versus Subject Number



Output 49.5.3. Plot of DFBETA Statistic for NPAP versus Subject Number



Example 49.6. Survivor Function Estimates for Specific Covariate Values

You may want to use your regression analysis results to generate predicted survival curves for subjects not in the study. This example illustrates how to use the BASELINE statement to obtain the survivor function for a new set of explanatory variable values. The various sets of explanatory variable values must be contained in a SAS data set.

In previous examples, LogBUN and HGB were identified as the most important prognostic factors for the myeloma data. Suppose you are interested in obtaining the survivor function estimates for the following two realizations of LogBUN and HGB, which are saved in a SAS data set called Inrisks.

```
data Inrisks;
   input LogBUN HGB;
   datalines;
1.00 10.0
1.80 12.0
;
```

In the BASELINE statement, you specify the name of the data set (COVARIATE=Inrisk) that contains the various sets of explanatory variable values and the name of the output SAS data set (OUT=Pred1) that contains the survivor function estimates. The option SURVIVAL=S puts the variable S containing the survivor function estimates in the output data set Pred1. Similarly, the options LOWER=S_lower and UPPER=S_upper put the variables S_lower and S_upper in Pred1; these variables contain, respectively, the lower and upper 95% confidence limits for the survival. The NOPRINT option in the PROC PHREG statement suppresses the displayed output (the analysis results are shown in Example 49.1). The PRINT procedure displays the observations in the data set Pred1.

```
proc phreg data=Myeloma noprint;
   model Time*VStatus(0)=LogBUN HGB;
   baseline covariates=Inrisks out=Pred1 survival=S
            lower=S_lower upper=S_upper;
run;
proc print data=Pred1;
run;
```

Output 49.6.1. Survivor Function Estimates for LogBUN=1.0 and HGB=10.0

```
Obs    LogBUN      HGB       Time       S       S_lower    S_upper

  1    1.0000    10.0000     0.00    1.00000     .          .
  2    1.0000    10.0000     1.25    0.98622    0.96600    1.00000
  3    1.0000    10.0000     2.00    0.96438    0.92775    1.00000
  4    1.0000    10.0000     3.00    0.95687    0.91513    1.00000
  5    1.0000    10.0000     5.00    0.93966    0.88745    0.99494
  6    1.0000    10.0000     6.00    0.90211    0.83101    0.97929
  7    1.0000    10.0000     7.00    0.87192    0.78793    0.96487
  8    1.0000    10.0000     9.00    0.86073    0.77215    0.95947
  9    1.0000    10.0000    11.00    0.80252    0.69458    0.92725
 10    1.0000    10.0000    13.00    0.78969    0.67751    0.92044
 11    1.0000    10.0000    14.00    0.77554    0.65896    0.91274
 12    1.0000    10.0000    15.00    0.76116    0.64048    0.90458
 13    1.0000    10.0000    16.00    0.73142    0.60343    0.88654
 14    1.0000    10.0000    17.00    0.69988    0.56494    0.86706
 15    1.0000    10.0000    18.00    0.68345    0.54525    0.85667
 16    1.0000    10.0000    19.00    0.64951    0.50561    0.83438
 17    1.0000    10.0000    24.00    0.63105    0.48401    0.82278
 18    1.0000    10.0000    25.00    0.61267    0.46287    0.81096
 19    1.0000    10.0000    26.00    0.59428    0.44209    0.79887
 20    1.0000    10.0000    32.00    0.57437    0.41972    0.78601
 21    1.0000    10.0000    35.00    0.55400    0.39725    0.77258
 22    1.0000    10.0000    37.00    0.53276    0.37421    0.75849
 23    1.0000    10.0000    41.00    0.48783    0.32796    0.72564
 24    1.0000    10.0000    51.00    0.45964    0.29978    0.70476
 25    1.0000    10.0000    52.00    0.42933    0.27013    0.68234
 26    1.0000    10.0000    54.00    0.39588    0.23828    0.65773
 27    1.0000    10.0000    58.00    0.35744    0.20219    0.63191
 28    1.0000    10.0000    66.00    0.31314    0.16511    0.59386
 29    1.0000    10.0000    67.00    0.26060    0.12215    0.55597
 30    1.0000    10.0000    88.00    0.19554    0.07520    0.50849
 31    1.0000    10.0000    89.00    0.12708    0.03552    0.45460
 32    1.0000    10.0000    92.00    0.00000     .          .
```

The first 32 observations of the data set Pred1 are shown in Output 49.6.1. They represent the survivor function for the realization LogBUN=1.00 and HGB=10.0. The first observation has survival time 0 and survivor function estimate 1.0. Each of the remaining 31 observations represents each unique event time in the input data set Myeloma. These observations are presented in ascending order of the event times.

Likewise, the next 32 observations of the data set Pred1 (starting from the 33rd observation) represent the survivor function for the realization LogBUN=1.80 and HGB=12.0.

By default, the procedure also outputs the set of survivor function estimates for LogBUN=1.3929 and HGB=10.2015, which are the sample means of LogBUN and HGB for the input data in Myeloma. (Note that in a stratified analysis, the sample means are calculated within each stratum.) The estimated survivor function estimates for these sample means are the last 32 observations in the data set Pred1. You can suppress this set of survival estimates by using the NOMEAN option in the BASELINE statement.

```
proc phreg data=Myeloma noprint;
   model Time*VStatus(0)=LogBUN HGB;
   baseline covariates=Inrisks out=Pred2 survival=S
            lower=S_lower upper=S_upper / nomean;
run;
```

The data set Pred2 consists of the first 64 observations of Pred1. If you are interested only in the survivor function estimates for the sample means of the explanatory variables, you can omit the COVARIATES= option in the BASELINE statement.

```
proc phreg data=Myeloma noprint;
   model Time*VStatus(0)=LogBUN HGB;
   baseline out=Pred3 survival=S lower=S_lower upper=S_upper;
run;
```

The data set Pred3 contains the last 32 observations of Pred1.

The following SAS statements are used to plot the survival curves in Pred1. For convenience, the variable Pattern is added to the data set Pred1 to identify the various patterns of explanatory variables.

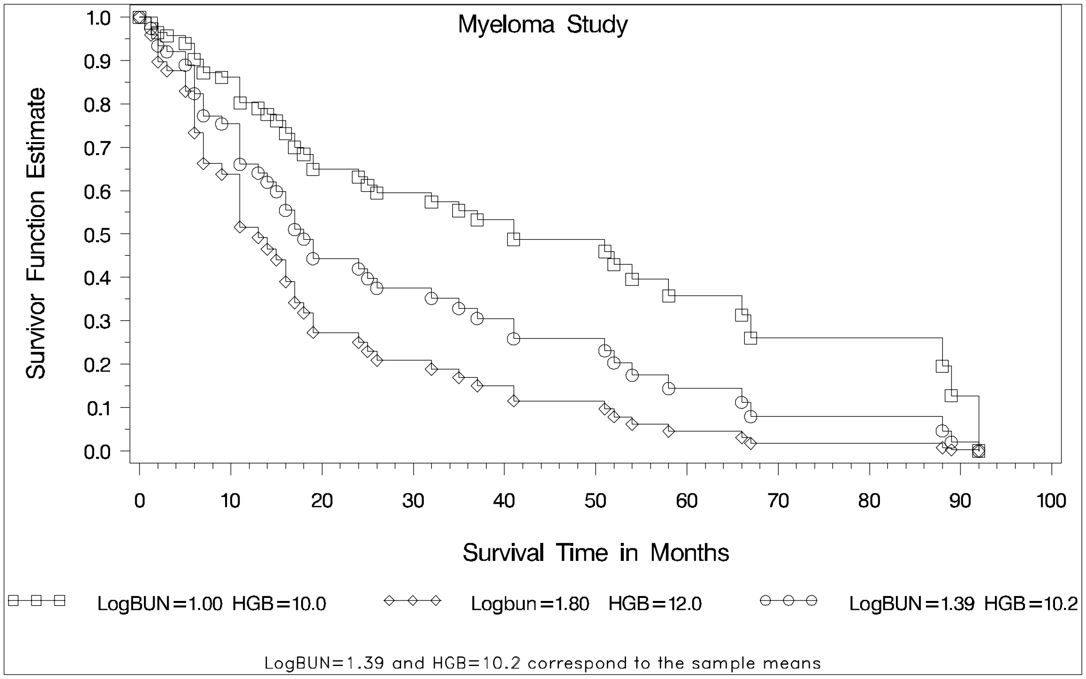
```
data Pred1;
   set Pred1;
   if      LogBUN= 1.0 and HGB=10.0 then Pattern=1;
   else if LogBUN= 1.8 and HGB=12.0 then Pattern=2;
   else                                  Pattern=3;

legend1 label=none shape=symbol(3, .8)
   value=(f=swiss h=.8 'LogBUN=1.00 HGB=10.0'
          'LogBUN=1.80  HGB=12.0' 'LogBUN=1.39 HGB=10.2');
axis1 label=(h=1 f=swiss a=90) minor=(n=1);
axis2 label=(h=1 f=swiss 'Survival Time in Months') minor=(n=4);
```

```
proc gplot data=Pred1;
   plot S*Time=Pattern / legend=legend1 vaxis=axis1
                         haxis=axis2    cframe=ligr;
   symbol1 interpol=stepLJ h=1 v=square  c=blue;
   symbol2 interpol=stepLJ h=1 v=diamond c=yellow;
   symbol3 interpol=stepLJ h=1 v=circle  c=red;
   note f=swiss h=1.5 j=c 'Myeloma Study';
   footnote h=.8 f=duplex
      'LogBUN=1.39 and HGB=10.2 correspond to the sample means';
run;
```

The survivor function estimates for these three patterns of explanatory variables are displayed in Output 49.6.2. Note that these survivor functions are portrayed as right-continuous functions.

Output 49.6.2. Survival Curves for Specific Covariate Patterns



Example 49.7. Analysis of Residuals

Residuals are used to investigate the lack of fit of a model to a given subject. You can obtain martingale and deviance residuals for the Cox proportional hazards regression analysis by requesting that they be included in the OUTPUT data set. You can plot these statistics and look for outliers.

Consider the stepwise regression analysis performed in Example 49.1. The final model included variables LogBUN and HGB. You can generate residual statistics for this analysis by refitting the model containing those variables and including an OUTPUT statement. The keywords XBETA, RESMART, and RESDEV identify new variables that contain the linear predictor scores $\mathbf{z}_j'\widehat{\boldsymbol{\beta}}$, martingale residuals, and deviance residuals. These variables are xb, mart, and dev, respectively.

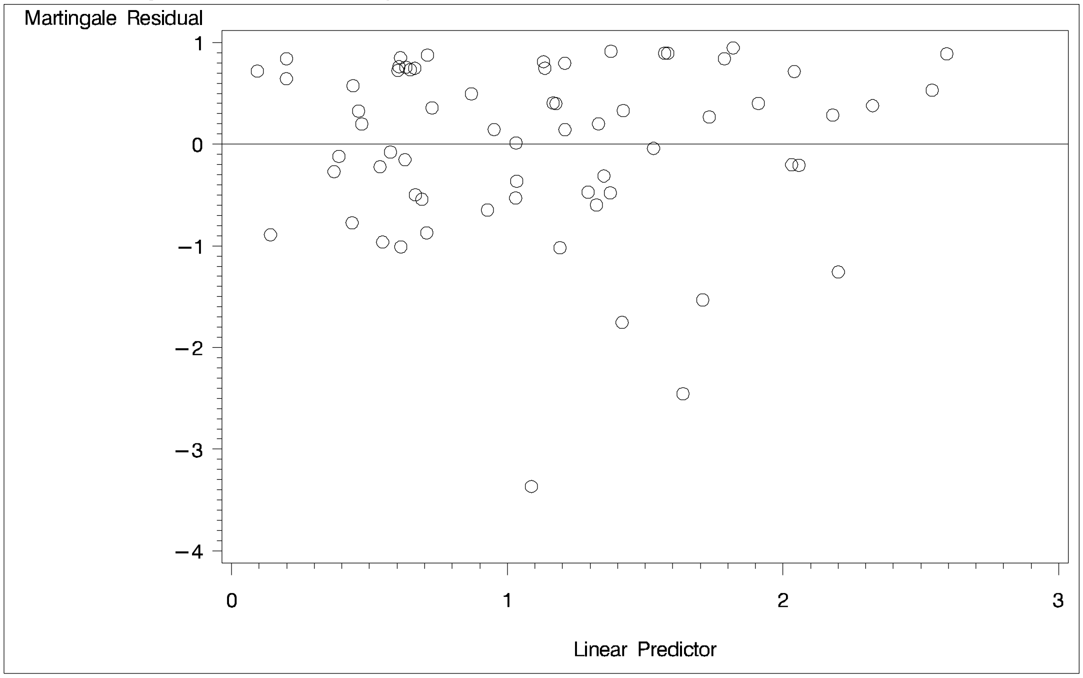
```
proc phreg data=Myeloma noprint;
   model Time*Vstatus(0)=LogBUN HGB;
   output out=Outp xbeta=xb resmart=mart resdev=dev;
run;
```

The following statements plot the residuals against the linear predictor scores:

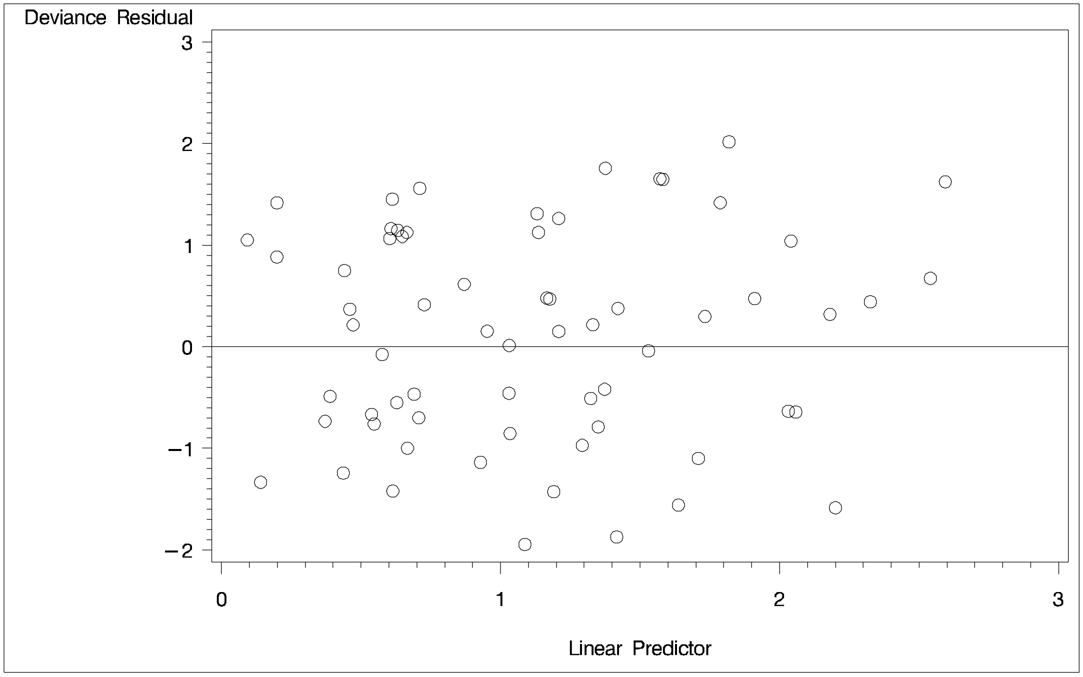
```
proc gplot data=Outp;
   plot (mart dev)*xb / vref=0 cframe=ligr;
   symbol1 value=circle c=blue;
run;
```

The resulting plots are shown in Output 49.7.1 and Output 49.7.2. The martingale residuals are skewed because of the single event setting of the Cox model. The martingale residual plot shows an isolation point (with linear predictor score 1.09 and martingale residual -3.37), but this observation is no longer distinguishable in the deviance residual plot. In conclusion, there is no indication of a lack of fit of the model to individual observations.

Output 49.7.1. Martingale Residual Plot



Output 49.7.2. Deviance Residual Plot



Example 49.8. Multiple Failure Outcomes

For survival data with multiple failure outcomes for each subject, the failures may be repetitions of the same kind of event or they may be events of different nature.

The Andersen-Gill (AG) model and the Prentice, Williams, and Peterson (1981) models, also referred to as the PWP models, can be used in the analysis of repeated failure outcomes of the same kind, while the marginal analysis approach of Wei, Lin, and Weissfeld (1989), also referred to as the WLW analysis, can be applied to both multiple events of the same types and multiple events of different types. The bladder cancer data listed in Wei, Lin, and Weissfeld (1989) are used to illustrate these methods of analyses.

The data consist of 86 patients with superficial bladder tumors, which were removed when they entered the study. Of these patients, 48 were randomized into the placebo group, and 38 were randomized into the thiotepa group. Many patients had multiple recurrences of tumors during the study, and new tumors were removed at each visit. The data set contains the first four recurrences of the tumor for each patient, and each recurrence time was measured from the patient's entry time into the study.

The input data consist of the following eight variables:

- Trt (treatment group, where 1=placebo and 2=thiotepa)
- Time (follow-up time)
- Number (number of initial tumors)
- Size (initial tumor size)
- T1, T2, T3, and T4 (times of the four possible recurrences of the bladder tumor. A patient with only two recurrences has missing values in T3 and T4.)

In the data set Bladder, four observations are created for each patient, each corresponding to one of the four tumor recurrences. In addition to values of Trt, Number, and Size for the patient, each observation contains the following variables:

- ID (patient's identification, which is the sequence number of the input data)
- Visit (event number, where 1=first recurrence, 2= second recurrence, and so on)
- TStart (time of the $(k-1)$ recurrence if Visit=k, or the entry time 0 if VISIT=1)
- TStop (time of the kth recurrence if Visit=k)
- Status (event status, where 1=recurrence and 0=censored)

For instance, a patient with only one recurrence time at month 6, who was followed until month 10, will have values for Visit, TStart, TStop, and Status of (1,0,6,1), (2,6,10,0), (3,10,10,0), and (4,10,10,0). If the follow-up time of a patient is beyond the time of the fourth tumor recurrence, an extra observation is created to represent this last follow-up period. For instance, a patient with recurrences at months 2, 15, 34, and 50, who was followed until month 61, will have five observations with values for Visit, TStart, TStop, and Status of (1,0,2,1), (2,2,15,1), (3,15,34,1), (4,34,50,1), and (5,50,61,0). In the former situation, the last two observations are redundant for the AG model, but they are important for the WLW analysis. In the latter situation, the fifth observation is not needed for the WLW model, but it is indispensable to the AG analysis.

The following SAS statements create the data set Bladder:

```
data Bladder;
   keep ID TStart TStop Status Trt Number Size Visit;
   retain ID TStart 0;
   array tt T1-T4;
   infile datalines missover;
   input Trt Time Number Size T1-T4;
   ID + 1;
   TStart=0;
   do over tt;
      Visit=_i_;
      if tt = . then do;
         TStop=Time;
         Status=0;
      end;
      else do;
         TStop=tt;
         Status=1;
      end;
      output;
      TStart=TStop;
   end;
   if (TStart < Time) then do;
      TStop= Time;
      Status=0;
      Visit=5;
      output;
   end;
   datalines;
1       0       1     1
1       1       1     3
1       4       2     1
1       7       1     1
1       10      5     1
1       10      4     1   6
1       14      1     1
1       18      1     1
1       18      1     3   5
1       18      1     1   12  16
1       23      3     3
1       23      1     3   10  15
1       23      1     1   3   16  23
1       23      3     1   3   9   21
1       24      2     3   7   10  16  24
1       25      1     1   3   15  25
1       26      1     2
1       26      8     1   1
1       26      1     4   2   26
1       28      1     2   25
1       29      1     4
1       29      1     2
```

```
1       29      4     1
1       30      1     6   28  30
1       30      1     5   2   17  22
1       30      2     1   3   6   8   12
1       31      1     3   12  15  24
1       32      1     2
1       34      2     1
1       36      2     1
1       36      3     1   29
1       37      1     2
1       40      4     1   9   17  22  24
1       40      5     1   16  19  23  29
1       41      1     2
1       43      1     1   3
1       43      2     6   6
1       44      2     1   3   6   9
1       45      1     1   9   11  20  26
1       48      1     1   18
1       49      1     3
1       51      3     1   35
1       53      1     7   17
1       53      3     1   3   15  46  51
1       59      1     1
1       61      3     2   2   15  24  30
1       64      1     3   5   14  19  27
1       64      2     3   2   8   12  13
2       1       1     3
2       1       1     1
2       5       8     1   5
2       9       1     2
2       10      1     1
2       13      1     1
2       14      2     6   3
2       17      5     3   1   3   5   7
2       18      5     1
2       18      1     3   17
2       19      5     1   2
2       21      1     1   17  19
2       22      1     1
2       25      1     3
2       25      1     5
2       25      1     1
2       26      1     1   6   12  13
2       27      1     1   6
2       29      2     1   2
2       36      8     3   26  35
2       38      1     1
2       39      1     1   22  23  27  32
2       39      6     1   4   16  23  27
2       40      3     1   24  26  29  40
2       41      3     2
2       41      1     1
2       43      1     1   1   27
2       44      1     1
2       44      6     1   2   20  23  27
2       45      1     2
2       46      1     4   2
2       46      1     4
2       49      3     3
```

```
2       50      1     1
2       50      4     1   4   24  47
2       54      3     4
2       54      2     1   38
2       59      1     3
;
```

The counting process MODEL specification is used to carry out the analysis of the AG model. Note that some of the observations in the data set Bladder have a degenerated interval of risk. The presence of these observations does not affect the results of the analysis since none of these observations are included in any of the risk sets. However, the procedure will run more efficiently without these observations; consequently, in the following SAS statements, the WHERE clause is used to eliminate these redundant observations.

```
*title 'Andersen-Gill Multiplicative Hazards Model';
proc phreg data=Bladder;
   model (TStart, TStop) * Status(0) = Trt Number Size;
   where TStart < TStop;
run;
```

Results of fitting the AG model are shown in Output 49.8.1.

Output 49.8.1. Fitting Andersen-Gill Multiplicative Hazards Model

```
                            The PHREG Procedure

                             Model Information

                   Data Set                 WORK.BLADDER
                   Dependent Variable       TStart
                   Dependent Variable       TStop
                   Censoring Variable       Status
                   Censoring Value(s)       0
                   Ties Handling            BRESLOW


             Summary of the Number of Event and Censored Values

                                                     Percent
                   Total       Event    Censored    Censored

                     190         112          78       41.05


                             Convergence Status

               Convergence criterion (GCONV=1E-8) satisfied.


                           Model Fit Statistics

                                   Without           With
                  Criterion     Covariates     Covariates

                  -2 LOG L         934.210        920.159
                  AIC              934.210        926.159
                  SBC              934.210        934.315


                  Testing Global Null Hypothesis: BETA=0

          Test                 Chi-Square       DF     Pr > ChiSq

          Likelihood Ratio        14.0509        3         0.0028
          Score                   15.4173        3         0.0015
          Wald                    15.1736        3         0.0017


                  Analysis of Maximum Likelihood Estimates

                 Parameter     Standard                                Hazard
Variable   DF     Estimate        Error   Chi-Square   Pr > ChiSq       Ratio

Trt         1     -0.40710      0.20007       4.1402       0.0419       0.666
Number      1      0.16065      0.04801      11.1980       0.0008       1.174
Size        1     -0.04009      0.07026       0.3256       0.5683       0.961
```

The WLW analysis regards the recurrence times as multivariate failure times, and it models the marginal distribution of each component time with the Cox proportional hazards model. No specific correlation structure is imposed on the multiple failure times. For the kth marginal model, let $\boldsymbol{\beta}_k$ denote the row vector of regression parameters, let $\hat{\boldsymbol{\beta}}_k$ denote the maximum likelihood estimate of $\boldsymbol{\beta}_k$, let $\hat{\mathbf{A}}_k$ denote the covariance matrix obtained by inverting the information matrix, and let $\mathbf{R}_i$ denote the matrix of score residuals. Wei, Lin, and Weissfeld (1989) showed that the joint distribution of $(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_4)'$ can be approximated by a multivariate normal distribution with mean vector $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_4)'$ and robust covariance matrix

$$\begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \mathbf{V}_{13} & \mathbf{V}_{14} \\ \mathbf{V}_{21} & \mathbf{V}_{22} & \mathbf{V}_{23} & \mathbf{V}_{24} \\ \mathbf{V}_{31} & \mathbf{V}_{32} & \mathbf{V}_{33} & \mathbf{V}_{34} \\ \mathbf{V}_{41} & \mathbf{V}_{42} & \mathbf{V}_{43} & \mathbf{V}_{44} \end{pmatrix}$$

with the submatrix $\mathbf{V}_{ij}$ given by

$$\mathbf{V}_{ij} = \hat{\mathbf{A}}_i(\mathbf{R}_i'\mathbf{R}_j)\hat{\mathbf{A}}_j$$

The PHREG procedure computes the DFBETA statistics, which are precisely the products $\mathbf{R}_k\hat{\mathbf{A}}_k$. By outputting the DFBETA statistics into an OUTPUT data set, you can subsequently use SAS/IML software to compute the robust covariance matrix in a straightforward manner.

In this example, there are four marginal proportional hazards models, one for each recurrence time. Instead of fitting one model at a time, you can fit all four marginal models in one analysis by using the STRATA statement. A new input data set (named Bladder2) is created from the data set Bladder by eliminating observations that have a Visit value of 5. These observations contain the follow-up information beyond the fourth recurrence time and are irrelevant in the fitting of the four marginal models. In addition, four treatment variables, Trt1, Trt2, Trt3, and Trt4, are created from variables Trt and Visit, representing the treatment group for each of the four values of Visit. For instance, Trt1=Trt when the Visit value is 1, and Trt1=0 when the Visit value is not 1. Likewise, variables Number1, Number2, Number3, and Number4 are created from the variables Number and Visit; variables Size1, Size2, Size3, and Size4 are created from the variables Size and Visit.

The following SAS statements create the data set Bladder2:

```
data Bladder2;
   set Bladder;
   if Visit < 5;
   Trt1= Trt * (Visit=1);
   Trt2= Trt * (Visit=2);
   Trt3= Trt * (Visit=3);
   Trt4= Trt * (Visit=4);
   Number1= Number * (Visit=1);
   Number2= Number * (Visit=2);
   Number3= Number * (Visit=3);
   Number4= Number * (Visit=4);
   Size1= Size * (Visit=1);
```

```
      Size2= Size * (Visit=2);
      Size3= Size * (Visit=3);
      Size4= Size * (Visit=4);
   run;
```

The following SAS statements fit the marginal models. The parameter estimates are output to a data set named Est1; the DFBETA statistics for the treatment variables are output to a data set named Out1.

```
   *title ’Fitting Marginal Proportional Hazards Models’;
   proc phreg data=Bladder2 outest=Est1;
      model TStop*Status(0)=Trt1-Trt4 Number1-Number4 Size1-Size4;
      output out=Out1 dfbeta=dt1-dt4 / order=data;
      strata Visit;
      id ID;
   run;
```

The output of this analysis is shown in Output 49.8.2

Output 49.8.2. Fitting Marginal Proportional Hazards Models

```
                            The PHREG Procedure

                             Model Information

                   Data Set                 WORK.BLADDER2
                   Dependent Variable       TStop
                   Censoring Variable       Status
                   Censoring Value(s)       0
                   Ties Handling            BRESLOW


             Summary of the Number of Event and Censored Values

                                                                Percent
    Stratum    Visit          Total       Event    Censored    Censored

          1    1                 86          47          39       45.35
          2    2                 86          29          57       66.28
          3    3                 86          22          64       74.42
          4    4                 86          14          72       83.72
    -------------------------------------------------------------------
      Total                     344         112         232       67.44


                             Convergence Status

               Convergence criterion (GCONV=1E-8) satisfied.


                           Model Fit Statistics

                                   Without           With
                  Criterion     Covariates     Covariates

                  -2 LOG L         880.828        851.435
                  AIC              880.828        875.435
                  SBC              880.828        908.057


                  Testing Global Null Hypothesis: BETA=0

          Test                 Chi-Square       DF     Pr > ChiSq

          Likelihood Ratio        29.3932       12         0.0034
          Score                   33.0747       12         0.0009
          Wald                    31.0544       12         0.0019


                  Analysis of Maximum Likelihood Estimates

                 Parameter     Standard                               Hazard
Variable   DF     Estimate        Error   Chi-Square   Pr > ChiSq      Ratio

Trt1        1     -0.51762      0.31576       2.6873       0.1012      0.596
Trt2        1     -0.61944      0.39318       2.4821       0.1151      0.538
Trt3        1     -0.69988      0.45994       2.3155       0.1281      0.497
Trt4        1     -0.65079      0.57744       1.2702       0.2597      0.522
Number1     1      0.23599      0.07608       9.6231       0.0019      1.266
Number2     1      0.13756      0.09190       2.2403       0.1345      1.147
Number3     1      0.16984      0.10521       2.6061       0.1065      1.185
Number4     1      0.32880      0.12528       6.8880       0.0087      1.389
Size1       1      0.06789      0.10125       0.4496       0.5025      1.070
Size2       1     -0.07612      0.13406       0.3224       0.5702      0.927
Size3       1     -0.21131      0.18240       1.3421       0.2467      0.810
Size4       1     -0.20317      0.23018       0.7791       0.3774      0.816
```

The following SAS statements calculate the robust covariance matrix for the treatment coefficients. The MEANS procedure sums up the DFBETA statistics for each subject and outputs the results to a SAS data set named Out2. The IML procedure then reads the DFBETA statistics from the data set Out2 and computes the robust variance, which is output to a SAS data set called RCov.

```
proc means data=Out1 noprint;
   by ID;
   var dt1-dt4;
   output out=Out2 sum=dt1-dt4;

proc iml;
   use Out2;
   read all var{dt1 dt2 dt3 dt4} into x;
   v=x‘ * x;
   reset noname;
   vname={"Trt1","Trt2","Trt3","Trt4"};
   print,"Estimated Covariance Matrix",, v[colname=vname
     rowname=vname format=10.5];
   create RCov from v[colname=vname rowname=vname];
   append from v[rowname=vname];
```

The estimated robust covariance matrix is displayed in Output 49.8.3.

Output 49.8.3. Robust Covariance Matrix for the Treatment Coefficients

```
                     Estimated Covariance Matrix


                      Trt1       Trt2       Trt3       Trt4

             Trt1    0.09456    0.06018    0.05677    0.04378
             Trt2    0.06018    0.13243    0.13012    0.11604
             Trt3    0.05677    0.13012    0.17236    0.15909
             Trt4    0.04378    0.11604    0.15909    0.23981
```

The approximate multivariate normal distribution of the parameter estimators provides a basis for simultaneous inferences about the parameters. For example, you can test jointly the null hypothesis of no treatment effect for each tumor recurrence, or you can estimate the coefficient for the common treatment effect. A detailed IML program for the analysis is given by the following SAS statements (results not shown):

```
proc iml;
   use Est1;
   read all var{Trt1 Trt2 Trt3 Trt4} into Trt;
   b= Trt‘;
   use Out2;
   read all var{dt1 dt2 dt3 dt4} into x;
   v=x‘ * x;
   nparm= nrow(b);
   se=sqrt(vecdiag(v));
   reset noname;
   stitle={"Estimate", "   Std Error"};
   vname={"Trt1","Trt2","Trt3","Trt4"};
   tmpprt= b || se;
   print,tmpprt[colname=stitle rowname=vname format=10.5];
   print,"Estimated Covariance Matrix",,
          v[colname=vname rowname=vname format=10.5];

   /* H0: beta11=beta12=beta13=beta14=0 */
   chisq= b‘ * inv(v) * b;
   df= nrow(b);
   p= 1-probchi(chisq,df);
   print ,,"Testing H0: no treatment effects", ,
         "Wald Chi-Square = " chisq, "DF = " df,
         "p-value = "p[format=5.4],;

   /* Assume beta11=beta12=beta13=beta14 and
      estimate the common value */
   c= {1 0 0 0, 0 1 0 0, 0 0 1 0, 0 0 0 1};
   cb= c * b;
   si= c * v * t(c);
   e= j(4,1,1);
   isi=inv(si);
   h= inv(e‘ * isi * e) * isi * e;
   b1= t(h) * cb;
   se= sqrt(t(h) * si * h);
   zscore= b1 / se;
   p= 1- probchi( zscore # zscore, 1);
   print ,"Estimation of the Common Parameter for Treatment",,
         "Optimal Weights = "h,
         "Estimate = " b1,
         "Standard Error = " se,
         "z-score =" zscore,
         "2-sided p-value = " p[format=5.4];
quit;
```

The PHREG procedure can also be used to fit the PWP model. In the PWP model, the risk set for the $(k+1)$ recurrence is restricted to those patients who have experienced the first k recurrences. For example, a patient who experienced only one recurrence is an event observation for the first recurrence; this patient is a censored observation for the second recurrence and should not be included in the risk set for the third or fourth recurrence. The following DATA step eliminates those observations that should not be in the risk sets, forming a new input data set (named Bladder3) for fitting the PWP models. The gap times between successive recurrences are also calculated.

The following SAS statements create the data set Bladder3:

```
data Bladder3(drop=lstatus);
   retain lstatus;
   set Bladder2;
   by ID;
   if first.ID then lstatus=1;
   if (Status=0 and lstatus=0) then delete;
   lstatus=Status;
```

```
      GapTime=TStop-TStart;
   run;
```

The following statements fit the PWP total time model with noncommon effects:

```
   *title2 'PWP Total Time Model with Noncommon Effects';
   proc phreg data=Bladder3;
      model TStop*Status(0)=Trt1-Trt4 Number1-Number4
                                      Size1-Size4;
      strata Visit;
   run;
```

The following statements fit the PWP gap time model with noncommon effects:

```
   *title2 'PWP Gap Time Model with Noncommon Effects';
   proc phreg data=Bladder3;
      model GapTime*Status(0)=Trt1-Trt4 Number1-Number4
                               Size1-Size4;
      strata Visit;
   run;
```

Results of these two analyses are shown in Output 49.8.4 and Output 49.8.5, respectively.

Output 49.8.4. Fitting PWP Total Time Model with Noncommon Effects

```
                              The PHREG Procedure

                               Model Information

                    Data Set                 WORK.BLADDER3
                    Dependent Variable       TStop
                    Censoring Variable       Status
                    Censoring Value(s)       0
                    Ties Handling            BRESLOW


              Summary of the Number of Event and Censored Values

                                                                Percent
      Stratum    Visit          Total       Event    Censored    Censored

            1    1                 86          47          39       45.35
            2    2                 47          29          18       38.30
            3    3                 29          22           7       24.14
            4    4                 22          14           8       36.36
      -------------------------------------------------------------------
        Total                     184         112          72       39.13


                               Convergence Status

                Convergence criterion (GCONV=1E-8) satisfied.
```

```
                              The PHREG Procedure

                             Model Fit Statistics

                                      Without           With
                  Criterion        Covariates     Covariates

                  -2 LOG L            743.098        725.677
                  AIC                 743.098        749.677
                  SBC                 743.098        782.299


                  Testing Global Null Hypothesis: BETA=0

          Test                  Chi-Square       DF     Pr > ChiSq

          Likelihood Ratio         17.4211       12         0.1344
          Score                    18.5546       12         0.0999
          Wald                     17.7388       12         0.1239


                  Analysis of Maximum Likelihood Estimates

                 Parameter     Standard                                Hazard
Variable   DF     Estimate        Error   Chi-Square     Pr > ChiSq     Ratio

Trt1        1     -0.51757      0.31576       2.6868         0.1012     0.596
Trt2        1     -0.42584      0.40227       1.1206         0.2898     0.653
Trt3        1     -0.89894      0.53956       2.7757         0.0957     0.407
Trt4        1     -0.23739      0.68274       0.1209         0.7281     0.789
Number1     1      0.23605      0.07607       9.6287         0.0019     1.266
Number2     1      0.00117      0.09372       0.0002         0.9900     1.001
Number3     1      0.01468      0.13253       0.0123         0.9118     1.015
Number4     1      0.29306      0.22067       1.7637         0.1842     1.341
Size1       1      0.06790      0.10125       0.4498         0.5024     1.070
Size2       1     -0.12515      0.11708       1.1425         0.2851     0.882
Size3       1     -0.21520      0.17801       1.4615         0.2267     0.806
Size4       1      0.25135      0.29077       0.7472         0.3874     1.286
```

Output 49.8.5. Fitting PWP Gap Time Model with Noncommon Effects

```
                        The PHREG Procedure

                         Model Information

               Data Set                 WORK.BLADDER3
               Dependent Variable       GapTime
               Censoring Variable       Status
               Censoring Value(s)       0
               Ties Handling            BRESLOW


         Summary of the Number of Event and Censored Values

                                                            Percent
Stratum    Visit          Total       Event    Censored    Censored

      1    1                 86          47          39       45.35
      2    2                 47          29          18       38.30
      3    3                 29          22           7       24.14
      4    4                 22          14           8       36.36
-------------------------------------------------------------------
  Total                     184         112          72       39.13


                         Convergence Status

           Convergence criterion (GCONV=1E-8) satisfied.


                       Model Fit Statistics

                               Without           With
              Criterion     Covariates     Covariates

              -2 LOG L         735.076        717.268
              AIC              735.076        741.268
              SBC              735.076        773.890
```

```
                              The PHREG Procedure

                    Testing Global Null Hypothesis: BETA=0

          Test                 Chi-Square       DF     Pr > ChiSq

          Likelihood Ratio        17.8089       12         0.1216
          Score                   19.6097       12         0.0748
          Wald                    18.4759       12         0.1020


                    Analysis of Maximum Likelihood Estimates

                   Parameter     Standard                                 Hazard
Variable   DF      Estimate        Error   Chi-Square   Pr > ChiSq        Ratio

Trt1        1      -0.51757      0.31576       2.6868       0.1012        0.596
Trt2        1      -0.25911      0.40511       0.4091       0.5224        0.772
Trt3        1       0.22105      0.54909       0.1621       0.6873        1.247
Trt4        1      -0.19498      0.64184       0.0923       0.7613        0.823
Number1     1       0.23605      0.07607       9.6287       0.0019        1.266
Number2     1      -0.00571      0.09667       0.0035       0.9529        0.994
Number3     1       0.12935      0.15970       0.6561       0.4180        1.138
Number4     1       0.42079      0.19816       4.5091       0.0337        1.523
Size1       1       0.06790      0.10125       0.4498       0.5024        1.070
Size2       1      -0.11636      0.11924       0.9524       0.3291        0.890
Size3       1       0.24995      0.23113       1.1695       0.2795        1.284
Size4       1       0.03557      0.29043       0.0150       0.9025        1.036
```

The following statements fit the PWP total time model with common effects:

```
*title2 'PWP Total Time Model with Common Effects';
proc phreg data=Bladder3;
   model TStop*Status(0)=Trt Number Size;
   strata Visit;
run;
```

The following statements fit the PWP gap time model with common effects:

```
*title2 'PWP Gap Time Model with Common Effects';
proc phreg data=Bladder3;
   model GapTime*Status(0)=Trt Number Size;;
   strata Visit;
run;
```

Results of these two analyses are not shown in this document.

References

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

Andersen, P.K., Borgan, é., Gill, R.D., and Keiding, N. (1992), *Statistical Models Based on Counting Processes*, New York: Springer-Verlag.

Andersen, P.K. and Gill, R.D. (1982), "Cox's Regression Model Counting Process: a Large Sample Study," *Annals of Statistics*, 10, 1100–1120.

Barlow, W.E. and Prentice, R.L. (1988), "Residuals for Relative Risk Regression," *Biometrika*, 75, 65–74.

Breslow, N.E. (1974), "Covariance Analysis of Censored Survival Data," *Biometrics*, 30, 89–99.

Cain, K.C. and Lange, N.T. (1984), "Approximate Case Influence for the Proportional Hazards Regression Model with Censored Data," *Biometrics*, 40, 493–499.

Collett, D. (1994), *Modelling Survival Data In Medical Research*, London: Chapman and Hall.

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. (1975), "Partial Likelihood," *Biometrika*, 62, 269–276.

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

Crowley, J. and Hu, M. (1977), "Covariance Analysis of Heart Transplant Survival Data," *Journal of the American Statistical Association*, 72, 27–36.

DeLong, D.M., Guirguis, G.H., and So, Y.C. (1994), "Efficient Computation of Subset Selection Probabilities with Application to Cox Regression," *Biometrika*, 81, 607–611.

Efron, B. (1977), "The Efficiency of Cox's Likelihood Function for Censored Data," *Journal of the American Statistical Association*, 72, 557–565.

Fleming, T.R. and Harrington, D.P (1991), *Counting Processes and Survival Analysis*, New York: John Wiley & Sons, Inc.

Furnival, G.M. and Wilson, R.W. (1974), "Regressions by Leaps and Bounds," *Technometrics*, 16, 499–511.

Gail, M.H., Lubin, J.H., and Rubinstein, L.V. (1981), "Likelihood Calculations for Matched Case-Control Studies and Survival Studies with Tied Death Times," *Biometrika*, 68, 703–707.

Grambsch, P.M. and Therneau, T.M. (1993), *Proportional Hazards Tests and Diagnostics Based on Weighted Residuals*, Research Report 93-002, Minneapolis: University of Minnesota.

Harrell, F.E. (1986), "The PHGLM Procedure," *SUGI Supplemental Library Guide, Version 5 Edition*, Cary, NC: SAS Institute Inc.

Hosmer, D.W., Jr., and Lemeshow, S. (1989), *Applied Logistic Regression*, 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.

Krall, J.M., Uthoff, V.A., and Harley, J. B. (1975), "A Step-up Procedure for Selecting Variables Associated with Survival," *Biometrics*, 31, 49–57.

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

Lin, D.Y. and Wei, L.J. (1989), "The Robust Inference for the Proportional Hazards Model," *Journal of the American Statistical Association*, 84, 1074–1078.

Pettitt, A.N. and Bin Daud, I. (1989), "Case-Weighted Measures of Influence for Proportional Hazards Regression," *Applied Statistics*, 38, 313–329.

Prentice, R.L., Williams, B.J., and Peterson, A.V. (1981), "On the Regression Analysis of Multivariate Failure Time Data," *Biometrika*, 68, 373–379.

Reid, N. and Crèpeau, H. (1985), "Influence Functions for Proportional Hazards Regression," *Biometrika*, 72, 1–9.

Schoenfeld, D. (1982), "Partial Residuals for the Proportional Hazards Regression Model," *Biometrika*, 69, 239–241.

Therneau, T.M. (1994), "A Package for Survival Analysis in S," Technical Report #53, Section of Biostatistics, Mayo Clinic, Rochester.

Therneau, T.M., Grambsch, P.M., and Fleming, T.R. (1990), "Martingale-Based Residuals and Survival Models," *Biometrika*, 77, 147–160.

Tsiatis, A. (1981), "A Large Sample Study of the Estimates for the Integrated Hazard Function in Cox's Regression Model for Survival Data," *Annals of Statistics*, 9, 93–108.

Wei, L.J., Lin, D.Y., and Weissfeld, L. (1989), "Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distribution," *Journal of the American Statistical Association*, 84, 1065–1073.

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.