SAS/STAT® 13.2 User’s Guide
The SURVEYPHREG Procedure
Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

support.sas.com/bookstore
for additional books and resources.
Chapter 100
The SURVEYPHREG Procedure

Contents

Overview: SURVEYPHREG Procedure ........................................... 8240
Getting Started: SURVEYPHREG Procedure ............................... 8241
Syntax: SURVEYPHREG Procedure .............................................. 8244
  PROC SURVEYPHREG Statement .............................................. 8245
  BY Statement ................................................................. 8251
  CLASS Statement ............................................................. 8251
  CLUSTER Statement ......................................................... 8254
  DOMAIN Statement ........................................................... 8254
  ESTIMATE Statement .......................................................... 8255
  FREQ Statement ............................................................... 8256
  LSMEANS Statement ............................................................ 8256
  LSMEANTE Statement ............................................................ 8257
  MODEL Statement ............................................................... 8258
  NLOPTIONS Statement .......................................................... 8262
  OUTPUT Statement .............................................................. 8262
  Programming Statements ....................................................... 8263
  REPWEIGHTS Statement .......................................................... 8265
  SLICE Statement ............................................................... 8266
  STORE Statement .............................................................. 8266
  STRATA Statement .............................................................. 8267
  TEST Statement ................................................................. 8267
  WEIGHT Statement ............................................................... 8268
Details: SURVEYPHREG Procedure ............................................... 8268
  Notation and Estimation ....................................................... 8268
  Failure Time Distribution ...................................................... 8270
  Time and CLASS Variables Usage ........................................... 8270
  Partial Likelihood Function for the Cox Model ................................ 8274
  Specifying the Sample Design ............................................... 8275
  Missing Values ................................................................. 8277
  Variance Estimation ............................................................ 8279
    Taylor Series Linearization .................................................. 8279
    Balanced Repeated Replication (BRR) Method ............................. 8280
    Jackknife Method ............................................................ 8283
    Degrees of Freedom .......................................................... 8284
    Variance Adjustment Factors ............................................... 8285
    Variance Ratios and Standard Error Ratios ............................... 8285
Overview: SURVEYPHREG Procedure

The SURVEYPHREG procedure performs regression analysis based on the Cox proportional hazards model for sample survey data. Cox’s semiparametric model is widely used in the analysis of survival data to estimate hazard rates when adequate explanatory variables are available. The procedure provides design-based variance estimates, confidence intervals, and hypothesis tests concerning the parameters and model effects. See Chapter 3, “Introduction to Statistical Modeling with SAS/STAT Software,” and Chapter 14, “Introduction to Survey Procedures,” for an introduction to the basic concepts of survey data analysis; see Chapter 13, “Introduction to Survival Analysis Procedures,” for an introduction to the basic concepts of survival analysis.

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

$$
\lambda_i(t) = \lambda_0(t) \exp(Z_i(t) \beta)
$$

where $\lambda_0(t)$ is an arbitrary and unspecified baseline hazard function, $Z_i(t)$ is the vector of explanatory variables for the $i$th population unit at time $t$, and $\beta$ is the vector of unknown regression parameters.

The finite population regression parameter $\hat{\beta}_N$ is defined as the maximizer of the partial log likelihood when the entire finite population is observed. The SURVEYPHREG procedure produces a sample-based estimate $\hat{\beta}$ of the proportional hazards regression parameters $\hat{\beta}_N$ for the finite population by maximizing the partial pseudo-log-likelihood $l_N(\beta; Z_i(t_i), t_i)$ based on observed covariates $Z_i(t_i)$ and observed survival time $t_i$. The procedure also produces an estimate of the sampling variance $V(\hat{\beta}|\mathcal{F}_N)$, which assumes that the values of
the finite population $F_N$ are fixed. For statistical inference, PROC SURVEYPHREG incorporates complex survey sample designs, including designs with stratification, clustering, and unequal weighting.

The procedure 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 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.

Several optimization techniques are available in SURVEYPHREG to maximize the log likelihood. Hazard ratio estimates can also be obtained along with parameter estimates. Sampling errors of the regression parameters and hazard ratios are computed by using either the Taylor series (linearization) method or one of the replication (resampling) methods that are based on complex sample designs (Binder 1983; Wolter 2007; Särndal, Swensson, and Wretman 1992; Binder 1992; Lohr 2010; Fuller 2009). These variance estimators essentially assume the finite population as fixed and estimate the variability due to the random sample selection mechanism.

The remaining sections of this chapter contain information about how to use PROC SURVEYPHREG, information about the underlying statistical methodology, and some applications of the procedure. The section “Getting Started: SURVEYPHREG Procedure” on page 8241 introduces PROC SURVEYPHREG with an example. The section “Syntax: SURVEYPHREG Procedure” on page 8244 describes the syntax of the procedure. The section “Details: SURVEYPHREG Procedure” on page 8268 summarizes the statistical techniques employed in PROC SURVEYPHREG. The section “Examples: SURVEYPHREG Procedure” on page 8297 includes some additional examples of useful applications. Experienced SAS/STAT software users might decide to proceed to the “Syntax” section, while other users might choose to read both the “Getting Started” and “Examples” sections before proceeding to “Syntax” and “Details.”

---

**Getting Started: SURVEYPHREG Procedure**

This section uses a data set that is obtained by stratified random sampling from a simulated finite population to illustrate some of the basic features of PROC SURVEYPHREG.

Suppose the library system for a small county wants to study the length of time that books are borrowed over a specified study period, adjusting for the age of the borrower and accounting for the fact that some books are never returned. Suppose there are 10 branch libraries in the county. Assume that a list of 11,617 (simulated) transactions is available for the study period October 1, 2008, to December 31, 2008, and assume that this list can be used as the sampling frame. A stratified random sample with replacement is used to select 100 transactions, where branch libraries are the strata. The total number of transactions within branches range from 510 to 2,011 for the study period. The total sample size of 100 transactions is allocated proportionally across branches based on the number of transactions. For each selected transaction, telephone interviews were conducted to find out additional characteristics of the borrower. The data set LibrarySurvey contains the following variables for all units (transactions) in the sample:

- **Branch**, the library branch from which the book was borrowed
- **SampleWeight**, the survey sampling weight for the transaction
- **CheckOut**, the date the book was borrowed
• CheckIn, the date the book was returned, with a missing value if the book was not returned by December 31, 2008

• Age, the age of the borrower

data LibrarySurvey;
  input Branch 2.
    SamplingWeight 7.2
    CheckOut date10.
    CheckIn date10.
    Age;
datalines;
  1 103.60 08NOV2008 13NOV2008 18
  1 103.60 01OCT2008 07OCT2008 30
  1 103.60 05NOV2008 06NOV2008 73
  1 103.60 25OCT2008 26OCT2008 53
  1 103.60 09NOV2008 10NOV2008 55
  2 127.50 10DEC2008 15DEC2008 39
  2 127.50 19DEC2008 . 33
  2 127.50 26NOV2008 27NOV2008 41
  2 127.50 03NOV2008 07NOV2008 33
  ... more lines ...
  10 118.35 14NOV2008 17NOV2008 29
  10 118.35 11DEC2008 13DEC2008 35
  10 118.35 21NOV2008 23NOV2008 46
;
data LibrarySurvey;
  set LibrarySurvey;
  Returned = (CheckIn ^= .);
  if (Returned) then
    lenBorrow = CheckIn - CheckOut;
  else
    lenBorrow = input('31Dec2008',date9.) - CheckOut;
run;

PROC SURVEYPHREG can be used to estimate the regression parameters of a proportional hazards model and the design-based variance of the estimated coefficients. The design-based variance is useful when the finite population is considered fixed, as in this example. See Lohr (2010) and Särndal, Swensson, and Wretman (1992) for details.

The following statements request a proportional hazards regression of lenBorrow on Age with Returned as the censor indicator. A transaction is considered to be censored if its check-in date is missing. The WEIGHT statement specifies the sampling weight variable (SamplingWeight), and the STRATA statement specifies the stratification variable (Branch).

proc surveyphreg data = LibrarySurvey;
  weight SamplingWeight;
  strata Branch;
  model lenBorrow*Returned(0) = Age;
run;
Summary information about the model, number of observations, survey design, censored values, and variance estimation method are shown in Output 100.1. The “Model Information” table summarizes the model you fit. The “Number of Observations” table displays the number of observations read and used by the procedure. This table also displays the sum of weights read and used. The sum of weights read (11,616.79) can be used as an estimator of the population size, and the sum of weights used can be used as an estimator of the respondent size in the population. The “Design Summary” table displays survey design information such as stratification and clustering. This example implements a stratified design with 10 strata. The “Censored Summary” and “Weighted Censored Summary” tables display the (weighted) number of censored and event units. Weighted counts can be used as estimators of the corresponding finite population quantities. For example, Output 100.1 shows that 10% of the sampled units are censored and an estimated 10.05% of the population units are censored.

Parameter estimates and their standard errors are shown in Output 100.2. The estimated regression coefficient is highly significant with a value of 0.062, indicating a positive association between age and the length of time books are borrowed (recall that these are simulated data). In this example, the procedure uses the
STRATA and WEIGHT statements to incorporate stratification and unequal weighting, respectively, into variance estimation. The degrees of freedom are calculated as the number of sampling units (100) minus the number of strata (10). Note that the estimated variance reported in Output 100.2 ignores the finite population correction (fpc). You can use the TOTAL= or RATE= option in the PROC statement to include an fpc in your variance estimator.

**Figure 100.2** Weighted Estimates and Their Standard Errors

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF Estimate</th>
<th>Standard Error t Value Pr &gt;</th>
<th>t</th>
<th>Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>90</td>
<td>0.061593</td>
<td>0.008366</td>
<td>7.36</td>
</tr>
</tbody>
</table>

**Syntax: SURVEYPHREG Procedure**

The following statements are available in the SURVEYPHREG procedure. Items within <> are optional.

```plaintext
PROC SURVEYPHREG <options> ;
  BY variables ;
  CLASS variable <options> < . . . variable <options>> </ options> ;
  CLUSTER variables ;
  DOMAIN variables = variable* variable = variable . . . ;
  ESTIMATE = 'label' = estimate-specification </ options> ;
  FREQ variable ;
  LSMEANS = model-effects </ options> ;
  LSMESTIMATE = model-effect lsmeans-specification </ options> ;
  MODEL response = * censor(list) > = effects </ options> ;
  NLOPTIONS <options> ;
  OUTPUT = OUT=SAS-data-set <keyword = name . . . keyword = name> </ options> ;
  Programming statements ;
  REPWEIGHTS variables </ options> ;
  SLICE model-effect </ options> ;
  STRATA variables </ option> ;
  STORE = OUT=item-store-name </ LABEL='label' > ;
  TEST = model-effects </ options> ;
  WEIGHT variable ;
```

The PROC SURVEYPHREG and MODEL statements are required. The CLASS statement, if present, must precede the MODEL statement.

The MODEL statement specifies the analysis model. The CLASS statement specifies the categorical variables. The STRATA statement lists the variables that form the strata in a stratified sample design. The CLUSTER statement specifies cluster identification variables in a clustered sample design. The WEIGHT statement names the sampling weight variable. The NLOPTIONS statement specifies the optimization techniques. The REPWEIGHTS statement names replicate weight variables for BRR or jackknife variance estimation. The DOMAIN statement lists the variables that define domains for subpopulation analysis. The BY statement requests completely separate analyses of groups defined by the BY variables.
The rest of this section provides detailed syntax information for each statement, beginning with the PROC SURVEYPHREG statement. The remaining statements are covered in alphabetical order.

The ESTIMATE, LSMEANS, LSMESTIMATE, SLICE, STORE, and TEST statements are also available in other procedures. Summary descriptions of functionality and syntax for these statements are provided in this chapter, and you can find full documentation about them in Chapter 19, “Shared Concepts and Topics.”

**PROC SURVEYPHREG Statement**

PROC SURVEYPHREG <options> ;

The PROC SURVEYPHREG statement invokes the SURVEYPHREG procedure. It also identifies the data set to be analyzed. Table 100.1 summarizes the options available in the PROC SURVEYPHREG statement.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA=</td>
<td>Names the input SAS data set</td>
</tr>
<tr>
<td>MISSING</td>
<td>Treats missing values as a valid category</td>
</tr>
<tr>
<td>NOPRINT</td>
<td>Suppresses all displayed output</td>
</tr>
<tr>
<td>NOMCAR</td>
<td>Uses missing observations specified as <em>not missing completely at random</em></td>
</tr>
<tr>
<td>ORDER=</td>
<td>Specifies the sort order of CLASS variables</td>
</tr>
<tr>
<td>RATE=</td>
<td>Specifies the sampling rate</td>
</tr>
<tr>
<td>TOTAL=</td>
<td>Specifies the total number of primary sampling units</td>
</tr>
<tr>
<td>VARMETHOD=</td>
<td>Specifies the variance estimation method</td>
</tr>
</tbody>
</table>

You can specify the following options in the PROC SURVEYPHREG statement:

**DATA=SAS-data-set**

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

**MISSING**

treats missing values as a valid (nonmissing) category for all categorical variables, which include CLASS, STRATA, CLUSTER, and DOMAIN variables. By default, if you do not specify the MISSING option, an observation is excluded from the analysis if it has a missing value for any of these categorical variables. For more information, see the section “Missing Values” on page 8277.

**NOPRINT**

suppresses all displayed output. Note that this option temporarily disables the Output Delivery System (ODS); see Chapter 20, “Using the Output Delivery System,” for more information.

**NOMCAR**

includes observations with missing values of the analysis variables that are specified in the MODEL statement as *not missing completely at random* (NOMCAR) for Taylor series variance estimation. When you specify the NOMCAR option, PROC SURVEYPHREG computes variance estimates by analyzing the nonmissing values as a domain (subpopulation), where the entire population includes both nonmissing and missing domains. See the section “Missing Values” on page 8277 for details.
By default, PROC SURVEYPHREG excludes an observation from analyses (and the corresponding variance computations) if that observation has a missing value for any of the variables in the MODEL statement. Note that if you specify the MISSING option for classification variables, then the procedure treats the missing values as a valid nonmissing level.

The NOMCAR option applies only to Taylor series variance estimation. The replication methods, which you request with the VARMETHOD=BRR and VARMETHOD=JACKKNIFE options, do not use the NOMCAR option.

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

specifies the sort order for the levels of the classification variables (which are specified in the CLASS statement).

This option applies to the levels for all classification variables, except when you use the (default) ORDER=FORMATTED option with numeric classification variables that have no explicit format. In that case, the levels of such variables are ordered by their internal value.

The ORDER= option can take the following values:

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

By default, ORDER=FORMATTED. For ORDER=FORMATTED and ORDER=INTERNAL, the sort order is machine-dependent.

For more information about sort order, see the chapter on the SORT procedure in the *Base SAS Procedures Guide* and the discussion of BY-group processing in *SAS Language Reference: Concepts*.

**RATE=value | SAS-data-set**

specifies the sampling rate as a nonnegative value, or identifies an input data set that gives the stratum sampling rates in a variable named _RATE_. PROC SURVEYPHREG uses this information to compute a finite population correction for Taylor series variance estimation. The procedure does not use the RATE= option for BRR or jackknife variance estimation, which you request with the VARMETHOD=BRR or VARMETHOD=JACKKNIFE option.

If your sample design has multiple stages, you should specify the first-stage sampling rate, which is the ratio of the number of primary sampling units (PSUs) that are selected to the total number of PSUs in the population.
For a nonstratified sample design, or for a stratified sample design with the same sampling rate in all strata, you should specify a nonnegative `value` for the RATE= option. If your design is stratified with different sampling rates in different strata, then you should name a SAS data set that contains the stratification variables and the stratum sampling rates. See the section “Population Totals and Sampling Rates” on page 8276 for details.

The sampling rate `value` must be a nonnegative number. You can specify `value` as a number between 0 and 1. Or you can specify `value` in percentage form as a number between 1 and 100, and PROC SURVEYPHREG converts that number to a proportion. The procedure treats the value 1 as 100% instead of 1%.

If you do not specify the RATE= or TOTAL= option, then the Taylor series variance estimation does not include a finite population correction. You cannot specify both the TOTAL= option and the RATE= option in the same PROC SURVEYPHREG statement.

**TOTAL=**`value` | **SAS-data-set**

**N=**`value` | **SAS-data-set**

specifies the total number of primary sampling units (PSUs) in the study population as a positive `value`, or identifies an input data set that gives the stratum population totals in a variable named `_TOTAL_`. PROC SURVEYPHREG uses this information to compute a finite population correction for Taylor series variance estimation. The procedure does not use the TOTAL= option for BRR or jackknife variance estimation, which you request with the **VARMETHOD=BRR** or **VARMETHOD=JACKKNIFE** option.

For a nonstratified sample design, or for a stratified sample design with the same population total in all strata, you should specify a positive `value` for the TOTAL= option, which refers to the total number of PSUs in each stratum. If your sample design is stratified with different population totals in different strata, then you should name a SAS data set that contains the stratification variables and the stratum totals. See the section “Population Totals and Sampling Rates” on page 8276 for details.

If you do not specify the TOTAL= or RATE= option, then the Taylor series variance estimation does not include a finite population correction. You cannot specify both the TOTAL= option and the RATE= option in the same PROC SURVEYPHREG statement.

**VARMETHOD=BRR (method-options) | JACKKNIFE (method-options) | TAYLOR**

specifies the variance estimation method. **VARMETHOD=TAYLOR** requests the Taylor series method, which is the default if you do not specify the VARMETHOD= option or a REPWEIGHTS statement. **VARMETHOD=BRR** requests variance estimation by balanced repeated replication (BRR), and **VARMETHOD=JACKKNIFE** requests variance estimation by the delete-1 jackknife method.

For **VARMETHOD=BRR** and **VARMETHOD=JACKKNIFE**, you can specify `method-options` in parentheses following the variance method name. Table 100.2 summarizes the available `method-options`.
### Table 100.2 Variance Estimation Options

<table>
<thead>
<tr>
<th><code>VARMETHOD=</code></th>
<th>Variance Estimation Method</th>
<th>Method Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRR</td>
<td>Balanced repeated replication</td>
<td>`CENTER=FULLSAMPLE</td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>FAY &lt;=value&gt;</code></td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>HADAMARD=SAS-data-set</code></td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>OUTWEIGHTS=SAS-data-set</code></td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>PRINTH</code></td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>REPS=number</code></td>
</tr>
<tr>
<td>JACKKNIFE</td>
<td>Jackknife</td>
<td>`CENTER=FULLSAMPLE</td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>OUTJKCOEFS=SAS-data-set</code></td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>OUTWEIGHTS=SAS-data-set</code></td>
</tr>
<tr>
<td>TAYLOR</td>
<td>Taylor series linearization</td>
<td>None</td>
</tr>
</tbody>
</table>

The following values are available for the `VARMETHOD=` option:

**BRR < (method-options) >**

requests variance estimation by balanced repeated replication (BRR). The BRR method requires a stratified sample design with two primary sampling units (PSUs) in each stratum. If you specify the `VARMETHOD=BRR` option, you must also specify a `STRATA` statement unless you provide replicate weights with a `REPWEIGHTS` statement. See the section “Balanced Repeated Replication (BRR) Method” on page 8280 for details.

You can specify the following `method-options` in parentheses after the `VARMETHOD=BRR` option:

**CENTER=FULLSAMPLE | REPLICATES**

defines how to compute the deviations for the BRR method. `CENTER=FULLSAMPLE` is the default, which computes the deviations of the replicate estimates from the full sample estimate. Alternatively, you can specify `CENTER=REPLICATES` to compute the deviations of the replicate estimates from the average of the replicate estimates. See the section “Balanced Repeated Replication (BRR) Method” on page 8280 for details.

**DETAILS**

displays the maximum likelihood estimates of model parameters for replicate samples when the replicate parameter estimates are available. A replicate sample might not provide useful parameter estimates (replicate estimates), for reasons such as nonconvergence of the optimization or inestimability of some parameters in that replicate sample.

**FAY <=value>**

requests Fay’s method, which is a modification of the BRR method. See the section “Fay’s BRR Method” on page 8282 for details.

You can specify the `value` of the Fay coefficient, which is used in converting the original sampling weights to replicate weights. The Fay coefficient must be a nonnegative number less than 1. By default, the value of the Fay coefficient equals 0.5.
The HADAMARD= input data set must contain at least \( H \) variables, where \( H \) denotes the number of first-stage strata in your design. If the data set contains more than \( H \) variables, PROC SURVEYPHREG uses only the first \( H \) variables. Similarly, the HADAMARD= input data set must contain at least \( H \) observations.

If you do not specify the REPS= method-option, then the number of replicates is taken to be the number of observations in the HADAMARD= input data set. If you specify the number of replicates—for example, REPS=\( nreps \)—then the first \( nreps \) observations in the HADAMARD= data set are used to construct the replicates.

You can specify the PRINTH method-option to display the Hadamard matrix that the procedure uses to construct replicates for BRR.

The OUTWEIGHTS= method-option is not available when you provide replicate weights with a REPWEIGHTS statement.

The PRINTH method-option is not available when you provide replicate weights with a REPWEIGHTS statement because the procedure does not use a Hadamard matrix in this case.
REPS=number
specifies the number of replicates for BRR variance estimation. The value of number must be an integer greater than 1.

If you do not provide a Hadamard matrix with the HADAMARD= method-option, the number of replicates should be greater than the number of strata and should be a multiple of 4. See the section “Balanced Repeated Replication (BRR) Method” on page 8280 for more information. If a Hadamard matrix cannot be constructed for the REPS= value that you specify, the value is increased until a Hadamard matrix of that dimension can be constructed. Therefore, it is possible for the actual number of replicates used to be larger than the REPS= value that you specify.

If you provide a Hadamard matrix with the HADAMARD= method-option, the value of REPS= must not be less than the number of rows in the Hadamard matrix. If you provide a Hadamard matrix and do not specify the REPS= method-option, the number of replicates equals the number of rows in the Hadamard matrix.

If you do not specify the REPS= or HADAMARD= method-option and do not include a REPWEIGHTS statement, the number of replicates equals the smallest multiple of 4 that is greater than the number of strata.

If you provide replicate weights with a REPWEIGHTS statement, the procedure does not use the REPS= method-option. With a REPWEIGHTS statement, the number of replicates equals the number of REPWEIGHTS variables.

JACKKNIFE | JK < (method-options)>
requests variance estimation by the delete-1 jackknife method. See the section “Jackknife Method” on page 8283 for details. If you provide replicate weights with a REPWEIGHTS statement, VARMETHOD=JACKKNIFE is the default variance estimation method. The JACKKNIFE method requires at least two primary sampling units (PSUs) in each stratum for stratified designs unless you provide replicate weights with a REPWEIGHTS statement.

You can specify the following method-options in parentheses following VARMETHOD=JACKKNIFE:

CENTER=FULLSAMPLE | REPLECTES
defines how to compute the deviations for the jackknife method. CENTER=FULLSAMPLE is the default, which computes the deviations of the replicate estimates from the full sample estimate. Alternatively, you can specify CENTER=REPLICATES to compute the deviations of the replicate estimates from the average of the replicate estimates. See the section “Jackknife Method” on page 8283 for details.

DETAILS
displays the maximum likelihood estimates of model parameters for replicate samples when the replicate parameter estimates are available. A replicate sample might not provide useful parameter estimates (replicate estimates), for reasons such as nonconvergence of the optimization or inestimability of some parameters in that replicate sample.

OUTWEIGHTS=SAS-data-set
names an output SAS data set that contains replicate weights. See the section “Jackknife Method” on page 8283 for information about replicate weights. See the section “Replicate Weights Output Data Set” on page 8291 for more details about the contents of the OUTWEIGHTS= data set.
The OUTWEIGHTS= method-option is not available when you provide replicate weights with the REPWEIGHTS statement.

OUTJKCOEFS=SAS-data-set

names an output SAS data set that contains jackknife coefficients. See the section “Jackknife Coefficients Output Data Set” on page 8292 for more details about the contents of the OUTJKCOEFS= data set.

TAYLOR

requests Taylor series variance estimation. This is the default method if you do not specify the VARMETHOD= option or a REPWEIGHTS statement. See the section “Taylor Series Linearization” on page 8279 for more information.

---

**BY Statement**

**BY** variables ;

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

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

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option in the BY statement for the SURVEYPHREG procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables by using the DATASETS procedure (in Base SAS software).

Note that using a BY statement provides completely separate analyses of the BY groups. It does not provide a domain (subpopulation) analysis, where the number of sampling units in the subpopulation is not known at the time the survey is designed. For such an analysis use the DOMAIN statement.

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

---

**CLASS Statement**

**CLASS** variable <(options)> . . . < variable <(options)> > </ options> ;

The CLASS statement names the classification variables to be used as explanatory variables in the analysis. The CLASS statement must precede the MODEL statement. Most options can be specified either as individual variable options or as global options. You can specify options for each variable by enclosing the options in parentheses after the variable name. You can also specify global options for the CLASS statement by placing
the options after a slash (/). Global options are applied to all the variables specified in the CLASS statement. If you specify more than one CLASS statement, the global options specified in any one CLASS statement apply to all CLASS statements. However, individual CLASS variable options override the global options. The following options are available:

**DESCENDING**

**DESC**

reverses the sort order of the classification variable. If both the DESCENDING and ORDER= options are specified, PROC SURVEYPHREG orders the categories according to the ORDER= option and then reverses that order.

**MISSING**

treats missing values (".", _, .A, . . . , .Z for numeric variables and blanks for character variables) as valid values for the CLASS variable.

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

specifies the sort order for the levels of classification variables. This ordering determines which parameters in the model correspond to each level in the data, so the ORDER= option can be useful when you use the CONTRAST statement. By default, ORDER=FORMATTED. For ORDER=FORMATTED and ORDER=INTERNAL, the sort order is machine-dependent. When ORDER=FORMATTED is in effect for numeric variables for which you have supplied no explicit format, the levels are ordered by their internal values.

The following table shows how PROC SURVEYPHREG interprets values of the ORDER= option.

<table>
<thead>
<tr>
<th>Value of ORDER=</th>
<th>Levels Sorted By</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA</td>
<td>Order of appearance in the input data set</td>
</tr>
<tr>
<td>FORMATTED</td>
<td>External formatted values, except for numeric variables with no explicit format, which are sorted by their unformatted (internal) values</td>
</tr>
<tr>
<td>FREQ</td>
<td>Descending frequency count; levels with more observations come earlier in the order</td>
</tr>
<tr>
<td>INTERNAL</td>
<td>Unformatted value</td>
</tr>
</tbody>
</table>

For more information about sort order, see the chapter on the SORT procedure in the *Base SAS Procedures Guide* and the discussion of BY-group processing in *SAS Language Reference: Concepts*.

**PARAM=keyword**

specifies the parameterization method for the classification variable or variables. You can specify any of the keywords shown in the following table;

Design matrix columns are created from CLASS variables according to the corresponding coding schemes:
### Value of PARAM= Coding

<table>
<thead>
<tr>
<th>PARAM=</th>
<th>Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>EFFECT</td>
<td>Effect coding</td>
</tr>
<tr>
<td>GLM</td>
<td>Less-than-full-rank reference cell coding (this <em>keyword</em> can be used only in a global option)</td>
</tr>
<tr>
<td>ORDINAL</td>
<td>Cumulative parameterization for an ordinal CLASS variable</td>
</tr>
<tr>
<td>THERMOMETER</td>
<td></td>
</tr>
<tr>
<td>POLYNOMIAL</td>
<td>Polynomial coding</td>
</tr>
<tr>
<td>POLY</td>
<td></td>
</tr>
<tr>
<td>REF</td>
<td>Reference cell coding</td>
</tr>
<tr>
<td>ORTHEFF</td>
<td>Orthogonalizes PARAM=EFFECT coding</td>
</tr>
<tr>
<td>ORTHORD</td>
<td>Orthogonalizes PARAM=ORDINAL coding</td>
</tr>
<tr>
<td>ORTHOTH</td>
<td></td>
</tr>
<tr>
<td>ORTHPOLY</td>
<td>Orthogonalizes PARAM=POLYNOMIAL coding</td>
</tr>
<tr>
<td>ORTHREF</td>
<td>Orthogonalizes PARAM=REFERENCE coding</td>
</tr>
</tbody>
</table>

All parameterizations are full rank, except for the GLM parameterization. The REF= option in the CLASS statement determines the reference level for EFFECT and REFERENCE coding and for their orthogonal parameterizations. It also indirectly determines the reference level for a singular GLM parameterization through the order of levels.

If PARAM=ORTHPOLY or PARAM=POLY and the classification variable is numeric, then the ORDER= option in the CLASS statement is ignored, and the internal unformatted values are used. See the section “Other Parameterizations” on page 391 in Chapter 19, “Shared Concepts and Topics,” for further details.

**REF= 'level' | keyword**

specifies the reference level for PARAM=EFFECT, PARAM=REFERENCE, and their orthogonalizations. For PARAM=GLM, the REF= option specifies a level of the classification variable to be put at the end of the list of levels. This level thus corresponds to the reference level in the usual interpretation of the linear estimates with a singular parameterization.

For an individual variable REF= option (but not for a global REF= option), you can specify the *level* of the variable to use as the reference level. Specify the formatted value of the variable if a format is assigned. For a global or individual variable REF= option, you can use one of the following *keywords*. The default is REF=LAST.

**FIRST** designates the first ordered level as reference.

**LAST** designates the last ordered level as reference.

**TRUNCATE< ==n>**

specifies the length *n* of CLASS variable values to use in determining CLASS variable levels. The default is to use the full formatted length of the CLASS variable. If you specify TRUNCATE without the length *n*, the first 16 characters of the formatted values are used. When formatted values are longer
than 16 characters, you can use this option to revert to the levels as determined in releases before SAS 9. The TRUNCATE option is available only as a global option.

---

**CLUSTER Statement**

`CLUSTER variables ;`

The CLUSTER statement names variables that identify the first-stage clusters in a clustered sample design. First-stage clusters are also known as primary sampling units (PSUs). The combinations of categories of CLUSTER variables define the clusters in the sample. If there is a STRATA statement, clusters are nested within strata.

If your sample design has clustering at multiple stages, you should specify only the first-stage clusters (PSUs) in the CLUSTER statement. See the section “Specifying the Sample Design” on page 8275 for more information.

If you provide replicate weights for BRR or jackknife variance estimation with a REPWEIGHTS statement, you do not need to specify a CLUSTER statement.

The CLUSTER variables are one or more variables in the DATA= input data set. These variables can be either character or numeric, but the procedure treats them as categorical variables. The formatted values of the CLUSTER variables determine the CLUSTER variable levels. Thus, you can use formats to group values into levels. See the discussion of the FORMAT procedure in the Base SAS Procedures Guide and the discussions of the FORMAT statement and SAS formats in SAS Formats and Informats: Reference.

You can use multiple CLUSTER statements to specify CLUSTER variables. The procedure uses variables from all CLUSTER statements to create clusters. Cluster variables must not occur in the CLASS statement.

---

**DOMAIN Statement**

`DOMAIN variables < variable* variable variable* variable variable variable ... > ;`

The DOMAIN statement requests analysis for domains (subpopulations), in addition to analysis for the entire study population. The DOMAIN statement names the variables that identify domains, which are called domain variables.

It is common practice to compute statistics for domains. The formation of these domains might not be known at the design stage. Therefore, the sample sizes for the domains are often random. Use a DOMAIN statement to incorporate this variability into the variance estimation.

Note that a DOMAIN statement is different from a BY statement. In a BY statement, you treat the sample sizes as fixed in each subpopulation, and you perform analysis within each BY group independently.

Use the DOMAIN statement on the entire data set to perform a domain analysis. Creating a new data set from a single domain and analyzing that with PROC SURVEYPHREG yields inappropriate estimates of variance.

A domain variable can be either character or numeric. The procedure treats domain variables as categorical variables. If a variable appears by itself in a DOMAIN statement, each level of this variable determines a domain in the study population. If two or more variables are joined by asterisks (*), then every possible combination of levels of these variables determines a domain. The procedure performs a descriptive analysis
ESTIMATE Statement

Within each domain that is defined by the domain variables. Domain variables must not occur in the CLASS statement.

The formatted values of the domain variables determine the categorical variable levels. Thus, you can use formats to group values into levels. For more information, see the FORMAT procedure in the Base SAS Procedures Guide and the FORMAT statement and SAS formats in the SAS Formats and Informats: Reference.

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

Table 100.3 summarizes the options available in the ESTIMATE statement.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIVISOR=</td>
<td>Specifies a list of values to divide the coefficients</td>
</tr>
<tr>
<td>NOFILL</td>
<td>Suppresses the automatic fill-in of coefficients for higher-order effects</td>
</tr>
<tr>
<td>SINGULAR=</td>
<td>Tunes the estimability checking difference</td>
</tr>
<tr>
<td>ADJUST=</td>
<td>Determines the method for multiple comparison adjustment of estimates</td>
</tr>
<tr>
<td>ALPHA=( \alpha )</td>
<td>Determines the confidence level ((1 - \alpha))</td>
</tr>
<tr>
<td>LOWER</td>
<td>Performs one-sided, lower-tailed inference</td>
</tr>
<tr>
<td>STEPDOWN</td>
<td>Adjusts multiplicity-corrected ( p )-values further in a step-down fashion</td>
</tr>
<tr>
<td>TESTVALUE=</td>
<td>Specifies values under the null hypothesis for tests</td>
</tr>
<tr>
<td>UPPER</td>
<td>Performs one-sided, upper-tailed inference</td>
</tr>
</tbody>
</table>

Statistical Output

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL</td>
<td>Constructs confidence limits</td>
</tr>
<tr>
<td>CORR</td>
<td>Displays the correlation matrix of estimates</td>
</tr>
<tr>
<td>COV</td>
<td>Displays the covariance matrix of estimates</td>
</tr>
<tr>
<td>E</td>
<td>Prints the ( L ) matrix</td>
</tr>
<tr>
<td>JOINT</td>
<td>Produces a joint ( F ) or chi-square test for the estimable functions</td>
</tr>
<tr>
<td>SEED=</td>
<td>Specifies the seed for computations that depend on random numbers</td>
</tr>
</tbody>
</table>
For details about the syntax of the ESTIMATE statement, see the section “ESTIMATE Statement” on page 444 in Chapter 19, “Shared Concepts and Topics.”

FREQ Statement

FREQ variable ;

The FREQ statement names a numeric variable that provides a frequency for each observation in the input data set. PROC SURVEYPHREG 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 analysis. The FREQ statement allows one frequency variable.

If you use the FREQ statement and request the jackknife or BRR variance estimator by specifying the VARMETHOD=JACKKNIFE or VARMETHOD=BRR option in the PROC SURVEYPHREG statement, then you must identify the primary sampling units with a CLUSTER statement unless you also provide replicate weights with a REPWEIGHTS statement.

LSMEANS Statement

LSMEANS <model-effects></options> ;

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

Table 100.4 summarizes the options available in the LSMEANS statement.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Construction and Computation of LS-Means</td>
<td></td>
</tr>
<tr>
<td>AT</td>
<td>Modifies the covariate value in computing LS-means</td>
</tr>
<tr>
<td>BYLEVEL</td>
<td>Computes separate margins</td>
</tr>
<tr>
<td>DIFF</td>
<td>Requests differences of LS-means</td>
</tr>
<tr>
<td>OM=</td>
<td>Specifies the weighting scheme for LS-means computation as determined by the input data set</td>
</tr>
<tr>
<td>SINGULAR=</td>
<td>Tunes estimability checking</td>
</tr>
<tr>
<td>Degrees of Freedom and p-values</td>
<td></td>
</tr>
<tr>
<td>ADJUST=</td>
<td>Determines the method for multiple-comparison adjustment of LS-means differences</td>
</tr>
<tr>
<td>ALPHA=α</td>
<td>Determines the confidence level (1 − α)</td>
</tr>
<tr>
<td>STEPDOWN</td>
<td>Adjusts multiple-comparison p-values further in a step-down fashion</td>
</tr>
</tbody>
</table>
Table 100.4 continued

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Statistical Output</strong></td>
<td></td>
</tr>
<tr>
<td>CL</td>
<td>Constructs confidence limits for means and mean differences</td>
</tr>
<tr>
<td>CORR</td>
<td>Displays the correlation matrix of LS-means</td>
</tr>
<tr>
<td>COV</td>
<td>Displays the covariance matrix of LS-means</td>
</tr>
<tr>
<td>E</td>
<td>Prints the L matrix</td>
</tr>
<tr>
<td>LINES</td>
<td>Produces a “Lines” display for pairwise LS-means differences</td>
</tr>
<tr>
<td>MEANS</td>
<td>Prints the LS-means</td>
</tr>
<tr>
<td>PLOTS=</td>
<td>Requests graphs of means and mean comparisons</td>
</tr>
<tr>
<td>SEED=</td>
<td>Specifies the seed for computations that depend on random numbers</td>
</tr>
</tbody>
</table>

For details about the syntax of the LSMEANS statement, see the section “LSMEANS Statement” on page 460 in Chapter 19, “Shared Concepts and Topics.”

**LSMESTITUTE Statement**

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

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

Table 100.5 summarizes the options available in the LSMESTITUTE statement.

Table 100.5 LSMESTITUTE Statement Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Construction and Computation of LS-Means</strong></td>
<td></td>
</tr>
<tr>
<td>AT</td>
<td>Modifies covariate values in computing LS-means</td>
</tr>
<tr>
<td>BYLEVEL</td>
<td>Computes separate margins</td>
</tr>
<tr>
<td>DIVISOR=</td>
<td>Specifies a list of values to divide the coefficients</td>
</tr>
<tr>
<td>OM=</td>
<td>Specifies the weighting scheme for LS-means computation as determined by a data set</td>
</tr>
<tr>
<td>SINGULAR=</td>
<td>Tunes estimability checking</td>
</tr>
</tbody>
</table>
### Table 100.5 continued

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degrees of Freedom and (p)-values</td>
<td></td>
</tr>
<tr>
<td>ADJUST=</td>
<td>Determines the method for multiple-comparison adjustment of LS-means differences</td>
</tr>
<tr>
<td>ALPHA=(\alpha)</td>
<td>Determines the confidence level ((1 - \alpha))</td>
</tr>
<tr>
<td>LOWER</td>
<td>Performs one-sided, lower-tailed inference</td>
</tr>
<tr>
<td>STEPDOWN</td>
<td>Adjusts multiple-comparison (p)-values further in a step-down fashion</td>
</tr>
<tr>
<td>TESTVALUE=</td>
<td>Specifies values under the null hypothesis for tests</td>
</tr>
<tr>
<td>UPPER</td>
<td>Performs one-sided, upper-tailed inference</td>
</tr>
<tr>
<td>Statistical Output</td>
<td></td>
</tr>
<tr>
<td>CL</td>
<td>Constructs confidence limits for means and mean differences</td>
</tr>
<tr>
<td>CORR</td>
<td>Displays the correlation matrix of LS-means</td>
</tr>
<tr>
<td>COV</td>
<td>Displays the covariance matrix of LS-means</td>
</tr>
<tr>
<td>E</td>
<td>Prints the (L) matrix</td>
</tr>
<tr>
<td>ELSM</td>
<td>Prints the (K) matrix</td>
</tr>
<tr>
<td>JOINT</td>
<td>Produces a joint (F) or chi-square test for the LS-means and LS-means differences</td>
</tr>
<tr>
<td>SEED=</td>
<td>Specifies the seed for computations that depend on random numbers</td>
</tr>
</tbody>
</table>

For details about the syntax of the LSMESTIMATE statement, see the section “LSMESTIMATE Statement” on page 476 in Chapter 19, “Shared Concepts and Topics.”

## MODEL Statement

```plaintext
MODEL response < * censor (list) > = effects < / options > ;
```

The MODEL statement identifies the variable to be used as the failure time variable, the optional censoring variable, and the explanatory effects, including covariates, main effects, and interactions; see the section “Specification of Effects” on page 3453 in Chapter 45, “The GLM Procedure,” for more information. A note of caution: specifying the effect \(T*A\) in the MODEL statement, where \(T\) is the time variable and \(A\) is a CLASS variable, does not make the effect time-dependent. You must specify exactly one MODEL statement.

The MODEL statement allows one response variable. In the MODEL statement, the failure time variable precedes the equal sign. 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.

The censoring variable must be numeric. The failure time variable 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. See “Missing Values” on page 8277 for details.
Table 100.6 summarizes the options available in the MODEL statement, which can be specified after a slash (/).

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALPHA=</td>
<td>Specifies $\alpha$ for the $100(1-\alpha)%$ confidence limits</td>
</tr>
<tr>
<td>CLPARM</td>
<td>Computes confidence limits for regression parameters</td>
</tr>
<tr>
<td>COVB</td>
<td>Displays covariance matrix</td>
</tr>
<tr>
<td>DF=</td>
<td>Specifies the denominator degrees of freedom</td>
</tr>
<tr>
<td>HESS</td>
<td>Displays the Hessian matrix</td>
</tr>
<tr>
<td>INVHESS</td>
<td>Displays the inverse of the Hessian matrix</td>
</tr>
<tr>
<td>RISKLIMITS</td>
<td>Computes confidence limits for the exponentials of the regression parameters</td>
</tr>
<tr>
<td>SERATIO=</td>
<td>Computes the ratio of two standard errors for the regression coefficients</td>
</tr>
<tr>
<td>SINGULAR=</td>
<td>Specifies tolerance for testing singularity</td>
</tr>
<tr>
<td>TIES=</td>
<td>Specifies the method of handling ties in failure times</td>
</tr>
<tr>
<td>VADJUST=</td>
<td>Specifies a variance adjustment factor</td>
</tr>
<tr>
<td>VARRATIO=</td>
<td>Computes the ratio of two variances for the regression coefficients</td>
</tr>
</tbody>
</table>

**ALPHA=** sets the level of the confidence limits for the estimated regression parameters and the hazard ratios. The value of alpha must be between 0 and 1, and the default is 0.05. A confidence level of $\alpha$ produces $100(1-\alpha)\%$ confidence limits. The default of ALPHA=0.05 produces 95% confidence limits.

The ALPHA= option has no effect unless the CLPARM or RISKLIMITS option is also specified.

**CLPARM**
produces confidence limits for regression parameters of Cox proportional hazards models. The confidence coefficient can be specified with the ALPHA= option. Classification main effects that use parameterizations other than REF, EFFECT, or GLM are ignored. See “Confidence Intervals” on page 8288 for details.

**COVB**
displays the estimated covariance matrix of the parameter estimates.

**DF=value | keyword**
specifies the denominator degrees of freedom for hypothesis tests and the degrees of freedom to use for confidence limits. If a value is specified, it must be a nonnegative number. By default, PROC SURVEYPHREG computes the degrees of freedom as described in the section “Degrees of Freedom” on page 8284. DF=value is also used to adjust the Wald test statistics that are reported in the “Testing Global Null Hypothesis” table. For more information, see the section “Testing the Global Null Hypothesis” on page 8287. Instead of a value, you can specify one of the following keywords:
NONE specifies the denominator degrees of freedom to be infinite. Use this option if you want to compute chi-square tests and normal confidence intervals. This option is applicable to both the Taylor series linearization and the replication methods.

PARMADJ computes the denominator degrees of freedom as the number of clusters (or observations if no CLUSTER statement is specified) minus the number of strata (or one if no STRATA statement is specified) minus the number of nonsingular parameters plus one in the model. This option can be useful if you are fitting a model with many parameters relative to the number of clusters minus the number of strata. See Korn and Graubard (1999, section 5.2) for further details. This option is applicable only for the Taylor series linearization method.

ALLREPS computes the denominator degrees of freedom for replication methods by using the total number of replicate samples. By default, PROC SURVEYPHREG computes the denominator degrees of freedom based on the number of replicate samples used. See “Degrees of Freedom” on page 8284 for details.

HESS displays the last evaluation of the Hessian matrix.

INVHESS displays the inverse of the Hessian matrix that is evaluated at the estimated regression parameters.

RISKLIMITS

RL produces confidence limits for hazard ratios and related quantities. See the section “Hazard Ratios” on page 8289 for details. The confidence coefficient can be specified with the ALPHA= option. Great care needs to be taken with any interpretation of the estimates and their confidence limits if interaction effects are involved in the model or if parameterizations other than REF, EFFECT, or GLM are used.

SERATIO=ALL | MODEL | IND computes the ratio of two standard errors for the regression parameters. The standard error in the numerator uses the complete design information that you specify. You can specify the following options to compute different standard errors for the denominator:

ALL requests both MODEL and IND standard error ratios.

MODEL computes the standard errors in the denominator as the square root of the diagonals of the inverse Hessian matrix evaluated at the estimated regression parameters. See the section “Variance Ratios and Standard Error Ratios” on page 8285 for more information.

IND computes the standard errors in the denominator by ignoring stratification and clustering. See the section “Variance Ratios and Standard Error Ratios” on page 8285 for more information.
MODEL Statement    8261

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

TIES=method
specifies how to handle ties in the failure time. The available methods are as follows:

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

**EFRON**
uses the approximate partial likelihood of Efron (1977).

If there are no ties, both methods result in the same likelihood and yield identical estimates. The default, TIES=BRESLOW, is the most efficient method when there are no ties.

VADJUST=DF | PARMADJ | NONE | AVGREPSS
specifies variance adjustment factors. You can use the following key words:

**DF**

**PARMADJ**
requests degrees of freedom adjustment $(n - 1)/(n - p)$ in the computation of the matrix $G$ for the Taylor series linearization variance estimation. By default, VADJUST=DF.

**NONE**
excludes the degrees of freedom adjustment $(n - 1)/(n - p)$ from the computation of the matrix $G$ for the Taylor series linearization variance estimation.

**AVGREPSS**
use the average sum of squares from all the usable replicate samples for the unusable replicates. This option is applicable only for the jackknife replication method. AVGREPSS multiplies the default jackknife variance estimator by the factor $R/R_a$, where $R_a$ is the number of usable replicates and $R$ is the total number of replicates. See the section “Variance Adjustment Factors” on page 8285 for details.

VARRATIO=ALL | MODEL | IND
computes the ratio of two variances for the regression parameters. The variance in the numerator uses the complete design information. You can specify the following options to compute different variances for the denominator:

**ALL**
requests both MODEL and IND variance ratios.

**MODEL**
computes the variances in the denominator as the diagonals of the inverse Hessian matrix evaluated at the estimated regression parameters. See the section “Variance Ratios and Standard Error Ratios” on page 8285 for more information.
computes the variances in the denominator by ignoring stratification and clustering. See the section “Variance Ratios and Standard Error Ratios” on page 8285 for more information.

**NLOPTIONS Statement**

**NLOPTIONS < options > ;**

The NLOPTIONS statement specifies details of the nonlinear optimization used by PROC SURVEYPHREG to maximize the log likelihood function. By default, the procedure uses the Newton-Raphson optimization technique. For more information about the NLOPTIONS statement, see the section “NLOPTIONS Statement” on page 488 in Chapter 19, “Shared Concepts and Topics.”

**OUTPUT Statement**

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

The OUTPUT statement creates a new SAS data set that contains statistics that are calculated for each observation unit. These statistics can include the estimated linear predictor ($z_j \hat{\beta}$) and its standard error, residuals, and influence statistics. In addition, this data set includes all the variables from the DATA= input data set.

Only score residuals are available in the OUTPUT data set if the model contains a time-dependent variable that is defined by means of programming statements.

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 named by using the DATA convention. See the section “OUT= Data Set for the OUTPUT statement” on page 8291 for more information.

**keyword=name**

specifies the statistics to include 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**), and optionally an equal sign with either a variable or a list of variables in parentheses to contain the statistics. If you specify a **keyword** without a variable name, then the procedure uses default names. The **keywords** that accept a list of variables are RESSCH, RESSCO, and WTRESSCH. For these **keywords**, you can specify as many names in **name** as the number of explanatory variables in the MODEL statement. If you specify **k** names and **k** is less than the total number of explanatory variables, only the first **k** names are taken from the list; the procedure assigns default names for the rest of the statistics. The **keywords** and the corresponding statistics are as follows:

**ATRISK**

specifies the number of subjects at risk at the observation time $\tau_j$. 
**RESDEV**

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

**RESMART**

specifies the martingale residual $\hat{M}_j$. The residual at the observation time $\tau_j$ 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**

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

**RESSCO**

specifies the score residuals. These residuals are a decomposition of the first partial derivative of the log likelihood. They can be used to assess the leverage that is exerted by each subject in the parameter estimation. They are also useful in constructing design-based variance estimators.

**STDXBETA**

specifies the standard error of the estimated linear predictor, $\sqrt{z_j' \hat{\Sigma}^{-1} \hat{\beta} z_j}$.

**WTATRISK**

specifies the weighted number of subjects at risk at the observation time $\tau_j$.

**XBETA**

specifies the estimate of the linear predictor, $z_j' \hat{\beta}$. 

---

**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. PROC SURVEYPHREG programming statements cannot be used to create or modify the values of the response variable, the censoring variable, the frequency variable, the WEIGHT variable, the CLASS variables, the STRATA variables, the CLUSTER variables, or the DOMAIN variables.

The following DATA step statements are available in PROC SURVEYPHREG:
By default, the PUT statement in PROC SURVEYPHREG writes results 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 statement:

```
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, see the *SAS Functions and CALL Routines: Reference*.

Consider the following example of using programming statements in PROC SURVEYPHREG. Suppose blood pressure is measured at multiple times during the course of a study that investigates the effect of blood pressure on some survival time. By treating the blood pressure as a time-dependent explanatory variable, you can 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.

<table>
<thead>
<tr>
<th>Time</th>
<th>survival time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Censor</td>
<td>censoring indicator (with 0 as the censoring value)</td>
</tr>
<tr>
<td>BP0</td>
<td>blood pressure on entry to the study</td>
</tr>
<tr>
<td>T1</td>
<td>time 1</td>
</tr>
<tr>
<td>BP1</td>
<td>blood pressure at T1</td>
</tr>
<tr>
<td>T2</td>
<td>time 2</td>
</tr>
<tr>
<td>BP2</td>
<td>blood pressure at T2</td>
</tr>
<tr>
<td>WT</td>
<td>design weight</td>
</tr>
</tbody>
</table>
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.

```r
proc surveyphreg;
  weight WT;
  model Time*Censor(0)=BP;
  cluster PSU;
  BP = BP0;
  if Time>T1 and T1^=. then BP=BP1;
  if Time=T2 and T2^=. then BP=BP2;
run;
```

**REPWEIGHTS Statement**

**REPWEIGHTS variables < / options > ;**

The REPWEIGHTS statement names variables that provide replicate weights for BRR or jackknife variance estimation, which you request with the `VARMETHOD=BRR` or `VARMETHOD=JACKKNIFE` option in the `PROC SURVEYPHREG` statement. If you do not provide replicate weights for these methods by using a REPWEIGHTS statement, then PROC SURVEYPHREG constructs replicate weights for the analysis. See the sections “Balanced Repeated Replication (BRR) Method” on page 8280 and “Jackknife Method” on page 8283 for more information.

Each REPWEIGHTS variable should contain the weights for a single replicate, and the number of replicates equals the number of REPWEIGHTS variables. The REPWEIGHTS variables must be numeric, and the variable values must be nonnegative numbers.

If you provide replicate weights with a REPWEIGHTS statement, you do not need to specify a CLUSTER or STRATA statement. If you use a REPWEIGHTS statement and do not specify the `VARMETHOD=` option in the PROC SURVEYPHREG statement, the procedure uses `VARMETHOD=JACKKNIFE` by default.

If you specify a REPWEIGHTS statement but do not include a WEIGHT statement, PROC SURVEYPHREG uses the average of each observation’s replicate weights as the observation’s weight.

You can specify the following options in the REPWEIGHTS statement after a slash (/):

- **DF=** `df`
  specifies the degrees of freedom for the analysis. The value of `df` must be a positive number. See the section “Degrees of Freedom” on page 8284 for details.
  
  PROC SURVEYPHREG also use the DF= value in computing the denominator degrees of freedom for the \( F \) statistics in Wald type tests and confidence intervals.

- **JKCOEFS=** `jackknife-coefficient-specification`
  specifies jackknife coefficients for `VARMETHOD=JACKKNIFE`. The default value for the jackknife coefficient is \((R-1)/R\), where \( R \) is the total number of replicates. You can specify an alternative value with one of the following three forms:
**JKCOEFS=**<em>value</em>
specifies a single jackknife coefficient for all replicates. The coefficient <em>value</em> must be a nonnegative number.

**JKCOEFS=(**<em>values</em>**)**
specifies jackknife coefficients for VARMETHOD=JACKKNIFE, where each coefficient corresponds to an individual replicate identified by a REPWEIGHTS variable. You can separate <em>values</em> with blanks or commas. The coefficient <em>values</em> must be nonnegative numbers. The number of <em>values</em> must equal the number of replicate weight variables named in the REPWEIGHTS statement. List these values in the same order in which you list the corresponding replicate weight variables in the REPWEIGHTS statement.

**JKCOEFS=**<em>SAS-data-set</em>
names a SAS data set that contains the jackknife coefficients for VARMETHOD=JACKKNIFE. You provide the jackknife coefficients in the JKCOEFS= data set variable JKCoefficient. Each coefficient value must be a nonnegative number. The observations in the JKCOEFS= data set should correspond to the replicates that are identified by the REPWEIGHTS variables. Arrange the coefficients or observations in the JKCOEFS= data set in the same order in which you list the corresponding replicate weight variables in the REPWEIGHTS statement. The number of observations in the JKCOEFS= data set must not be less than the number of REPWEIGHTS variables.

See the section “Jackknife Method” on page 8283 for details about jackknife coefficients.

---

**SLICE Statement**

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

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

The SLICE statement uses the same <em>options</em> as the LSMEANS statement, which are summarized in Table 19.21. For details about the syntax of the SLICE statement, see the section “SLICE Statement” on page 505 in Chapter 19, “Shared Concepts and Topics.”

---

**STORE Statement**

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

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

For details about the syntax of the STORE statement, see the section “STORE Statement” on page 508 in Chapter 19, “Shared Concepts and Topics.”
STRATA Statement

    STRATA variables < / option > ;

The STRATA statement names variables that form the strata in a stratified sample design. The combinations of levels of STRATA variables define the strata in the sample, where strata are nonoverlapping subgroups that were sampled independently.

If your sample design has stratification at multiple stages, you should identify only the first-stage strata in the STRATA statement. See the section “Specifying the Sample Design” on page 8275 for more information.

If you provide replicate weights for BRR or jackknife variance estimation with a REPWEIGHTS statement, you do not need to specify a STRATA statement.

The STRATA variables are one or more variables in the DATA= input data set. These variables can be either character or numeric, but the procedure treats them as categorical variables. The formatted values of the STRATA variables determine the STRATA variable levels. Thus, you can use formats to group values into levels. See the discussion of the FORMAT procedure in the Base SAS Procedures Guide and the discussions of the FORMAT statement and SAS formats in the SAS Formats and Informats: Reference. Strata variables must not occur in the CLASS statement.

The STRATA statement in PROC SURVEYPHREG is different from the STRATA statement in PROC PHREG (Chapter 73, “The PHREG Procedure”). PROC PHREG fits different baseline hazard functions in different strata, which is useful if the proportional hazards assumption is not satisfied.

You can specify the following option in the STRATA statement after a slash (/):

LIST

displays a “Stratum Information” table, which lists all strata together with the corresponding values of the STRATA variables. This table provides the number of observations and the number of clusters in each stratum, as well as the sampling fraction if you specify the RATE= or TOTAL= option.

TEST Statement

    TEST < model-effects > < / options > ;

The TEST statement enables you to perform F tests for model effects that test Type I, Type II, or Type III hypotheses. See Chapter 15, “The Four Types of Estimable Functions,” for details about the construction of Type I, II, and III estimable functions.

Table 100.7 summarizes the options available in the TEST statement.
Table 100.7  TEST Statement Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHISQ</td>
<td>Requests chi-square tests</td>
</tr>
<tr>
<td>DDF=</td>
<td>Specifies denominator degrees of freedom for fixed effects</td>
</tr>
<tr>
<td>E</td>
<td>Requests Type I, Type II, and Type III coefficients</td>
</tr>
<tr>
<td>E1</td>
<td>Requests Type I coefficients</td>
</tr>
<tr>
<td>E2</td>
<td>Requests Type II coefficients</td>
</tr>
<tr>
<td>E3</td>
<td>Requests Type III coefficients</td>
</tr>
<tr>
<td>HTYPE=</td>
<td>Indicates the type of hypothesis test to perform</td>
</tr>
<tr>
<td>INTERCEPT</td>
<td>Adds a row that corresponds to the overall intercept</td>
</tr>
</tbody>
</table>

For details about the syntax of the TEST statement, see the section “TEST Statement” on page 509 in Chapter 19, “Shared Concepts and Topics.”

**WEIGHT Statement**

WEIGHT variable ;

The WEIGHT statement names the variable that contains the sampling weights. This variable must be numeric, and the sampling weights must be positive numbers. If an observation has a weight that is nonpositive or missing, then the procedure omits that observation from the analysis. See the section “Missing Values” on page 8277 for more information. The WEIGHT statement allows one weight variable.

If you do not specify a WEIGHT statement but provide replicate weights with a REPWEIGHTS statement, PROC SURVEYPHREG uses the average of each observation’s replicate weights as the observation’s weight.

If you specify neither a WEIGHT statement nor a REPWEIGHTS statement, PROC SURVEYPHREG assigns all observations a weight of one.

**Details: SURVEYPHREG Procedure**

**Notation and Estimation**

Let \( U = \{1, 2, \ldots, N\} \) be the set of indices and let \( \mathcal{F}_N \) be the set of values for a finite population of size \( N \). The survival time of each member of the finite population is assumed to follow its own hazard function, \( \lambda_i(t) \), expressed as

\[
\lambda_i(t) = \lambda(t; \mathbf{Z}_i(t)) = \lambda_0(t) \exp(\mathbf{Z}_i^t(\mathbf{t})\mathbf{\beta})
\]

where \( \lambda_0(t) \) is an arbitrary and unspecified baseline hazard function, \( \mathbf{Z}_i(t) \) is the vector of explanatory variables for the \( i \)th unit at time \( t \), and \( \mathbf{\beta} \) is the vector of unknown regression parameters that are associated with the explanatory variables. The vector \( \mathbf{\beta} \) is assumed to be the same for all individuals.
The partial likelihood function introduced by Cox (1972, 1975) eliminates the unknown baseline hazard $\lambda_0(t)$ and accounts for censored survival times. If the entire population is observed, then this partial likelihood can be used to estimate $\beta$. Let $\hat{\beta}_N$ be the desired estimator. Assuming a working model with uncorrelated responses, $\hat{\beta}_N$ is obtained by maximizing the partial log likelihood,

$$l(\beta) = \sum_{i \in U} \log \left\{ L(\beta; Z_i(t), t_i) \right\}$$

with respect to $\beta$, where $L(\beta; Z_i(t), t_i)$ is Cox’s partial likelihood function.

Assume that probability sample $A$ is selected from the finite population $U$ and $\pi_i$ is the selection probability for unit $i$. Further assume that covariates $Z_i(t)$ and survival time $t_i$ are available for every unit in the sample $A$. An estimator of the finite population log likelihood is

$$l_{\pi}(\beta) = \sum_{i \in A} \pi_i^{-1} \log \left\{ L(\beta; Z_i(t), t_i) \right\}$$

See “Partial Likelihood Function for the Cox Model” on page 8274 for more details.

A sample-based estimator $\hat{\beta}$ for the finite population quantity $\beta_N$ can be obtained by maximizing the partial pseudo-log-likelihood $l_{\pi}(\beta; Z_i(t), t_i)$ with respect to $\beta$. The design-based variance for $\hat{\beta}$ is obtained by assuming the set of finite population values $F_N$ as fixed. For more information about maximum pseudo-likelihood estimators and other inferential approaches for survey data, see Kish and Frankel (1974); Godambe and Thompson (1986); Pfeffermann (1993), Korn and Graubard (1999, chapter 3), Chambers and Skinner (2003, chapter 2), and Fuller (2009, section 6.5). Maximum pseudo-likelihood estimators and their properties for Cox’s proportional hazards model for survey data are discussed in Binder (1990, 1992); Lin and Wei (1989); Lin (2000); Boudreau and Lawless (2006).

Without loss of generality, the rest of this section uses indices for stratified clustered designs. For a stratified clustered sample design, observations are represented by a matrix

$$(w, t, \Delta, Z) = (w_{hij}, t_{hij}, \Delta_{hij}, z_{hij})$$

where

- $w$ denotes the vector of sampling weights
- $t$ denotes the event time variable
- $\Delta$ denotes the event indicator
- $Z$ denotes the $n \times p$ matrix of auxiliary information
- $h = 1, 2, \ldots, H$ is the stratum index
- $i = 1, 2, \ldots, n_h$ is the cluster index within stratum $h$
- $j = 1, 2, \ldots, m_{hi}$ is the unit index within cluster $i$ of stratum $h$
- $p$ is the total number of parameters
- $n = \sum_{h=1}^{H} \sum_{i=1}^{n_h} m_{hi}$ is the total number of observations in the sample
- $y_{hij}(t) = I(t_{hij} \geq t)$, where $I(\cdot)$ is an indicator function
- $n_{hij}(t) = I(t_{hij} \leq t)$, where $I(\cdot)$ is an indicator function
Let $\sum_B = \sum_{(h,i,j) \in B}$ denote the summation over the set of indices such that the observation unit $j$ in PSU $i$ and stratum $h$ belongs to the index set $B$. Typically, $B$ is the set of all population indices that are in the sample, the risk set, or the set of all units with a failure.

The first-stage sampling rate (fraction of PSUs selected for the sample) is denoted by $f_h$. The first-stage sampling rate is used in Taylor series variance estimation. You can specify the stratum sampling rates with the RATE= option. Or if you specify population totals with the TOTAL= option, PROC SURVEYFREQ computes $f_h$ as the ratio of stratum sample size to the stratum total, in terms of PSUs. See the section “Population Totals and Sampling Rates” on page 8276 for details. If you do not specify the RATE= option or the TOTAL= option, then the procedure assumes that the stratum sampling rates $f_h$ are negligible and does not use a finite population correction when computing variances.

**Failure Time Distribution**

Let $T$ be a nonnegative random variable that represents the failure time of an individual from a homogeneous superpopulation. 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 $\lambda(t)$ specifies the instantaneous failure rate at $t$. If $T$ is a continuous random variable, $\lambda(t)$ is expressed as

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

where $f(t)$ is the probability density function of $T$.

**Time and CLASS Variables Usage**

The following DATA step creates an artificial data set, Test, to be used in this section. There are six variables in Test: the variable T contains the failure times; the variable Status is the censoring indicator variable with the value 1 for an uncensored failure time and the value 0 for a censored time; the variable A is a categorical variable with values 1, 2, and 3 representing three different categories; the variable MirrorT is an exact copy of T; the variable W is the observation weight; and the variable S is the strata indicator.

```plaintext
data Test;
  input T Status A W S @@;
  MirrorT = T;
datalines;
23 1 1 10 1 7 0 1 20 2
23 1 1 10 1 10 1 1 20 2
20 0 1 10 1 13 0 1 20 2
24 1 1 10 1 10 1 1 20 2
18 1 2 10 1 6 1 2 20 2
18 0 2 10 1 6 1 2 20 2
13 0 2 10 1 13 1 2 20 2
9 0 2 10 1 15 1 2 20 2
8 1 3 10 1 6 1 3 20 2
```
Time and CLASS Variables Usage

The time variable cannot be used explicitly as an explanatory effect in the MODEL statement. The following statements produce an error message:

```
proc surveyphreg data=Test;
  weight W;
  strata S;
  class A;
  model T*Status(0)=T*A;
run;
```

To use the time variable as an explanatory effect, replace T by MirrorT as an effect, which is an exact copy of T, as in the following statements:

```
proc surveyphreg data=Test;
  weight W;
  strata S;
  class A;
  model T*Status(0)=A*MirrorT;
run;
```

Note that neither T*A nor MirrorT*A in the MODEL statement is time-dependent. The results of fitting this model are shown in Figure 100.3.

**Figure 100.3**  T*A Effect

The SURVEYPHREG Procedure

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>MirrorT*A 1</td>
<td>30</td>
<td>-17.560699</td>
<td>0.337239</td>
<td>-52.07</td>
<td>&lt;.0001</td>
<td>0.000</td>
</tr>
<tr>
<td>MirrorT*A 2</td>
<td>30</td>
<td>-17.424235</td>
<td>0.331186</td>
<td>-52.61</td>
<td>&lt;.0001</td>
<td>0.000</td>
</tr>
<tr>
<td>MirrorT*A 3</td>
<td>30</td>
<td>-17.448672</td>
<td>0.290159</td>
<td>-60.13</td>
<td>&lt;.0001</td>
<td>0.000</td>
</tr>
</tbody>
</table>

CLASS Variables and Programming Statements

In PROC SURVEYPHREG, the levels of CLASS variables are determined by the CLASS statement and the input data and are not affected by user-supplied programming statements. Consider the following statements, which produce the results in Figure 100.4. Variable A is declared as a CLASS variable in the CLASS statement.
proc surveyphreg data=Test;
    weight W;
    strata S;
    class A;
    model T*Status(0)=A;
run;

Figure 100.4 shows the parameters that correspond to A and their respective regression coefficients estimates.

Figure 100.4
Design Variable and Regression Coefficient Estimates

The SURVEYPHREG Procedure

| Parameter | DF | Estimate | Standard Error | t Value | Pr > | Hazard | Ratio |
|-----------|----|----------|----------------|---------|------|--------|
| A 1       | 30 | -1.162184| 0.655136       | -1.77   | 0.0862 | 0.313  |
| A 2       | 30 | -0.616962| 0.521841       | -1.18   | 0.2464 | 0.540  |
| A 3       | 30 | 0        | .              | .       | .     | 1.000  |

Now consider the programming statement that attempts to change the value of the CLASS variable A as in the following specification:

proc surveyphreg data=Test;
    weight W;
    strata S;
    class A;
    model T*Status(0)=A;
    if A=3 then A=2;
run;

Results of this analysis are shown in Figure 100.5 and are identical to those in Figure 100.4. The `if A=3 then A=2` programming statement has no effect on the explanatory variable for A, which have already been determined.

Figure 100.5
Design Variable and Regression Coefficient Estimates

The SURVEYPHREG Procedure

<table>
<thead>
<tr>
<th>Class Level Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class Levels Values</td>
</tr>
<tr>
<td>A 3 1 2 3</td>
</tr>
</tbody>
</table>
Additionally any variable used in a programming statement that has already been declared in the CLASS statement is not treated as a collection of the corresponding design variables. Consider the following statements:

```
proc surveyphreg data=Test;
  class A;
  model T*Status(0)=A X;
  X=T*A;
run;
```

The CLASS variable `A` generates two design variables as explanatory variables. The variable `X` created by the `X=T*A` programming statement is a single time-dependent covariate whose values are evaluated using the exact values of `A` given in the data, not the dummy coded values that represent `A`. In the data set `Test`, `A` has the values of 1, 2, and 3, and these values are multiplied by the values of `T` to produce `X`. If `A` were a character variable with values ‘Bird’, ‘Cat’, and ‘Dog’, the programming statement `X=T*A` would have produced an error in the attempt to multiply a number with a character value.

```
proc surveyphreg data=Test;
  class A;
  model T*Status(0)=A X1 X2;
  X1 = T*(A=1);
  X2 = T*(A=2);
run;
```

The following statements are not the same as in the preceding program. If you want to create time-dependent covariates from the values of a CLASS variable, you could use syntax like the following:

```
proc surveyphreg data=Test;
  class A;
  model T*Status(0)=A X1 X2;
  X1 = T*(A=1);
  X2 = T*(A=2);
run;
```

The Boolean parenthetical expressions `(A=1)` and `(A=2)` resolve to a value of 1 or 0, depending on whether the expression is true or false, respectively.

Results of this test are shown in Figure 100.7.
In general, when your model contains a categorical explanatory variable that is time-dependent, it might be necessary to use hardcoded dummy variables to represent the categories of the categorical variable.

### Partial Likelihood Function for the Cox Model

Let $t_{(1)} < t_{(2)} < \ldots < t_{(K)}$ denote the $K$ distinct, ordered event times. Let $d_k$ denote the multiplicity of failures at $t_{(k)}$; that is, $d_k$ is the size of the set $D_k$ of individuals that fail at $t_{(k)}$. Let $w_{hij}$ be the weight associated with the $j$th observation unit in the $i$th cluster in stratum $h$. Using this notation, the pseudo-likelihood functions used in PROC SURVEYPHREG to estimate $\beta_N$ are described in the following sections.

#### Continuous Time Scale

Let $\mathcal{R}_k$ denote the risk set just before the $k$th ordered event time $t_{(k)}$.

The Breslow likelihood is expressed as

$$ L_{\text{Breslow}}(\beta) = \prod_{k=1}^{K} \exp \left( \beta' \sum_{D_k} w_{hij} Z_{hij}(t) \right) / \sum_{\mathcal{R}_k} w_{hij} \exp(\beta' Z_{hij}(t)) $$

The Efron likelihood is expressed as

$$ L_{\text{Efron}}(\beta) = \prod_{k=1}^{K} \exp \left( \beta' \sum_{D_k} w_{hij} Z_{hij}(t) \right) \left\{ \phi(\beta, Z, w, k) \right\}^{d_k} / \sum_{D_k} w_{hij} $$

where $\phi(\beta, Z, w, k)$ is

$$ \phi(\beta, Z, w, k) = \prod_{l=1}^{d_k} \left( \sum_{\mathcal{R}_k} w_{hij} \exp(\beta' Z_{hij}(t)) \right) - \frac{1}{d_k} \sum_{D_k} w_{hij} \exp(\beta' Z_{hij}(t)) $$
Specifying the Sample Design

PROC SURVEYPHREG produces statistics that are based on the sample design used to obtain the survey data. PROC SURVEYPHREG can be used for single-stage or multistage designs, with or without stratification, and with or without unequal weighting. To analyze your survey data with PROC SURVEYPHREG, you need to provide sample design information for the procedure. This information can include design (or variance) strata, clusters, and sampling weights. You provide sample design information with the STRATA, CLUSTER, and WEIGHT statements, and with the RATE= or TOTAL= option in the PROC SURVEYPHREG statement.

If you provide replicate weights for BRR or jackknife variance estimation, you do not need to specify a STRATA or CLUSTER statement. Otherwise, you should specify STRATA and CLUSTER statements whenever your design includes stratification and clustering.

When there are clusters (PSUs) in the sample design, the procedure estimates variance by using the PSUs, as described in the section “Variance Estimation” on page 8279. For a multistage sample design, PROC SURVEYPHREG uses only the first stage of the sample design for variance estimation. So, the required input includes only first-stage cluster (PSU) and first-stage stratum identification. You do not need to input design information about any additional stages of sampling.

Stratification

If your sample design is stratified at the first stage of sampling, use the STRATA statement to name the variables that form the strata. The combinations of categories of STRATA variables define the strata in the sample, where strata are nonoverlapping subgroups that were sampled independently. If your sample design has stratification at multiple stages, you should identify only the first-stage strata in the STRATA statement.

If you use a REPWEIGHTS statement to provide replicate weights for BRR or jackknife variance estimation, you do not need to specify a STRATA statement. Otherwise, you should specify a STRATA statement whenever your design includes stratification. If you do not specify a STRATA statement or a REPWEIGHTS statement, then PROC SURVEYPHREG assumes there is no stratification at the first stage. In other words, in this case, the procedure assumes that all observation units are in the same stratum.

Clustering

If your sample design selects clusters at the first stage of sampling, use the CLUSTER statement to name the variables that identify the first-stage clusters, which are also called primary sampling units (PSUs). The combinations of categories of CLUSTER variables define the clusters in the sample. If there is a STRATA statement, clusters are nested within strata. If your sample design has clustering at multiple stages, you should specify only the first-stage clusters (PSUs) in the CLUSTER statement. PROC SURVEYPHREG assumes that each cluster that is defined by the CLUSTER statement variables represents a PSU in the sample.

If you use a REPWEIGHTS statement to provide replicate weights for BRR or jackknife variance estimation, you do not need to specify a CLUSTER statement. Otherwise, you should specify a CLUSTER statement whenever your design includes clustering at the first stage of sampling. If you do not specify a CLUSTER statement, then PROC SURVEYPHREG treats each observation as a PSU.
Weighting

If your sample design includes unequal weighting, use the WEIGHT statement to name the variable that contains the sampling weights. Sampling weights must be positive numbers. If an observation has a weight that is nonpositive or missing, then the procedure omits that observation from the analysis. See the section “Missing Values” on page 8277 for more information.

If you do not specify a WEIGHT statement but include a REPWEIGHTS statement, PROC SURVEYPHREG uses the average of each observation’s replicate weights as the observation’s weight. If you specify neither a WEIGHT statement nor a REPWEIGHTS statement, PROC SURVEYPHREG assumes all observations have a weight of one.

Population Totals and Sampling Rates

To include a finite population correction (fpc) in Taylor series variance estimation, you can input either the sampling rate or the population total by using the RATE= or TOTAL= option in the PROC SURVEYPHREG statement. You cannot specify both of these options in the same PROC SURVEYPHREG statement. The RATE= and TOTAL= options apply only to Taylor series variance estimation. The procedure does not use a finite population correction for BRR or jackknife variance estimation.

If you do not specify the RATE= or TOTAL= option, the Taylor series variance estimation does not include a finite population correction. For fairly small sampling fractions, this correction is often ignored. See Cochran (1977) and Kish (1965) for more information.

If your design has multiple stages of selection and you are specifying the RATE= option, you should input the first-stage sampling rate, which is the ratio of the number of PSUs in the sample to the total number of PSUs in the study population. If you are specifying the TOTAL= option for a multistage design, you should input the total number of PSUs in the study population.

For a nonstratified sample design, or for a stratified sample design with the same sampling rate or the same population total in all strata, you can use the RATE=value or TOTAL=value option. If your sample design is stratified with different sampling rates or population totals in different strata, use the RATE=SAS-data-set or TOTAL=SAS-data-set option to name a SAS data set that contains the stratum sampling rates or totals. This data set is called a secondary data set, as opposed to the primary data set that you specify with the DATA= option.

The secondary data set must contain all the stratification variables listed in the STRATA statement and all the variables in the BY statement. Furthermore, the BY groups must appear in the same order as in the primary data set. If there are formats associated with the STRATA variables and the BY variables, then the formats must be consistent in the primary and the secondary data sets. If you specify the TOTAL=SAS-data-set option, the secondary data set must have a variable named _TOTAL_ that contains the stratum population totals. If you specify the RATE=SAS-data-set option, the secondary data set must have a variable named _RATE_ that contains the stratum sampling rates. If the secondary data set contains more than one observation for any one stratum, the procedure uses the first value of _TOTAL_ or _RATE_ for that stratum and ignores the rest.

The value in the RATE= option or the values of _RATE_ in the secondary data set must be nonnegative numbers. You can specify value as a number between 0 and 1. Or you can specify value in percentage form as a number between 1 and 100, and PROC SURVEYPHREG converts that number to a proportion. The procedure treats the value 1 as 100% instead of 1%.
If you specify the TOTAL=\textit{value} option, \textit{value} must not be less than the sample size. If you provide stratum population totals in a secondary data set, these values must not be less than the corresponding stratum sample sizes.

**Missing Values**

Missing values in your survey data can compromise the quality of your survey results. Some missing values for survey data are because of nonresponses. An observation whose response to every survey item is available is called a \textit{complete respondent}, and an observation whose response to one or more survey items are missing is called an \textit{incomplete respondent}. If the complete respondents are different from the incomplete respondents with regard to a survey effect or outcome, then survey estimates will be biased and will not accurately represent the survey population. There are a variety of techniques in sample design and survey operations that can reduce nonresponse. After data collection is complete, you can use imputation to replace missing values with acceptable values, and you can use sampling weight adjustments to compensate for nonresponse. You should complete this data preparation and adjustment before you analyze your data with \texttt{PROC SURVEYPHREG}. For more details, see Cochran (1977); Kalton and Kasprzyk (1986); Brick and Kalton (1996).

**WEIGHT Variable**

If an observation has a missing value or a nonpositive value for the \texttt{WEIGHT} variable, then \texttt{PROC SURVEYPHREG} excludes that observation from the analysis.

**REPWEIGHTS Variables**

If you provide replicate weights with a \texttt{REPWEIGHTS} statement for BRR or jackknife variance estimation, all \texttt{REPWEIGHTS} variable values must be nonmissing. Similarly, if you provide jackknife coefficients with the \texttt{JKCOEFS=} option in the \texttt{REPWEIGHTS} statement, all values of the \texttt{JKCoefficient} variable must be nonmissing. The procedure does not perform the analysis when any replicate weight or jackknife coefficient value is missing.

**CLASS, STRATA, CLUSTER, and DOMAIN Variables**

An observation is excluded from the analysis if it has a missing value for any \texttt{CLASS}, \texttt{STRATA}, \texttt{CLUSTER}, or \texttt{DOMAIN} variable, unless you specify the \texttt{MISSING} option in the \texttt{PROC SURVEYPHREG} statement. If you specify the \texttt{MISSING} option, the procedure treats missing values as a valid (nonmissing) category for all categorical variables, which include \texttt{STRATA} variables, \texttt{CLUSTER} variables, \texttt{CLASS} variables, and \texttt{DOMAIN} variables.

**Analysis Variables**

By default, \texttt{PROC SURVEYPHREG} excludes an observation from the likelihood estimation and all associated analyses if the observation has a missing value for any of the variables in the \texttt{MODEL} statement, unless you specify the \texttt{MISSING} or \texttt{NOMCAR} option in the \texttt{PROC SURVEYPHREG} statement. When the procedure excludes observations with missing values from analyses, it displays the total frequency of observations used in the \texttt{NObs} table.
If you specify time-dependent covariates by using programming statements, the procedure computes the values of the covariates for all observations in the risk set at every event time. If an observation contains missing values for any of the time-dependent covariates at a given event time, then the observation is not used at that event time. However, that same observation can be used at some other event times where it contains no missing values. Therefore, an observation with missing time-dependent covariates can be used at some event times but ignored at other event times, depending on whether any of the corresponding time-dependent covariates are missing.

If you specify the MISSING option, the procedure treats missing levels as a valid (nonmissing) level for each categorical analysis variable.

If you specify the NOMCAR option for Taylor series variance estimation, the procedure includes observations with missing values of analysis variables in the variance computations.

The NOMCAR Option

When you specify the NOMCAR option, PROC SURVEYPHREG computes variance estimates by analyzing the nonmissing values for variables in the regression model as a domain or subpopulation, where the entire population includes both nonmissing and missing domains. By default, if an observation contains missing values for the dependent variable or for any variable used in the independent effects, the observation is excluded from the analysis. See the section “Missing Values” on page 8277 for more information.

Note that the NOMCAR option has no effect on categorical predictors when you specify the MISSING option, which treats missing values as a valid nonmissing level. The NOMCAR option does not affect the inclusion of observations with missing values of the WEIGHT, FREQ, CLUSTER, STRATA, or DOMAIN variables. Observations with missing values of the WEIGHT and FREQ variables are always excluded from the analysis. Observations with missing values of the CLUSTER, DOMAIN, or STRATA variables are excluded unless you specify the MISSING option.

The NOMCAR option applies only to Taylor series variance estimation. The replication methods, which you request with the VARMETHOD=BRR and VARMETHOD=JACKKNIFE options, do not use the NOMCAR option.

Degrees of Freedom

PROC SURVEYPHREG computes degrees of freedom to compute confidence limits and $F$ statistics. The degrees of freedom computation depends on the variance estimation method that you request. See the section “Degrees of Freedom” on page 8284 for details. Missing values can affect the degrees of freedom computation.

Taylor Series Variance Estimation

The degrees of freedom can depend on the number of clusters, the number of strata, and the number of observations. For Taylor series variance estimation, these numbers are based on the observations included in the analysis. These numbers do not count observations that are excluded from the analysis due to missing values. If all values in a stratum are excluded from the analysis as missing values, then that stratum is called an empty stratum. Empty strata are not counted in the total number of strata for the analysis. Similarly, empty clusters and missing observations are not included in the totals counts of clusters and observations that are used to compute the degrees of freedom for the analysis.

If you specify the MISSING option, missing values are treated as valid nonmissing levels and are included in computing degrees of freedom. If you specify the NOMCAR option for Taylor series variance estimation,
observations with missing values for variables in the regression model are included in computing degrees of freedom.

**Replicate-Based Variance Estimation**

For BRR or jackknife variance estimation, by default PROC SURVEYPHREG computes the degrees of freedom by using all valid observations in the input data set. A valid observation is an observation that has a positive value of the WEIGHT variable and nonmissing values of the STRATA and CLUSTER variables unless you specify the MISSING option.

---

**Variance Estimation**

PROC SURVEYPHREG uses the Taylor series method or replication (resampling) methods to estimate sampling errors of estimators that are based on complex sample designs (Fuller 1975; Särndal, Swensson, and Wretman 1992; Wolter 2007; Rust 1985; Dippo, Fay, and Morganstein 1984; Rao and Shao 1999, 1996; and Binder 1992). You can use the VARMETHOD= option in the PROC statement to specify the variance estimation method. By default, PROC SURVEYPHREG uses the Taylor series method.

However, replication methods have recently gained popularity for estimating variances in complex survey data analysis. One reason for this popularity is the relative simplicity of replication-based estimates, especially for nonlinear estimators; another is that modern computational capacity has made replication methods feasible for practical survey analysis.

Replication methods draw multiple replicates (also called subsamples) from a full sample according to a specific resampling scheme. The most commonly used resampling schemes are the balanced repeated replication (BRR) method and the jackknife method. For each replicate, the original weights are modified for the PSUs in the replicates to create replicate weights. The parameters of interest are estimated by using the replicate weights for each replicate. Then the variances of parameters of interest are estimated by the variability among the estimates derived from these replicates. The procedure automatically creates replicate weights based on the replication method you specify; alternatively you can use the REPWEIGHTS statement to provide your own replicate weights for variance estimation.

The following sections provide details about how the variance-covariance matrix of the estimated regression coefficients is estimated for each variance estimation method.

**Taylor Series Linearization**

The Taylor series linearization method is the default variance estimation method used by PROC SURVEYPHREG. See the section “Notation and Estimation” on page 8268 for definitions of the notation used in this section. Let

\[
S^{(r)}(\beta, t) = \sum_{A} w_{hij} y_{hij}(t) \exp(\beta' Z_{hij}(t)) Z_{hij}(t)^{\otimes r}
\]

where \( r = 0, 1 \). Let \( A \) be the set of indices in the selected sample. Let

\[
a^{\otimes r} = \begin{cases} aa' & r = 1 \\ I_{\text{dim}(a)} & r = 0 \end{cases}
\]

and let \( I_{\text{dim}(a)} \) be the identity matrix of appropriate dimension.
Chapter 100: The SURVEYPHREG Procedure

Let $\tilde{Z}(\beta, t) = \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)}$. The score residual for the $(h, i, j)$ subject is

$$L_{hij}(\beta) = \Delta_{hij}\left\{Z_{hij}(t_{hij}) - \tilde{Z}(\beta, t_{hij})\right\}$$

$$- \sum_{(h, i, j) \in A} \Delta_{hij} \frac{w_{hij}Y_{hij}(t_{hij})}{S^{(0)}(\beta, t_{hij})} \exp\left(\beta'Z_{hij}(t_{hij})\right) \left\{Z_{hij}(t_{hij}) - \tilde{Z}(\beta, t_{hij})\right\}$$

For TIES=EFRON, the computation of the score residuals is modified to comply with the Efron partial likelihood. See the section “Residuals” on page 8289 for more information.

The Taylor series estimate of the covariance matrix of $\hat{\beta}$ is

$$\hat{V}(\hat{\beta}) = \mathcal{I}^{-1}(\hat{\beta}) G \mathcal{I}^{-1}(\hat{\beta})$$

where $\mathcal{I}(\hat{\beta})$ is the observed information matrix and the $p \times p$ matrix $G$ is defined as

$$G = \frac{n - 1}{n - p} \sum_{h=1}^{H} n_h (1 - f_h) \sum_{i=1}^{n_h} (e_{hi+} - \bar{e}_{h-})^t (e_{hi+} - \bar{e}_{h-})$$

The observed residuals, their sums and means are defined as follows:

$$e_{hij} = w_{hij}L_{hij}(\hat{\beta})$$

$$e_{hi+} = \sum_{j=1}^{m_{hi}} e_{hij}$$

$$\bar{e}_{h-} = \frac{1}{n_h} \sum_{i=1}^{n_h} e_{hi+}$$

The factor $(n - 1)/(n - p)$ in the computation of the matrix $G$ reduces the small sample bias that is associated with using the estimated function to calculate deviations (Fuller et al. (1989), pp. 77–81). For simple random sampling, this factor contributes to the degrees of freedom correction applied to the residual mean square for ordinary least squares in which $p$ parameters are estimated. By default, the procedure uses this adjustment in the variance estimation. If you do not want to use this multiplier in the variance estimator, then specify the VADJUST=NONE option in the MODEL statement.

**Balanced Repeated Replication (BRR) Method**

The balanced repeated replication (BRR) method requires that the full sample be drawn by using a stratified sample design with two primary sampling units (PSUs) per stratum. The BRR method constructs half-sample replicates by deleting one PSU per stratum according to a Hadamard matrix and doubling the original weight of the other PSU in that stratum. Let $H$ be the total number of strata. The total number of replicates $R$ is the smallest multiple of 4 that is greater than $H$. However, if you prefer a larger number of replicates, you can specify the REPS=$n$ method-option. If a $n \times n$ Hadamard matrix cannot be constructed, the number of replicates is increased until a Hadamard matrix becomes available.

Each replicate is obtained by deleting one PSU per stratum according to a corresponding Hadamard matrix and adjusting the original weights for the remaining PSUs. The new weights are called replicate weights.
Replicates are constructed by using the first $H$ columns of the $R \times R$ Hadamard matrix. The $r$th ($r = 1, 2, \ldots, R$) replicate is drawn from the full sample according to the $r$th row of the Hadamard matrix as follows:

- If the $(r, h)$ element of the Hadamard matrix is 1, then the first PSU of stratum $h$ is included in the $r$th replicate and the second PSU of stratum $h$ is excluded.
- If the $(r, h)$ element of the Hadamard matrix is $-1$, then the second PSU of stratum $h$ is included in the $r$th replicate and the first PSU of stratum $h$ is excluded.

The replicate weights of the remaining PSUs in each half sample are then doubled to their original weights. For more detail about the BRR method, see Wolter (2007) and Lohr (2010).

By default, an appropriate Hadamard matrix is generated automatically to create the replicates. You can display the Hadamard matrix by specifying the VARMETHOD=BRR(PRINTH) method-option. If you provide a Hadamard matrix by specifying the VARMETHOD=BRR(HADAMARD=) method-option, then the replicates are generated according to the provided Hadamard matrix. You can use the VARMETHOD=BRR(OUTWEIGHTS=) method-option to store the replicate weights in a SAS data set.

Let $\hat{\beta}$ be the estimated proportional hazards regression coefficients from the full sample, and let $\hat{\beta}_r$ be the estimated proportional hazards regression coefficients from the $r$th replicate by using replicate weights. PROC SURVEYPHREG estimates the covariance matrix of $\hat{\beta}$ by

$$\hat{V}(\hat{\beta}) = \frac{1}{R} \sum_{r=1}^{R} \left( \hat{\beta}_r - \bar{\beta} \right) \left( \hat{\beta}_r - \bar{\beta} \right)'$$

with $H$ degrees of freedom, where $H$ is the number of strata.

If you specify the CENTER=REPLICATES method-option, then PROC SURVEYPHREG computes the covariance matrix of $\hat{\beta}$ by

$$\hat{V}(\hat{\beta}) = \frac{1}{R} \sum_{r=1}^{R} \left( \hat{\beta}_r - \bar{\beta}_r \right) \left( \hat{\beta}_r - \bar{\beta}_r \right)'$$

where $\bar{\beta}_r$ is the average of the replicate estimates as follows:

$$\bar{\beta}_r = \frac{1}{R} \sum_{r=1}^{R} \hat{\beta}_r$$

If one or more components of $\hat{\beta}_r$ cannot be calculated for some replicates, then the variance estimate is computed by using only the replicates for which the proportional hazards regression coefficients can be estimated. Estimability and nonconvergence are the two most common reasons why $\hat{\beta}_r$ might not be available for a replicate sample even if $\hat{\beta}$ is defined for the full sample. Let $R_a$ be the number of replicates where $\hat{\beta}_r$ is available, and let $R - R_a$ be the number of replicates where $\hat{\beta}_r$ is not available. Without loss of generality, assume that $\hat{\beta}_r$ is available only for the first $R_a$ replicates; then the BRR variance estimator is

$$\hat{V}(\hat{\beta}) = \frac{1}{R_a} \sum_{r=1}^{R_a} \left( \hat{\beta}_r - \bar{\beta} \right) \left( \hat{\beta}_r - \bar{\beta} \right)'$$

with degrees of freedom equal to the minimum of $H$ and $R_a$, where $H$ is the number of strata. Alternatively, you can use the FAY= method-option to request Fay’s BRR method, as discussed in the following section.
**Fay’s BRR Method**

The traditional BRR method constructs half-sample replicates by deleting one PSU per stratum according to a Hadamard matrix and doubling the original weight of the other PSU. Fay’s BRR method uses the Fay coefficient, \( \epsilon \) (0 \( \leq \epsilon < 1 \)), and instead of deleting one PSU per stratum, it multiplies the original weight by the coefficient \( \epsilon \). The original weight of the remaining PSU in that stratum is multiplied by 2 \( - \epsilon \). PROC SURVEYPHREG uses \( \epsilon = 0.5 \) as the default value; alternatively, you can specify a value for \( \epsilon \) with the **FAY=** method-option. When \( \epsilon = 0 \), Fay’s method becomes the traditional BRR method. For more details, see Dippo, Fay, and Morganstein (1984); Fay (1984, 1989); Judkins (1990). Because the traditional BRR method uses only half of the total sample in every replicate, several replicate estimators (\( \hat{\beta}_r \)) might be undefined even when the full sample estimator (\( \hat{\beta} \)) is defined. Fay’s BRR method is especially useful for this situation because it uses all the sampled units in every replicate.

Let \( \hat{\beta} \) be the estimated proportional hazards regression coefficients from the full sample, and let \( \hat{\beta}_r \) be the estimated regression coefficients that are obtained from the \( r \)th replicate by using replicate weights. PROC SURVEYPHREG estimates the covariance matrix of \( \hat{\beta} \) by

\[
\hat{V}(\hat{\beta}) = \frac{1}{R(1-\epsilon)^2} \sum_{r=1}^{R} (\hat{\beta}_r - \hat{\beta})(\hat{\beta}_r - \hat{\beta})'
\]

with \( H \) degrees of freedom, where \( H \) is the number of strata.

**Hadamard Matrix**

PROC SURVEYPHREG uses a Hadamard matrix to construct replicates for BRR variance estimation. You can provide a Hadamard matrix for replicate construction by using the **HADAMARD=** method-option for **VARMETHOD=BRR**. Otherwise, PROC SURVEYPHREG generates an appropriate Hadamard matrix. You can display the Hadamard matrix by specifying the **PRINTH** method-option.

A Hadamard matrix \( A \) of dimension \( R \) is a square matrix that has all elements equal to 1 or –1 such that \( A'A = RL \), where \( I \) is an identity matrix of appropriate order. The dimension of a Hadamard matrix must equal 1, 2, or a multiple of 4.

For example, the following matrix is a Hadamard matrix of dimension \( k = 8 \):

\[
\begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\
1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\
1 & 1 & 1 & -1 & -1 & 1 & -1 & 1 \\
1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 \\
1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 \\
1 & -1 & -1 & 1 & 1 & 1 & 1 & -1
\end{bmatrix}
\]

For BRR replicate construction, the dimension of the Hadamard matrix must be at least \( H \), where \( H \) denotes the number of first-stage strata in your design. If a Hadamard matrix of a given dimension exists, it is not necessarily unique. Therefore, if you want to use a specific Hadamard matrix, you must provide the matrix as a SAS data set in the **HADAMARD=** method-option. You must ensure that the matrix that you provide is actually a Hadamard matrix; PROC SURVEYPHREG does not check the validity of your Hadamard matrix.

See the section “**Balanced Repeated Replication (BRR) Method**” on page 8280 for details about how the Hadamard matrix is used to construct replicates for BRR variance estimation.
Jackknife Method

The jackknife method of variance estimation deletes one PSU at a time from the full sample to create replicates. This method is also known as the delete-1 jackknife method because it deletes exactly one PSU in every replicate. The total number of replicates $R$ is the same as the total number of PSUs. In each replicate, the sampling weights of the remaining PSUs are modified by the jackknife coefficient $\alpha_r$. The modified weights are called replicate weights.

Let PSU $i$ in stratum $h_r$ be omitted for the $r$th replicate; then the jackknife coefficient and replicate weights are computed as

$$\alpha_r = \begin{cases} \frac{n_{hr} - 1}{n_{hr}} & \text{for a stratified design} \\ \frac{R-1}{R} & \text{for designs without stratification} \end{cases}$$

and

$$w_{hij}^{(r)} = \begin{cases} w_{hij} & \text{if observation unit } j \text{ is not in donor stratum } h_r \\ 0 & \text{if observation unit } j \text{ is in PSU } i \text{ of donor stratum } h_r \\ w_{hij}/\alpha_r & \text{if observation unit } j \text{ is not in PSU } i \text{ but in donor stratum } h_r \end{cases}$$

You can use the `VARMETHOD=JACKKNIFE(OUTJKCOEFS=)` method-option to store the jackknife coefficients in a SAS data set and use the `VARMETHOD=JACKKNIFE(OUTWEIGHTS=)` method-option to store the replicate weights in a SAS data set.

If you provide your own replicate weights with a `REPWEIGHTS` statement, then you can also provide corresponding jackknife coefficients with the `JKCOEFS=` option. If you provide replicate weights with a `REPWEIGHTS` statement but do not provide jackknife coefficients, then the procedure uses $(R - 1)/R$ as the default jackknife coefficient for every replicate, where $R$ is the total number of replicates.

Let $\hat{\beta}$ be the estimated proportional hazards regression coefficients from the full sample, and let $\hat{\beta}_r$ be the estimated regression coefficients for the $r$th replicate. PROC SURVEYPHREG estimates the covariance matrix of $\hat{\beta}$ by

$$\hat{V}(\hat{\beta}) = \sum_{r=1}^{R} \alpha_r \left( \hat{\beta}_r - \overline{\hat{\beta}} \right) \left( \hat{\beta}_r - \overline{\hat{\beta}} \right)'$$

with $R - H$ degrees of freedom, where $R$ is the number of replicates and $H$ is the number of strata, or $R - 1$ when there is no stratification.

If you specify the `CENTER=REPLICATES` method-option, then PROC SURVEYPHREG computes the covariance matrix of $\hat{\beta}$ by

$$\hat{V}(\hat{\beta}) = \sum_{r=1}^{R} \alpha_r \left( \hat{\beta}_r - \overline{\hat{\beta}}_r \right) \left( \hat{\beta}_r - \overline{\hat{\beta}}_r \right)'$$

where $\overline{\hat{\beta}}_r$ is the average of the replicate estimates as follows:

$$\overline{\hat{\beta}}_r = \frac{1}{R} \sum_{r=1}^{R} \hat{\beta}_r$$
Chapter 100: The SURVEYPHREG Procedure

If one or more components of $\hat{\beta}_r$ cannot be calculated for some replicates, then the variance estimator uses only the replicates for which the proportional hazards regression coefficients can be estimated. Estimability and nonconvergence are two common reasons why $\hat{\beta}_r$ might not be available for a replicate sample even if $\hat{\beta}$ is defined for the full sample. Let $R_a$ be the number of replicates where $\hat{\beta}_r$ are available, and let $R - R_a$ be the number of replicates where $\hat{\beta}_r$ are not available. Without loss of generality, assume that $\hat{\beta}_r$ is available only for the first $R_a$ replicates; then the jackknife variance estimator is

$$\hat{V}(\hat{\beta}) = \sum_{r=1}^{R_a} \alpha_r \left( \hat{\beta}_r - \hat{\beta} \right) \left( \hat{\beta}_r - \hat{\beta} \right)'$$

with $R_a - H$ degrees of freedom, where $H$ is the number of strata. Alternatively, you can use the VADJUST=AVGREPSS option in the MODEL statement to use the average sum of squares for the invalid replicate samples. See “Variance Adjustment Factors” on page 8285 for details.

**Degrees of Freedom**

PROC SURVEYPHREG uses the degrees of freedom of the variance estimator to obtain $t$ confidence limits and Wald type $F$ tests. PROC SURVEYPHREG computes the degrees of freedom based on the variance estimation method and the sample design. Alternatively, you can specify the degrees of freedom in the DF= option in the MODEL statement.

For Taylor series variance estimation, PROC SURVEYPHREG calculates the degrees of freedom ($df$) as the number of clusters minus the number of strata. If the CLUSTER statement is not specified, then the procedure treats each observation as a cluster. If the STRATA statement is not specified, then the procedure assumes that all observations are in the same stratum. These numbers are based on the observations included in the analysis. These numbers do not count observations that are excluded from the analysis due to missing values. See the section “Missing Values” on page 8277 for details. If you specify the MISSING option in the CLASS statement, missing values are treated as valid nonmissing levels and are included in computing degrees of freedom. If you specify the NOMCAR option for Taylor series variance estimation, observations with missing values of the analysis variables are included in computing the degrees of freedom.

If you provide replicate weights with a REPWEIGHTS statement, the degrees of freedom is equal the number of replicates used, which is the number of REPWEIGHTS variables that provide replicate estimates. Alternatively, you can specify the degrees of freedom in the DF= option in the REPWEIGHTS or MODEL statement.

For BRR variance estimation (when you do not use a REPWEIGHTS statement), PROC SURVEYPHREG calculates the degrees of freedom as the number of strata. PROC SURVEYPHREG bases the number of strata on all valid observations in the data set. If some replicate samples are not usable in the sense that these replicate samples cannot be used for parameter estimation (say, for nonconvergence or inestimability), then $df$ equals the minimum of the number of strata and the number of replicates used. Alternatively, you can use the DF=ALLREPS option in the MODEL statement to specify that $df$ equals the number of strata.

For jackknife variance estimation (when you do not use a REPWEIGHTS statement), PROC SURVEYPHREG calculates the degrees of freedom as the number of clusters minus the number of strata. If the CLUSTER statement is not specified, then the procedure treats each observation as a cluster. If the STRATA statement is not specified, then the procedure assumes that all observations are in the same stratum. For jackknife variance estimation, PROC SURVEYPHREG bases the number of strata and clusters on all valid observations in the data set. If some replicate samples are not usable in the sense that these replicate samples cannot be used for parameter estimation (say, for nonconvergence or inestimability), then
**Variance Estimation**

\( df \) equals the number of clusters (or observations if no CLUSTER statement is specified) minus the number of strata (or one in if no STRATA statement is specified) minus the number of replicate samples that are not used. Alternatively, you can use the \( \text{DF=ALLREPS} \) option in the \texttt{MODEL} statement to specify that \( df \) equals the number of clusters minus the number of strata.

**Variance Adjustment Factors**

PROC SURVEYPHREG provides options for adjustment of the default variance estimators. \texttt{VADJUST=NONE} and \texttt{VADJUST=DF} are available for the Taylor series linearization variance estimator. \texttt{VADJUST=AVGREPSS} is available for the jackknife replication variance estimators.

For models with large number of parameters, it is reasonable to adjust the Taylor series linearized variance estimator by the number of estimable parameters in the analysis model. Fuller et al. (1989, pp. 77–81) use an adjustment factor \( \frac{n}{n-1} \) to estimate the linearized variance for regression coefficients, where \( n \) is the total number of observation units and \( p \) is the number of estimable parameters in the analysis model. By default, PROC SURVEYPHREG uses this adjustment in the computation of the matrix \( G \) for the Taylor series linearization variance estimation. If you do not want to use this adjustment, then specify \texttt{VADJUST=NONE}.

Variance adjustment factors can be useful for replication variance estimations, especially if some replicate samples are not usable. A replicate sample might not provide useful parameter estimates (replicate estimates) for reasons such as nonconvergence of the optimization or inestimability of some parameters in that subsample. For example, consider the jackknife variance estimator with \( R \) replicates. Suppose that only \( R_a (< R) \) replicates are used to obtain replicate estimates and \( R - R_a \) replicates cannot be used due to, say, nonconvergence of the optimization. Without loss of generality, assume that the first \( R_a \) replicates are used. By default SURVEYPHREG uses

\[
\tilde{V}(\hat{\beta}) = \sum_{r=1}^{R_a} \alpha_r (\hat{\beta}_r - \hat{\beta})(\hat{\beta}_r - \hat{\beta})',
\]

as the jackknife variance estimator. An alternative estimator is

\[
\tilde{V}(\hat{\beta}) = \sum_{r=1}^{R_a} \alpha_r (\hat{\beta}_r - \hat{\beta})(\hat{\beta}_r - \hat{\beta})' + (R - R_a) \left\{ \frac{1}{R_a} \sum_{r=1}^{R_a} \alpha_r (\hat{\beta}_r - \hat{\beta})(\hat{\beta}_r - \hat{\beta})' \right\}
\]

\[
= \frac{R}{R_a} \sum_{r=1}^{R_a} \alpha_r (\hat{\beta}_r - \hat{\beta})(\hat{\beta}_r - \hat{\beta})'
\]

which uses the average replicate sum of squares for the \( R - R_a \) unusable replicate samples. If you specify the \texttt{VADJUST=AVGREPSS} option, PROC SURVEYPHREG uses the second variance estimator for the jackknife replication method. Note that you can specify the \texttt{FAY method-option} for the BRR method to avoid nonconvergence of the optimization or inestimability of some parameters in subsamples.

**Variance Ratios and Standard Error Ratios**

PROC SURVEYPHREG provides options to compute different variance ratios and standard error ratios.

If you specify the \texttt{VARRATIO=MODEL} option, then the procedure computes the variance ratio of the estimated regression parameter \( \hat{\beta}_j \) as \( \frac{\hat{V}(\hat{\beta}_j)}{\hat{V}_M(\hat{\beta}_j)} \), where \( \hat{V}(\hat{\beta}_j) \), the estimated variance of \( \hat{\beta}_j \), uses the complete design information and \( \hat{V}_M(\hat{\beta}_j) \) is the \( j \)th diagonal element of the observed information matrix \( I^{-1}(\hat{\beta}) \).
If you specify the VARRATIO=IND option, then the procedure computes the variance ratio of the estimated regression parameter $\hat{\beta}_j$ as $\frac{\hat{V}(\hat{\beta}_j)}{\hat{V}_{IND}(\hat{\beta}_j)}$, where $\hat{V}(\hat{\beta}_j)$, the estimated variance of $\hat{\beta}_j$, uses the complete design information and $\hat{V}_{IND}(\hat{\beta}_j)$ is the $j$th diagonal element of $\hat{V}_{IND}(\hat{\beta})$. $\hat{V}_{IND}(\hat{\beta})$ is the sandwich variance estimator, which ignores the strata and the clusters and is computed as

$$\hat{V}_{IND}(\hat{\beta}) = \mathcal{I}^{-1}(\hat{\beta}) \left\{ \frac{n}{n-1} (1 - f) \sum_h \sum_i \sum_j (e_{hij} - \bar{e}_h) (e_{hij} - \bar{e}_h) \right\} \mathcal{I}^{-1}(\hat{\beta})$$

where $e_{hij}$ are the weighted score residuals, $f$ is the overall sampling fraction, and $n$ is the number of observation units. The three sums are over the observation units ($j$) across the PSUs ($i$) and the strata ($h$).

For Taylor series variance estimation, PROC SURVEYPHREG determines the value of $f$, the overall sampling fraction, based on the RATE= or TOTAL= option. If you do not specify either of these options, PROC SURVEYPHREG assumes that the value of $f$ is negligible and does not use a finite population correction in the analysis. If you specify RATE=value, PROC SURVEYPHREG uses value as the overall sampling fraction $f$. If you specify TOTAL=value, PROC SURVEYPHREG computes $f$ as the ratio of the number of PSUs in the sample to the specified total.

If you specify stratum sampling rates by using the RATE=SAS-data-set option, then PROC SURVEYPHREG computes stratum totals based on these stratum sampling rates and the number of sample PSUs in each stratum. The procedure sums the stratum totals to form the overall total and then computes $f$ as the ratio of the number of sample PSUs to the overall total. Alternatively, if you specify stratum totals with the TOTAL=SAS-data-set option, then PROC SURVEYPHREG sums these totals to compute the overall total. The overall sampling fraction $f$ is then computed as the ratio of the number of sample PSUs to the overall total.

The replication methods do not use the finite population correction factor $(1 - f)$ in the denominator.

Standard error ratios are computed as the square root of the variance ratios.

**Domain Analysis**

Domain analysis refers to the computation of statistics for domains (subpopulations). Formation of subpopulations can be unrelated to the sample design, and so the domain sample sizes can actually be random variables. Domain analysis takes this variability into account to compute variance estimates for estimated model parameters. Domain analysis is also known as subgroup analysis, subpopulation analysis, and subdomain analysis. For more information about domain analysis, see Lohr (2010); Särndal, Swensson, and Wretman (1992); Cochran (1977).

To request domain analysis with PROC SURVEYPHREG, use the DOMAIN statement. If your domains are formed by more than one variable, you can specify DomainVariable_1 * DomainVariable_2 in the DOMAIN statement. If you use the DOMAIN statement, the procedure performs separate analyses for all domains, in addition to the overall analysis.

Including the domain variables in a DOMAIN statement request provides a different analysis from that obtained by using a BY statement, which provides completely separate analyses of the BY groups. The BY statement can also be used to analyze the data set by subgroups, but it is critical to note that this does not account for random sample sizes that often occur for domain analyses. The BY statement is appropriate only when the number of units in each subgroup is known with certainty. For example, the BY statement can be
used to obtain stratum level estimates when you have fixed sample sizes for the strata. When the subgroup sample size is random, include the domain variables in DOMAIN statement.

**Hypothesis Tests, Confidence Intervals, and Residuals**

**Testing the Global Null Hypothesis**

The following statistics can be used to test the global null hypothesis $H_0: \beta = 0$. Let $d$ be the number of clusters (or observations) minus the number of strata (or one) and $p$ be the number of estimable parameters in the analysis model.

The likelihood ratio test is expressed as

$$
\chi^2_{LR} = 2 \left[ \log \left\{ L(\hat{\beta}) \right\} - \log \left\{ L(0) \right\} \right]
$$

where $L(\cdot)$ denotes the partial pseudo-likelihood described in “Partial Likelihood Function for the Cox Model” on page 8274. The $p$-value is computed by using a chi-square distribution with $p$ degrees of freedom. The usual assumptions required for a likelihood ratio test do not hold for the pseudo-likelihood that is used by PROC SURVEYPHREG, leading to other methods for testing the global null hypothesis, such as the Wald test discussed below.

Wald’s test is expressed as

$$
W_F = \left( \frac{d - p + 1}{d} \right) \hat{\beta}' \left[ \sqrt{\hat{V}(\hat{\beta})} \right]^{-1} \hat{\beta}
$$

The $p$-value is computed by using an $F$ distribution with $(p, d)$ degrees of freedom. For the Taylor series linearization method, the DF=PARMADJ option in the MODEL statement computes the $p$-value by using an $F$ distribution with $(p, d - p + 1)$ degrees of freedom. The DF=q option in the MODEL statement replaces the $(d - p + 1)$ factor in the numerator of the Wald’s statistics by $q$ and computes the $p$-value by using an $F$ distribution with $(p, q)$ degrees of freedom.

If you specify the DF=NONE option in the MODEL statement, then the procedure computes

$$
W_{\chi^2} = \hat{\beta}' \left[ \sqrt{\hat{V}(\hat{\beta})} \right]^{-1} \hat{\beta}
$$

and the $p$-value is computed by using a chi-square distribution with $p$ degrees of freedom.

**Model Fit Statistics**

Suppose the model contains $p$ estimable parameters. Then the following two criteria are displayed for model fit statistics:

- $-2$ log likelihood:

$$
-2 \log L = -2 \log (L(\hat{\beta}))
$$

where $L(\cdot)$ is a partial pseudo-likelihood function for the corresponding TIES= option as described in the section “Partial Likelihood Function for the Cox Model” on page 8274, and $\hat{\beta}$ is the maximum pseudo-log-likelihood estimate of the proportional hazards regression coefficients.
• Akaike’s information criterion (AIC):

\[ \text{AIC} = -2 \log L + 2p \]

The AIC statistics gives a different way of adjusting the \(-2 \log \) likelihood statistic for the number of estimable parameters in the model.

**Contrasts**

For a testable hypothesis \( H_0: \mathbf{L} \beta = 0 \), the Wald \( F \) statistic is computed as

\[
F_{\text{Wald}} = \frac{\left( \mathbf{L}^* \hat{\beta} \right)^\top (\mathbf{L}^*\hat{\mathbf{V}}\mathbf{L}^*)^{-1} (\mathbf{L}^* \hat{\beta})}{\text{rank}(\mathbf{L})}
\]

where \( \mathbf{L} \) is a contrast vector or matrix that you specify, \( \beta \) is the vector of regression parameters, \( \hat{\beta} \) is the estimated regression coefficients, \( \hat{\mathbf{V}} \) is the estimated covariance matrix of \( \hat{\beta} \), \( \text{rank}(\mathbf{L}) \) is the rank of \( \mathbf{L} \), and \( \mathbf{L}^* \) is a matrix such that

• \( \mathbf{L}^* \) has the same number of columns as \( \mathbf{L} \)
• \( \mathbf{L}^* \) has full row rank
• the rank of \( \mathbf{L}^* \) equals the rank of the \( \mathbf{L} \) matrix
• all rows of \( \mathbf{L}^* \) are estimable functions
• the Wald \( F \) statistic computed by using the \( \mathbf{L}^* \) matrix is equivalent to the Wald \( F \) statistic computed by using the \( \mathbf{L} \) matrix with any row deleted that is a linear combination of previous rows

If \( \mathbf{L} \) is a full-rank matrix and all rows of \( \mathbf{L} \) are estimable functions, then \( \mathbf{L}^* \) is the same as \( \mathbf{L} \). It is possible that \( \mathbf{L} \) matrix cannot be constructed for a given set of linear contrasts, in which case the contrasts are not testable.

If the **DF=NONE** option in the MODEL statement is specified, then the procedure performs a chi-square significance test.

**Confidence Intervals**

By default, the SURVEYPHREG procedure computes \( t \) confidence limits for the estimated regression coefficients. Alternatively, you can specify the **DF=NONE** option in the MODEL statement to request standard normal confidence intervals. The \( t \) confidence interval for a linear combination \( \mathbf{y} \beta \) of the regression coefficients is computed as

\[
\left( \mathbf{y} \hat{\beta} \pm t_{df, \alpha/2} \sqrt{\mathbf{V}(\hat{\beta})^{\mathbf{y}}} \right)
\]

where \( t_{df, \alpha/2} \) is the \( 100(1 - \alpha/2) \) percentile point of the \( t \) distribution with \( df \) degrees of freedom. See the section “Degrees of Freedom” on page 8284 for more information about \( df \). If you use the **DF=NONE** option in the MODEL statement, then the procedure uses the \( 100(1 - \alpha/2) \) percentile point of the standard normal distribution.
Hazard Ratios

The hazard ratio for a quantitative effect with regression coefficient $\beta_j = e'_j \beta$ is defined as $\exp(\beta_j)$, where $e_j$ denotes the $j$th unit vector. In general, a log-hazard ratio can be written as $l\beta$, a linear combination of the regression coefficients, and the hazard ratio $\exp(l\beta)$ is obtained by replacing $e_j$ with $l$.

The confidence intervals for hazard ratios are obtained by exponentiating the confidence limits of the corresponding linear combination. Thus, the $100(1 - \alpha)$ confidence limits are

$$\exp\left(e'_j \hat{\beta} \pm t_{df, \alpha/2} \sqrt{e'_j \hat{V}(\hat{\beta}) e_j}\right)$$

where $t_{df, \alpha/2}$ is the $100(1 - \alpha/2)$ percentile point of the $t$ distribution with $df$ degrees of freedom. See the section “Degrees of Freedom” on page 8284 for more information about $df$. If you use the DF=NONE option in the MODEL statement, then the procedure uses the $100(1 - \alpha/2)$ percentile point of the standard normal distribution.

Residuals

This section describes the computation of residuals (RESMART, RESDEV, RESSCH, and RESSCO in the OUTPUT statement). See the section “Notation and Estimation” on page 8268 for definition of notation that is used in this section. The residuals are calculated based on the TIES= option in the MODEL statement.

**TIES=BRESLOW**

This is the default option. Let

$$S^{(r)}(\beta, t) = \sum_A w_{hij} y_{hij}(t) \exp(\beta'Z_{hij}(t)) Z_{hij}^{\otimes r}(t)$$

$$\tilde{Z}(\beta, t) = \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)}$$

where $r = 0, 1$; and $A$ be the set of indices in the selected sample.

Further let

$$d\Lambda_0(\beta, t) = \sum_A w_{hij} \frac{d n_{hij}(t)}{S^{(0)}(\beta, t)}$$

$$dM_{hij}(\beta, t) = d n_{hij}(t) - y_{hij}(t) \exp(\beta'Z_{hij}(t)) d\Lambda_0(\beta, t)$$

The martingale residual at $t$ is defined as

$$\hat{M}_{hij}(t) = \int_0^t dM_{hij}(\hat{\beta}, \tau)$$

$$= n_{hij}(t) - \int_0^t y_{hij}(\tau) \exp(\hat{\beta}'Z_{hij}(\tau)) d\Lambda_0(\hat{\beta}, \tau)$$

Here $\hat{M}_{hij}(t)$ estimates the difference over $(0, t]$ between the observed number of events for the $(h, i, j)$ observation unit and a conditional expected number of events. The quantity $\hat{M}_{hij} \equiv \hat{M}_{hij}(\infty)$ is referred to as the martingale residual for the $(h, i, j)$ observation unit. For the Cox model with no time-dependent
Chapter 100: The SURVEYPHREG Procedure

explanatory variables, the martingale residual for the \((h, i, j)\) unit with observation time \(t_{(h,i,j)}\) and event status \(\Delta_{(h,i,j)}\) is

\[
\hat{M}_{(h,i,j)} = \Delta_{(h,i,j)} - e^{\hat{\beta}'Z_{(h,i,j)}} \int_0^{t_{(h,i,j)}} d\Lambda_0(\hat{\beta}, s)
\]

The deviance residual \(D_{hij}\) for the \((h, i, j)\) observation unit is a transformation of the corresponding martingale residuals,

\[
D_{hij} = \text{sign}(\hat{M}_{hij}) \sqrt{2 \left[ -\hat{M}_{hij} - n_{hij}(\infty) \log \left( \frac{n_{hij}(\infty) - \hat{M}_{hij}}{n_{hij}(\infty)} \right) \right]}
\]

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 around zero than the martingale residuals. For the Cox model, the deviance residual reduces to the form

\[
D_{hij} = \text{sign}(\hat{M}_{hij}) \sqrt{2 [-\hat{M}_{hij} - \Delta_{hij} \log(\Delta_{hij} - \hat{M}_{hij})]}
\]

The Schoenfeld (1982) residual vector is calculated on a per-event-time basis. At the \(k\)th event time \(t_{hij,k}\) of the \((h, i, j)\) observation unit, the Schoenfeld residual

\[
\hat{U}_{hij}(t_{hij,k}) = Z_{hij}(t_{hij,k}) - Z(\hat{\beta}, t_{hij,k})
\]

is the difference between the observed covariate vector for the \((h, i, j)\) observation unit and the average of the covariate vectors over the risk set at \(t_{hij,k}\). 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.

The score process for the \((h, i, j)\) subject at time \(t\) is

\[
L_{hij}(\hat{\beta}, t) = \int_0^t [Z_{hij}(\tau) - Z(\hat{\beta}, \tau)] d\hat{M}_{hij}(\hat{\beta}, \tau)
\]

The vector \(\hat{L}_{hij} \equiv L_{hij}(\hat{\beta}, \infty)\) is the score residual for the \((h, i, j)\) observation unit.

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 variance estimators.

**TIES=EFRON**

For TIES=EFRON, the preceding computation is modified to comply with the Efron partial likelihood. For a given uncensored time \(t\), let \(\delta_{hij}(t) = 1\) if \(t\) is an event time for the \((h, i, j)\) observation, and 0 otherwise. Let \(d(t) = \sum_{hij \in A} \delta_{hij}(t)\), which is the number of observation units that have an event at \(t\). For \(1 \leq l \leq d(t)\), let

\[
S^{(r)}(\beta, l, t) = \sum_A w_{hij} y_{hij}(t) \left\{ 1 - \frac{l-1}{d(t)} \delta_{hij}(t) \right\} \exp (\beta'Z_{hij}(t)) Z_{hij}^{\otimes r}(t)
\]

\[
\bar{Z}(\beta, l, t) = \frac{S^{(1)}(\beta, l, t)}{S^{(0)}(\beta, l, t)}
\]

\[
d\Lambda_0(\beta, l, t) = \sum_A \frac{w_{hij} d n_{hij}(t)}{S^{(0)}(\beta, l, t)}
\]

\[
dM_{hij}(\beta, l, t) = d n_{hij}(t) - y_{hij}(t) \left( 1 - \delta_{hij}(t) \frac{l-1}{d(t)} \right) \exp (\beta'Z_{hij}(t)) d\Lambda_0(\beta, l, t)
\]
where \( r = 0, 1, \) and \( A \) are the set of indices in the selected sample.

The martingale residual at \( t \) for the \((h, i, j)\) observation unit is defined as

\[
\hat{M}_{hij}(t) = \int_0^t \frac{1}{d(t)} \sum_{l=1}^{d(t)} dM_{hij}(\hat{\beta}, l, \tau) \\
= n_{hij}(t) - \int_0^t \frac{1}{d(t)} \sum_{l=1}^{d(t)} y_{hij}(\tau) \left( 1 - \delta_{hij}(\tau) \frac{l - 1}{d(\tau)} \right) \exp \left( \hat{\beta}' \mathbf{Z}_{hij}(\tau) \right) d\Lambda_0(\hat{\beta}, l, \tau)
\]

Deviance residuals are computed by using the same transform on the corresponding martingale residuals as in TIES=BRESLOW.

The Schoenfeld residual vector for the \((h, i, j)\) observation unit at event time \( t_{hij,k} \) is

\[
\hat{U}_{hij}(t_{hij,k}) = \mathbf{Z}_{hij}(t_{hij,k}) - \frac{1}{d(t_{hij,k})} \sum_{l=1}^{d(t_{hij,k})} \frac{d(t_{hij,k})}{d(l_{hij,k})} \mathbf{Z}(\hat{\beta}, l, t_{hij,k})
\]

The score process for the \((h, i, j)\) observation unit at time \( t \) is

\[
L_{hij}(\beta, t) = \int_0^t \frac{1}{d(t)} \sum_{l=1}^{d(t)} \left( \mathbf{Z}_{hij}(\tau) - \mathbf{Z}(\beta, l, \tau) \right) dM_{hij}(\beta, l, \tau)
\]

Output Data Sets

You can use the Output Delivery System to create a SAS data set from any piece of PROC SURVEYPHREG output. See the section “ODS Table Names” on page 8295 for more information. PROC SURVEYPHREG also provides an output data set to store observation-level statistics, an output data set to store the replicate weights for BRR or jackknife variance estimation, and an output data set to store the jackknife coefficients for jackknife variance estimation.

OUT= Data Set for the OUTPUT statement

The OUTPUT statement can be used to store observation-level statistics, such as the predicted values and their standard errors, the (weighted) number of observation units at risk, martingale residuals, Schoenfeld residuals, score residuals, and deviance residuals. See the section “Residuals” on page 8289 for details about how these statistics are calculated.

Replicate Weights Output Data Set

If you specify the OUTWEIGHTS= method-option for VARMETHOD=BRR or JACKKNIFE, PROC SURVEYPHREG stores the replicate weights in an output data set. The OUTWEIGHTS= output data set contains all observations that are used in the analysis or all valid observations in the DATA= input data set. See the section “Missing Values” on page 8277 for details about valid observations.
The OUTWEIGHTS= data set contains the following variables:

- all variables in the DATA= input data set
- RepWt_1, RepWt_2, . . . , RepWt_R, which are the replicate weight variables, where R is the total number of replicates in the analysis

Each replicate weight variable contains the replicate weights for the corresponding replicate. Replicate weights equal zero for those observations not included in the replicate.

After the procedure creates replicate weights for a particular input data set and survey design, you can use the OUTWEIGHTS= method-option to store these replicate weights and then use them again in subsequent analyses, either in PROC SURVEYPHREG or in the other survey procedures. You can use a REPWEIGHTS statement to provide replicate weights for the procedure.

**Jackknife Coefficients Output Data Set**

If you specify the OUTJKCOEFS= method-option for VARMETHOD=JACKKNIFE, PROC SURVEYPHREG stores the jackknife coefficients in an output data set. The OUTJKCOEFS= output data set contains one observation for each replicate. The OUTJKCOEFS= data set contains the following variables:

- Replicate, which is the replicate number for the jackknife coefficient
- JKCoefficient, which is the jackknife coefficient for the replicate
- DonorStratum, which is the stratum of the PSU that was deleted to construct the replicate, if you specify a STRATA statement

After the procedure creates jackknife coefficients for a particular input data set and survey design, you can use the OUTJKCOEFS= method-option to store these coefficients and then use them again in subsequent analyses, either in PROC SURVEYPHREG or in the other survey procedures. You can use the JKCOEFS= option in the REPWEIGHTS statement to provide jackknife coefficients for the procedure.

**Displayed Output**

If you use the NOPRINT option in the PROC SURVEYPHREG statement, the procedure does not display any output. Otherwise, PROC SURVEYPHREG displays results of the analysis in a collection of tables.

**Model Information**

The “Model Information” table displays 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 and the values that indicate censored times, the model, the name and label of the FREQ variable, the name and label of the WEIGHT variable, the name and label of the STRATA variables, the name and label of the CLUSTER variables, and the method of handling ties in the failure time for the Cox model. The ODS name of the “Model Information” table is ModelInfo.
Number of Observations

The “Number of Observations” table displays the number of observations that are read and used, the sum of frequencies read and used, the sum of weights read and used, and the weighted sum of frequencies that are read and used in the analysis. The ODS name of the “Number of Observations” table is NObs.

Summary of the Number of Event and Censored Values

The “Summary of the Number of Event and Censored Values” table displays the number of events and censored values. The ODS name of the “Summary of the Number of Event and Censored Values” table is CensoredSummary.

Summary of the Weighted Number of Event and Censored Values

The “Summary of the Weighted Number of Event and Censored Values” table displays the weighted number of events and censored values. The ODS name of the “Summary of the Weighted Number of Event and Censored Values” table is WeightedCensoredSummary.

Class Level Information

The “Class Level Information” table is displayed when there are CLASS variables in the model. The table lists the categories of every CLASS variable that is used in the model and the corresponding design variable values. The ODS name of the “Class Level Information” table is ClassLevelInfo.

Design Summary Table

The “Design Summary” table provides information about the sample design. The table displays the total number of strata that are read and used, and the total number of clusters read and used. The table is displayed only if you specify a STRATA or CLUSTER statement. The ODS name of the “Design Summary” table is DesignSummary.

Stratum Information Table

If you specify the LIST option in the STRATA statement, PROC SURVEYPHREG displays a “Stratum Information” table. The ODS name of the “Stratum Information Table” is StrataInfo. This table provides the following information for each stratum:

• Stratum Index, which is a sequential stratum identification number
• STRATA variables, which list the levels of STRATA variables for the stratum
• Number of Observations, which is the number of observations used in the stratum
• Population Total for the stratum, if you specify the TOTAL= option
• Sampling Rate for the stratum, if you specify the TOTAL= or RATE= option. If you specify the TOTAL= option, the sampling rate is based on the number of valid observations in the stratum.
• Number of Clusters, which is the number of clusters in the stratum, if you specify a CLUSTER statement
Convergence Status

The “Convergence Status” table displays the convergence status of the optimization routine. The procedure displays this table only when you specify the NLOPTI ONS statement. The ODS name of the “Convergence Status” table is ConvergenceStatus.

Model Fit Statistics

The “Model Fit Statistics” table displays the values of –2 log likelihood and the AIC for the null model and the fitted model. The ODS name of the “Model Fit Statistics” table is FitStatistics.

Testing Global Null Hypothesis: BETA=0

The “Testing Global Null Hypothesis: BETA=0” table displays results of the likelihood ratio test and the Wald test for testing the hypothesis that all parameters are zero. The ODS name of the “Testing Global Null Hypothesis: BETA=0” table is GlobalTests.

Analysis of Maximum Likelihood Estimates

The “Analysis of Maximum Likelihood Estimates” table displays the denominator degrees of freedom, which is computed as described in the section “Degrees of Freedom” on page 8284; the maximum likelihood estimate of the parameter; the estimated standard error, computed as the square root of the corresponding diagonal element of the estimated covariance matrix; the \(t\) statistic, computed as the parameter estimate divided by the standard error; the \(p\)-value of the \(t\) statistic with respect to a \(t\) distribution with denominator degrees of freedom; and the hazard ratio estimate. The \(t\) confidence limits for the parameter estimates and estimated hazard ratios are displayed if you specify the CLPARM or RISKLIMITS option in the MODEL statement. You can specify the DF=NONE option in the MODEL statement to request \(p\)-values and confidence intervals from a standard normal distribution. If you specify the VARRATIO=ALL | MODEL | IND option in the MODEL statement, then the variance ratios for model or independence (or both) are displayed. If you specify the SERATIO=ALL | MODEL | IND option in the MODEL statement, then the standard error ratios for model or independence (or both) are displayed.

The ODS name of the “Analysis of Maximum Likelihood Estimates” table is ParameterEstimates.

Covariance Matrix

The “Covariance Matrix” table is displayed if you specify the COVB option in the MODEL statement. The table contains the estimated covariance matrix for the parameter estimates. The ODS name of the “Covariance Matrix” table is CovB.

Hessian Matrix

The “Hessian Matrix” table is displayed if you specify the HESS option in the MODEL statement. The table contains the Hessian matrix that is evaluated at the estimated regression parameters. The ODS name of the “Hessian Matrix” table is Hessian.

Inverse Hessian Matrix

The “Inverse Hessian Matrix” table is displayed if you specify the INVHESS option in the MODEL statement. The table contains the inverse of the Hessian matrix evaluated at the estimated regression parameters. The ODS name of the “Inverse Hessian Matrix” table is InvHessian.
Variance Estimation Table

The “Variance Estimation” table provides the following information:

- Method, which is the variance estimation method—Taylor Series, Balanced Repeated Replication, or Jackknife
- Replicate Weights input data set name, if you provide replicate weights with a REPWEIGHTS statement
- Number of Replicates, for VARMETHOD=BRR or VARMETHOD=JACKKNIFE
- Hadamard Data Set name, if you specify the HADAMARD= method-option for VARMETHOD=BRR
- Fay Coefficient, if you specify the FAY method-option for VARMETHOD=BRR
- Missing Values Included (MISSING), if you specify the MISSING option
- Missing Values Included (NOMCAR), if you specify the NOMCAR option
- Missing Values Excluded, if you have missing values and you do not specify the NOMCAR option

The ODS name of the “Variance Estimation” table is VarianceEstimation.

Hadamard Matrix

If you specify the PRINTH method-option for VARMETHOD=BRR, PROC SURVEYPHREG displays the Hadamard matrix that is used to construct replicates for BRR variance estimation. If you provide a Hadamard matrix with the HADAMARD= method-option for VARMETHOD=BRR but the procedure does not use the entire matrix, the procedure displays only the rows and columns that are actually used to construct replicates. The ODS name of the “Hadamard Matrix” table is HadamardMatrix.

Maximum Likelihood Estimates for Replicate Samples

If you specify the DETAILS method-option for VARMETHOD=BRR or the DETAILS method-option for VARMETHOD=JACKKNIFE, PROC SURVEYPHREG displays the “Maximum Likelihood Estimates for Replicate Samples” table. The Replicate Number column displays the replication number, the Replicate Weight column displays the name of the replicate weight variable, and the Status column displays the convergence status. The replicate number for the full sample is set to 0. If you do not specify replicate weights, then PROC SURVEYPHREG uses default names to identify the replicate weights. For more information, see “Replicate Weights Output Data Set” on page 8291. The convergence status is 1 if the maximum likelihood estimates are available for a replicate sample and 0 otherwise. If the maximum likelihood estimates are not available for a replicate sample, then the parameter estimates are set to missing for that replicate. The ODS name of the “Maximum Likelihood Estimates for Replicate Samples” table is RepEstimates.

ODS Table Names

PROC SURVEYPHREG assigns a name to each table it creates. You can use this name to refer to the table when using the Output Delivery System (ODS) to select tables and create output data sets. For more
information about ODS, see Chapter 20, “Using the Output Delivery System.” Table 100.8 lists the table names, along with the corresponding analysis options.

<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Statement / Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>CensoredSummary</td>
<td>Summary of event and censored observations</td>
<td>Default</td>
</tr>
<tr>
<td>ClassLevelInfo</td>
<td>CLASS variable levels</td>
<td>CLASS</td>
</tr>
<tr>
<td>ConvergenceStatus</td>
<td>Convergence status</td>
<td>NLOPTIONS / PALL</td>
</tr>
<tr>
<td>CovB</td>
<td>Covariance of parameter estimates</td>
<td>MODEL / COVB</td>
</tr>
<tr>
<td>DesignSummary</td>
<td>Design summary</td>
<td>STRATA or CLUSTER</td>
</tr>
<tr>
<td>FitStatistics</td>
<td>Model fit statistics</td>
<td>Default</td>
</tr>
<tr>
<td>GlobalTests</td>
<td>Tests of the global null hypothesis</td>
<td>Default</td>
</tr>
<tr>
<td>Hadamard</td>
<td>Hadamard matrix</td>
<td>PROC / VARMETHOD=BRR(PRINTH)</td>
</tr>
<tr>
<td>Hessian</td>
<td>Observed Hessian matrix</td>
<td>MODEL / HESSIAN</td>
</tr>
<tr>
<td>InvHessian</td>
<td>Inverse Hessian matrix</td>
<td>MODEL / INVHESS</td>
</tr>
<tr>
<td>IterHist</td>
<td>Iteration history</td>
<td>NLOPTIONS / PHISTORY</td>
</tr>
<tr>
<td>ModelInfo</td>
<td>Model information</td>
<td>Default</td>
</tr>
<tr>
<td>NObs</td>
<td>Number of observations</td>
<td>Default</td>
</tr>
<tr>
<td>ParameterEstimates</td>
<td>Maximum likelihood estimates</td>
<td>Default</td>
</tr>
<tr>
<td>ParameterEstimatesStart</td>
<td>Initial parameter values</td>
<td>NLOPTIONS / PALL</td>
</tr>
<tr>
<td>RepEstimates</td>
<td>Maximum likelihood estimates for replicate samples</td>
<td>PROC / VARMETHOD=BRR(DETAILS) or VARMETHOD=JACKKNIFE(DETAILS)</td>
</tr>
<tr>
<td>StrataInfo</td>
<td>Stratum information</td>
<td>STRATA / LIST</td>
</tr>
<tr>
<td>VarianceEstimation</td>
<td>Variance estimation</td>
<td>Default</td>
</tr>
<tr>
<td>WeightedCensoredSummary</td>
<td>Summary of weighted number of event and censored observations</td>
<td>WEIGHT</td>
</tr>
</tbody>
</table>

**ODS Graphics**

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

Before you create graphs, ODS Graphics must be enabled (for example, by specifying the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section “Enabling and Disabling ODS Graphics” on page 606 in Chapter 21, “Statistical Graphics Using ODS.”

The overall appearance of graphs is controlled by ODS styles. Styles and other aspects of using ODS Graphics are discussed in the section “A Primer on ODS Statistical Graphics” on page 605 in Chapter 21, “Statistical Graphics Using ODS.”
When ODS Graphics is enabled, the ESTIMATE, LSMEANS, LSMESTIMATE, and SLICE statements can produce plots that are associated with their analyses. For information about these plots, see the corresponding sections of Chapter 19, “Shared Concepts and Topics.”

### Example 100.1: Analysis of Clustered Data

When experimental units are naturally or artificially clustered, failure times of experimental units within a cluster are correlated. Lee, Wei, and Amato (1992) estimate the regression parameters in the Cox model by maximizing a partial likelihood function under an independent working correlation assumption and estimate the variance of the estimated regression coefficients by using a robust sandwich variance estimator that accounts for the intracluster dependence.

The Diabetic Retinopathy Study (DRS) is a randomized, controlled clinical trial of more than 1,700 patients across 15 medical centers. One objective of this study was to determine if photocoagulation treatment delays the occurrence of blindness. One eye of each patient was randomly assigned to treatment and the other eye to control. See Example 73.11 in Chapter 73, “The PHREG Procedure,” for more information about the data set and a similar analysis; see [http://www.nei.nih.gov/neitrials/static/study62.asp](http://www.nei.nih.gov/neitrials/static/study62.asp) for more information about the DRS.

Each patient is a cluster that contributes two observations to the input data set, one for each eye. The following variables are available:

- **ID**, patient’s identification
- **Time**, failure time
- **Status**, event indicator (0=censored, and 1=uncensored)
- **Treatment**, treatment received (1=laser photocoagulation, and 0=otherwise)
- **DiabeticType**, type of diabetes (0=juvenile onset with age of onset at 20 or under, and 1= adult onset with age of onset over 20)

The following DATA step creates the data set **Blind**, which represents 197 diabetic patients from the DRS:

```plaintext
data Blind;
  input ID Time Status DiabeticType Treatment @@;
datalines;
  5 46.23 0 1 1  5 46.23 0 1 0  14 42.50 0 0 1  14 31.30 1 0 0
  16 42.27 0 0 1  16 42.27 0 0 0  25 20.60 0 0 1  25 20.60 0 0 0
  29 38.77 0 0 1  29 0.30 1 0 0  46 65.23 0 0 1  46 54.27 1 0 0
  49 63.50 0 0 1  49 10.80 1 0 0  56 23.17 0 0 1  56 23.17 0 0 0
  61 1.47 0 0 1  61 1.47 0 0 0  71 58.07 0 1 1  71 13.83 1 1 0
  100 46.43 1 1 1  100 48.53 0 1 0  112 44.40 0 1 1  112 7.90 1 1 0
  120 39.57 0 1 1  120 39.57 0 1 0  127 30.83 1 1 1  127 38.57 1 1 0
  133 66.27 0 1 1  133 14.10 1 1 0  150 20.17 1 0 1  150 6.90 1 0 0
```

The following SAS statements request a proportional hazards regression of Time on Treatment, DiabeticType, and the Treatment × DiabeticType interaction, with Status as the censoring indicator. The CLUSTER statement indicates the observations that came from the same patient.

```sas
proc surveyphreg data=Blind;
   model Time*Status(0) = Treatment DiabeticType Treatment*DiabeticType;
   cluster id;
run;
```

Output 100.1.1 displays some summary information. There are 394 observations and 197 patients (clusters). Almost 61% of the observations are censored. The p-values for the null model are less than 0.0001 for both the likelihood ratio test and the Wald test (Output 100.1.2), which indicates that the survival time is highly dependent on Treatment and DiabeticType. In this example, the likelihood ratio statistic has a chi-square distribution with 3 degrees of freedom and the Wald statistics has the F distribution with the numerator degrees of freedom 3 and the denominator degrees of freedom 196. The denominator degrees of freedom are calculated as the number of clusters (197) minus one.

**Output 100.1.1 Summary Information**

<table>
<thead>
<tr>
<th>The SURVEYPHREG Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observations Read</td>
</tr>
<tr>
<td>Number of Observations Used</td>
</tr>
</tbody>
</table>

**Design Summary**

| Number of Clusters | 197 |

<table>
<thead>
<tr>
<th>Summary of the Number of Event and Censored Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>394</td>
</tr>
</tbody>
</table>

**Variance Estimation**

| Method | Taylor Series |

**Output 100.1.2 Global Test Results**

<table>
<thead>
<tr>
<th>Testing Global Null Hypothesis: BETA=0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Statistic</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
</tr>
<tr>
<td>Wald</td>
</tr>
</tbody>
</table>
Output 100.1.3 displays parameter estimates, standard errors, $t$ statistics, denominator degrees of freedom, $p$-values, and hazard ratios. In this example data set, Treatment and Treatment $\times$ DiabeticType interaction are significant with $p$-values 0.023 and 0.006, respectively. Since the model contains Treatment $\times$ DiabeticType interaction, the exponential of the estimated regression coefficient is not the hazard ratio. Use the ESTIMATE statement to calculate the hazard ratios.

### Output 100.1.3 Parameter Estimates

| Parameter                        | DF   | Estimate | Standard Error | $t$ Value | $p > |t|$ | Hazard Ratio |
|----------------------------------|------|----------|----------------|-----------|-------|--------------|
| Treatment                        | 196  | -0.424672| 0.185912       | -2.28     | 0.0234| 0.654        |
| DiabeticType                     | 196  | 0.340841 | 0.196577       | 1.73      | 0.0845| 1.406        |
| Treatment$\times$DiabeticTy      | 196  | -0.845665| 0.305081       | -2.77     | 0.0061| 0.429        |

### Example 100.2: Stratification, Clustering, and Unequal Weights

This example uses a data set from the National Health and Nutrition Examination Survey I (NHANES I) Epidemiologic Followup Study (NHEFS). The NHEFS is a national longitudinal survey that is conducted by the National Center for Health Statistics, the National Institute on Aging, and some other agencies of the Public Health Service in the United States. Some important objectives of this survey are to determine the relationships between clinical, nutritional, and behavioral factors; to determine mortality and hospital utilizations; and to monitor changes in risk factors for the initial cohort that represents the NHANES I population. A cohort of size 14,407, which includes all persons 25 to 74 years old who completed a medical examination at NHANES I in 1971–1975, was selected for the NHEFS. Personal interviews were conducted for every selected unit during the first wave of data collection from the year 1982 to 1984. Follow-up studies were conducted in 1986, 1987, and 1992. In the year 1986, only nondeceased persons 55 to 74 years old (as reported in the base year survey) were interviewed. The 1987 and 1992 NHEFS contain the entire nondeceased NHEFS cohort. Vital and tracing status data, interview data, health care facility stay data, and mortality data for all four waves are available for public use. See [http://www.cdc.gov/nchs/nhanes/nhefs/nhefs.htm](http://www.cdc.gov/nchs/nhanes/nhefs/nhefs.htm) for more information about the survey and the data sets.

For illustration purposes, 1,018 observations from the 1987 NHEFS public use interview data are used to create the data set cancer. The observations are obtained from 10 strata that contain 596 PSUs. The sum of observation weights for these selected units is over 19 million. Observation weights range from 359 to 129,359 with a mean of 18,747.69 and a median of 11,414. Several observation weights have large values; therefore it is reasonable to rescale the observation weights to facilitate the optimization routine. Different scaling techniques are proposed in the literature. For example, Binder (1992) uses scaled weights such that the sum of weights over the sampled units is one. Without loss of generality, the analysis weights in this example are obtained by dividing each observation weight by a large number (130,000). Because of this rescaling, you must be careful interpreting some results from PROC SURVEYPHREG.

The following variables are used in this example:

- ObsNo, unit identification
- Strata, stratum identification
- PSU, identification for primary sampling units
- ObservationWt, sampling weight associated with each unit
• AnalysisWt, obtained from the sampling weights by dividing each ObservationWt by 130,000

• Smoke, smoking status (−1 = not applicable, 1 = never smoked, 2 = current or former smoker in 1982–1984 follow-up, and 3 = current or former smoker in 1987 follow-up)

• Age, the event-time variable, defined as follows:
  – age of the subject when the first cancer was reported for subjects with reported cancer
  – age of the subject at death for deceased subjects without reported cancer
  – age of the subject as reported in 1987 follow-up (this value is used for nondeceased subjects who never reported cancer)
  – age of the subject for the entry year 1971–1975 survey if the subject has cancer (or is deceased) but the date of incident is not reported

• Cancer, cancer indicator (1 = cancer reported, 0 = cancer not reported)

• BodyWeight, body weight of the subject as reported in the 1987 follow-up, or an imputed body weight based on the subject’s age in the entry year 1971–1975 survey

The following SAS statements create the data set cancer. Note that BodyWeight for a few observations (8%) is imputed based on Age by using a deterministic regression imputation model (Särndal and Lundström (2005, chapter 12)). The imputed values are treated as observed values in this example. In other words, this example treats the data set cancer as the observed data set.

```sas
data cancer;
  input ObsNo Strata PSU AnalysisWt ObservationWt Smoke Age Cancer BodyWeight;
  datalines;
  1 3 002 0.02927 3805 2 53 1 175
  2 3 002 0.04698 6107 2 77 0 175
  3 3 039 0.02283 2968 2 50 0 160
  4 3 084 0.23414 30438 2 52 0 145
  5 3 007 0.03908 5081 1 80 0 127
  6 3 009 0.02993 3891 1 62 0 180
  7 3 050 0.18268 23748 2 60 0 140
  8 3 022 0.02283 2968 2 56 0 142
  9 3 050 0.18268 23748 2 60 0 140
... more lines ...
  1016 4 002 0.02068 2689 2 40 0 120
  1017 4 092 0.35298 45888 2 52 0 166
  1018 4 035 0.03344 4347 -1 58 0 156
;```

Suppose you want to study the occurrence of cancer for the base year survey population and its relation to smoking status and body weight. The following statements request a proportional hazards regression of Age on BodyWeight and Smoke with Cancer as the censor indicator. The STRATA, CLUSTER, and WEIGHT statements identify the variance strata, PSUs, and analysis weights respectively. The CLASS statement specifies that Smoke is a categorical variable, and the MODEL statement provides information about the analysis model. The TIES= option in the MODEL statement requests the Efron likelihood to handle tied events. If you do not specify the TIES= option in the MODEL statement, then the procedure uses the Breslow likelihood. The PHISTORY option in the NLOPTIONS statement is used to display the iteration history table.
The ESTIMATE statement computes a contrast between subjects who are reported as current (or former) smokers and the others. The EXP option in the ESTIMATE statement requests that the linear contrast be estimated in the exponential scale, which is the hazard ratio. The TEST statement requests the Type 3 test for each effect that is specified in the MODEL statement.

```plaintext
proc surveyphreg data = cancer;
strata strata;
cluster psu;
weight analysiswt;
class smoke;
model age*cancer(0) = bodyweight smoke / ties = efron;
nloptions phistory;
estimate smoke 0.5 0.5 -0.5 -0.5 / exp;
test ;
run;
```

Some summary statistics are shown in Output 100.2.1. The “Model Information” table contains information about the model such as the names for the dependent and censoring variables, and the likelihood. The “Number of Observations” table displays the number of observations and the sum of weights. A total of 1,018 observations are read from the cancer data set, but one observation is not used in the analysis because it has a zero sampling weight. The sum of weights is 146.81, which gives an estimated population size of 19,085,105 (= 146.8085 × 130,000). Note that the estimated population size would be 19,085,151 if you use the sampling weights (ObservationWt) instead of the analysis weights (AnalysisWt). The difference is due to the rounding errors in AnalysisWt. For simplicity, analysis weights are rounded at the fifth decimal place. The “Design Summary” table shows that there are 596 PSUs and 10 strata. From the censored summary tables, 11.7% subjects in the sample have reported cancer and an estimated 11.6% subjects in the study population have cancer. The “Variance Estimation” table shows that the Taylor series linearization variance estimation method is used and the observation units with missing values are excluded from the analysis. Note that the only missing unit in this data set has a zero sampling weight and hence it is not included in the analysis.

Output 100.2.1 Model Information, Data Summary, Design Summary, and Information about Variance Estimation

<table>
<thead>
<tr>
<th>The SURVEYPHREG Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model Information</strong></td>
</tr>
<tr>
<td>Data Set</td>
</tr>
<tr>
<td>Dependent Variable</td>
</tr>
<tr>
<td>Censoring Variable</td>
</tr>
<tr>
<td>Censoring Value(s)</td>
</tr>
<tr>
<td>Weight Variable</td>
</tr>
<tr>
<td>Stratum Variable</td>
</tr>
<tr>
<td>Cluster Variable</td>
</tr>
<tr>
<td>Ties Handling</td>
</tr>
</tbody>
</table>

| Number of Observations Read | 1018 |
| Number of Observations Used | 1017 |
| Sum of Weights Read         | 146.8085 |
| Sum of Weights Used         | 146.8085 |

<table>
<thead>
<tr>
<th>Design Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Strata</td>
</tr>
<tr>
<td>Number of Clusters</td>
</tr>
</tbody>
</table>
The “Iteration History” table in Output 100.2.2 shows that the procedure converged after four iterations. The “Objective Function” column contains the value of the likelihood after every iteration. The “Objective Function Change” column measures the change in the objective function between iterations; however, this is not the monitored convergence criterion. The SURVEYPHREG procedure monitors several features simultaneously to determine whether to stop an optimization.

Estimates for proportional hazards regression coefficients and their standard errors are shown in Output 100.2.3. The categorical variable *Smoke* has four levels, and GLM parameterization is used by PROC SURVEYPHREG. You can use the PARAM= option in the CLASS statement to specify other types of parameterizations. The estimated regression coefficient for *BodyWeight* is 0.012 with a standard error of 0.003. The degrees of freedom for the *t* test are equal to the number of PSUs (596) minus the number of strata (10). The “Estimates” table displays the estimated contrast and the corresponding hypothesis test. The estimated value for the contrast is –0.75. The estimated hazard for the nonsmokers is 0.47 times the estimated hazard for the current or former smokers. In this example data set, the contrast of interest is not significant at 0.05 levels. The “Type III Tests of Model Effects” table displays the Type 3 analysis. The effect variable *Smoke* has four levels. The F Value for *Smoke* is 1.49 with three numerator degrees of freedom and 586 denominator degrees of freedom.
Example 100.3: Domain Analysis

This example uses a data set from the NHANES I Epidemiologic Followup Study (NHEFS); see Example 100.2 for more information about the NHEFS.

For illustration purposes, 1,891 observations from the 1992 NHEFS vital and tracing status data set are used to estimate the regression coefficients of a proportional hazards model. The observations are obtained from 22 strata; each stratum contains either two or three primary sampling units. The sum of observation weights for these selected units is almost 103 million. Observation weights range from 1,498 to 470,154 with a mean of 54,457.11 and a median of 45,246. The following variables are used in this example. Although this example uses the observation weights directly, Binder (1992) suggests that a scaled version of the observation weights would be useful to improve the performance of the optimization routine.

The following variables are created in the data set mortality:

- **ID**, unit identification
- **VARSTRATA**, stratum identification
- **VARPSU**, identification for primary sampling units
- **SWEIGHT**, sampling weight associated with each unit
- **AGE**, the subject’s reported age at the 1992 interview if the subject was alive at that time; otherwise, the subject’s age at death
- **VITALSTATUS**, vital status of subject in 1992 (1 = alive, 3 = dead, 4 = unknown, 5 = traced alive with direct subject contact, 6 = traced alive without direct subject contact)

---

### Output 100.2.3 Parameter Estimates

#### Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>BodyWeight</td>
<td>586</td>
<td>0.011920</td>
<td>0.003155</td>
<td>3.78</td>
<td>0.0002</td>
<td>1.012</td>
</tr>
<tr>
<td>Smoke -1</td>
<td>586</td>
<td>-1.174048</td>
<td>0.739450</td>
<td>-1.59</td>
<td>0.1129</td>
<td>0.309</td>
</tr>
<tr>
<td>Smoke 1</td>
<td>586</td>
<td>-1.006515</td>
<td>0.578810</td>
<td>-1.74</td>
<td>0.0826</td>
<td>0.365</td>
</tr>
<tr>
<td>Smoke 2</td>
<td>586</td>
<td>-0.674183</td>
<td>0.558412</td>
<td>-1.21</td>
<td>0.2278</td>
<td>0.510</td>
</tr>
<tr>
<td>Smoke 3</td>
<td>586</td>
<td>0</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>1.000</td>
</tr>
</tbody>
</table>

#### Type III Tests of Model Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num</th>
<th>Den</th>
<th>DF</th>
<th>DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>BodyWeight</td>
<td>1</td>
<td>586</td>
<td>14.27</td>
<td>0.0002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoke</td>
<td>3</td>
<td>586</td>
<td>1.49</td>
<td>0.2160</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Estimate

<table>
<thead>
<tr>
<th>Label</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>DF</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>Exponentiated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row 1</td>
<td>-0.7532</td>
<td>0.3870</td>
<td>586</td>
<td>-1.95</td>
<td>0.0521</td>
<td>0.4709</td>
</tr>
</tbody>
</table>

---
• **POVARIND**, indicator for poverty area where subject’s household was located at NHANES I (1971–1975) exam, (1 = poverty area, 2 = non-poverty area)

• **GENDER**, (1 = male, 2 = female)

```sas
data mortality;
  input ID VARSTRATA VARPSU SWEIGHT AGE VITALSTATUS POVARIND GENDER;
  datalines;
   1 03 1 13312 66 1 1 1
   2 03 1 7941 71 3 1 2
   3 03 1 16048 . 4 1 1
   4 03 3 9298 58 3 1 1
   5 03 2 15336 56 3 1 2
   6 03 1 14744 63 1 1 1
   7 03 2 83729 70 1 2 2
   8 03 3 106492 57 1 2 1
   9 03 3 78083 81 3 2 2
  10 03 3 55957 79 3 2 1
  ... more lines ...
  1890 13 1 88939 59 1 2 1
  1891 13 1 59218 75 1 2 2
;
```

Suppose you want to estimate the hazard function for mortality time after adjusting for the poverty area indicator in the base year survey population. The following SAS statements request a proportional hazards regression of age (**AGE**) on poverty indicator (**POVARIND**):

```sas
proc surveyphreg data = mortality nomcar;
  class povarind;
  strata varstrata;
  cluster varpsu;
  weight sweight;
  model age*vitalstatus(1 4 5 6) = povarind;
  domain gender;
run;
```

Subjects with **VITALSTATUS** 1, 4, 5, or 6 are considered alive. The **CLASS** statement specifies that **POVARIND** is a categorical variable, the **WEIGHT** statement identifies the sampling weights, the **STRATA** statement identifies variance strata, and the **CLUSTER** statement identifies variance PSUs. The **DOMAIN** statement requests three separate analyses: for the overall data set, the male subpopulation, and the female subpopulation respectively. There are 223 observation units with missing values on age. All the units with missing age have vital status 1, 4, 5, or 6. Therefore, these subjects are considered to be alive in the current survey year 1992. Age for every observation unit in the base year survey was known from 1971–1975 NHANES I. One reasonable approach is to determine the age of these 223 units based on their age from the NHANES I data set. However, for illustration purposes, this example does not include the observation units with missing age when estimating the regression coefficients. Instead, an analysis of just the set of respondents is requested by specifying the **NOMCAR** option in the **PROC SURVEYPHREG** statement. This option uses a variance estimator that accounts for the random size of the set of respondents.

*Output 100.3.1* shows summary statistics for the overall analysis. A total of 1,891 observations are read from the input **DATA=** data set, but only 1,668 observations are used in the analysis. The remaining 223 observations have missing values in the variable age. The respondent data set represents almost 89.5 million units in the population. There are 22 strata and 55 clusters. Although only 57% observation units in the sample are alive, an estimated 69% observation units in the population are alive. This difference is reasonable
because selection probabilities for observation units are not the same. If you do not use the sampling weights, then your sample-based estimators might be biased for the corresponding finite population quantities. The “Variance Estimation” table indicates that the NOMCAR option is used for variance estimation.

**Output 100.3.1** Summary Statistics for the Entire Population

<table>
<thead>
<tr>
<th>The SURVEYPHREG Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observations Read</td>
</tr>
<tr>
<td>Number of Observations Used</td>
</tr>
<tr>
<td>Sum of Weights Read</td>
</tr>
<tr>
<td>Sum of Weights Used</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Design Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Strata</td>
</tr>
<tr>
<td>Number of Clusters</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Summary of the Number of Event and Censored Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1668</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Summary of the Weighted Number of Event and Censored Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>89439590</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
</tr>
<tr>
<td>Missing Values</td>
</tr>
</tbody>
</table>

Output 100.3.2 displays the estimated regression coefficients and their standard errors. Poverty index has two levels, and only one level is estimable. By default, PROC SURVEYPHREG estimates the first level (POVARIND 1) and assigns a zero value for the second level. The estimated regression coefficient is 0.385 with a standard error of 0.078. The estimated hazard for the poverty areas is 1.47 times higher than the estimated hazard for the non-poverty areas. The degrees of freedom are equal to the number of PSUs (55) minus the number of strata (22).

**Output 100.3.2** Inference for the Entire Population

<table>
<thead>
<tr>
<th>Analysis of Maximum Likelihood Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>POVARIND 1</td>
</tr>
<tr>
<td>POVARIND 2</td>
</tr>
</tbody>
</table>

Output 100.3.3 shows that 813 observation units in the sample are male, and they account for over 42.6 million males in the base year survey population. Approximately half of these observation units in the sample are censored, and an estimated 64.5% observation units are censored for the male subpopulation.
Output 100.3.3  Summary Statistics for the Male Subpopulation

The SURVEYPHREG Procedure

Domain Analysis for domain GENDER=1

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observations Read</td>
<td>1891</td>
</tr>
<tr>
<td>Number of Observations Used</td>
<td>813</td>
</tr>
<tr>
<td>Sum of Weights Read</td>
<td>48887067</td>
</tr>
<tr>
<td>Sum of Weights Used</td>
<td>42629905</td>
</tr>
</tbody>
</table>

Summary of the Number of Event and Censored Values

<table>
<thead>
<tr>
<th>Total</th>
<th>Event</th>
<th>Censored</th>
<th>Percent Censored</th>
</tr>
</thead>
<tbody>
<tr>
<td>813</td>
<td>404</td>
<td>409</td>
<td>50.31</td>
</tr>
</tbody>
</table>

Summary of the Weighted Number of Event and Censored Values

<table>
<thead>
<tr>
<th>Total</th>
<th>Event</th>
<th>Censored</th>
<th>Percent Censored</th>
</tr>
</thead>
<tbody>
<tr>
<td>42629905</td>
<td>15126321</td>
<td>27503584</td>
<td>64.52</td>
</tr>
</tbody>
</table>

Output 100.3.4 shows that the estimated regression coefficient for POVARIND 1 is 0.425 with a standard error of 0.157. The estimated hazard for the males in the poverty areas is 1.53 times higher than the estimated hazard for the males in the non-poverty areas. The degrees of freedom for the $t$ significant test for the male subpopulation are equal to the total number of PSUs (55) minus the total number of strata (22).

Output 100.3.4  Inference for the Male Subpopulation

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>POVARIND 1</td>
<td>33</td>
<td>0.424922</td>
<td>0.156583</td>
<td>2.71</td>
<td>0.0105</td>
<td>1.529</td>
</tr>
<tr>
<td>POVARIND 2</td>
<td>33</td>
<td>0</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Output 100.3.5 displays some summary statistics for the female subpopulation. There are 855 observation units for females in the sample, and they represent over 46.8 million females in the base year survey population. Although 63.4% females in the sample are alive, an estimated 73.2% females in the subpopulation are alive.

Output 100.3.5  Summary Statistics for the Female Subpopulation

The SURVEYPHREG Procedure

Domain Analysis for domain GENDER=2

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observations Read</td>
<td>1891</td>
</tr>
<tr>
<td>Number of Observations Used</td>
<td>855</td>
</tr>
<tr>
<td>Sum of Weights Read</td>
<td>54091604</td>
</tr>
<tr>
<td>Sum of Weights Used</td>
<td>46809685</td>
</tr>
</tbody>
</table>
Output 100.3.5 continued

Summary of the Number of Event and Censored Values

<table>
<thead>
<tr>
<th>Total</th>
<th>Event</th>
<th>Censored</th>
<th>Percent Censored</th>
</tr>
</thead>
<tbody>
<tr>
<td>855</td>
<td>313</td>
<td>542</td>
<td>63.39</td>
</tr>
</tbody>
</table>

Summary of the Weighted Number of Event and Censored Values

<table>
<thead>
<tr>
<th>Total</th>
<th>Event</th>
<th>Censored</th>
<th>Percent Censored</th>
</tr>
</thead>
<tbody>
<tr>
<td>46809685</td>
<td>12524027</td>
<td>34285658</td>
<td>73.24</td>
</tr>
</tbody>
</table>

Output 100.3.6 shows that the estimated proportional hazards regression coefficients for POVARIND for the females subpopulation (0.435) is higher than the estimated proportional hazards regression coefficients for POVARIND for the males subpopulation. The estimated hazard for the females in the poverty areas is 1.54 times higher than the estimated hazard for the females in the non-poverty areas. The degrees of freedom for the $t$ significant test for the female subpopulation are equal to the total number of PSUs (55) minus the total number of strata (22).

Output 100.3.6 Inference for the Female Subpopulation

<table>
<thead>
<tr>
<th>Analysis of Maximum Likelihood Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>POVARIND 1</td>
</tr>
<tr>
<td>POVARIND 2</td>
</tr>
</tbody>
</table>

Example 100.4: Variance Estimation by Using Replicate Weights

Consider the data set LibrarySurvey from “Getting Started: SURVEYPHREG Procedure” on page 8241. The selected sample contains 100 transactions from ten branch libraries. A set of replicate weights and jackknife coefficients are created by randomly assigning observation units in disjoint groups of nearly equal size within each stratum. A total of 46 different groups are created. The data set LibraryRepWeights is similar to the data set LibrarySurvey except that it also contains replicate weights repwt_1 to repwt_46. Each column of replicate weights is obtained by deleting one group of observations and adjusting the sampling weights for the other groups in that stratum (Rust 1985).

The data set LibraryJKCOEF contains the jackknife coefficient for every replicate sample. The variable replicate denotes the replicate number, donorstratum denotes the stratum identification for that replicate, and jkcoefficient denotes the jackknife coefficient for that replicate sample.

``` SAS
data LibrarySurvey;
  set LibrarySurvey;
  randomorder = ranuni(12345);
  run;
proc sort data = LibrarySurvey out = LibrarySurvey;
  by Branch randomorder;
  run;
data LibrarySurvey;
```
set LibrarySurvey;
array nGroup{10} (2 2 4 4 4 8 8 8);
GroupPSU = mod(_N_,nGroup/Branch);
drop randomorder nGroup1 nGroup2 nGroup3 nGroup4
    nGroup5 nGroup6 nGroup7 nGroup8 nGroup9 nGroup10;
run;

proc surveymeans data = LibrarySurvey varmethod = jk
    (outweights = LibraryRepWeights outjkcoefs = LibraryJKCOEF);
    weight SamplingWeight;
    strata Branch;
    cluster GroupPSU;
    var Age;
run;

It is not necessary to provide replicate weights to compute jackknife variance estimates using the SURVEYPHREG procedure. If you do not specify the replicate weights, then the procedure creates replicate weights for you. For this illustration, assume that LibraryRepWeights and LibraryJKCOEF are the only two data sets available for analysis.

The following SAS statements request a proportional hazards regression of `lenBorrow` on `Age`. The variable `Returned` is the censor indicator, and the value 0 indicates a censored observation. The WEIGHT statement specifies the sampling weight variable, and the REPWEIGHTS statement specifies replicate weight variables `RepWt_1` to `RepWt_46`. The JKCOEFS= option in the REPWEIGHTS statement specifies the jackknife coefficient for each replicate sample. The VARMETHOD= option in the MODEL statement requests the jackknife variance estimation method. A STRATA statement is not required when the REPWEIGHTS statement is specified.

```
proc surveyphreg data = LibraryRepWeights varmethod = jk;
    weight SamplingWeight;
    repweights RepWt_: / jkcoefs = LibraryJKCOEF;
    model lenBorrow*Returned(0) = Age;
run;
```

Output 100.4.1 displays some summary information. The “Number of Observations,” “Censored Summary,” and “Weighted Censored Summary” tables are exactly the same as in the example discussed in “Getting Started: SURVEYPHREG Procedure” on page 8241. The “Variance Estimation” table displays information about the variance estimation, such as the name of the variance estimation method and the number of replicate samples.

```
Output 100.4.1 Summary Statistics for Overall Analysis

The SURVEYPHREG Procedure

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observations Read</td>
<td>100</td>
</tr>
<tr>
<td>Number of Observations Used</td>
<td>100</td>
</tr>
<tr>
<td>Sum of Weights Read</td>
<td>11616.79</td>
</tr>
<tr>
<td>Sum of Weights Used</td>
<td>11616.79</td>
</tr>
</tbody>
</table>

Summary of the Number of Event and Censored Values

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Event</th>
<th>Censored</th>
<th>Percent Censored</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>90</td>
<td>10</td>
<td>10.00</td>
</tr>
</tbody>
</table>
```
Output 100.4.2 shows that the estimated regression coefficient is 0.0616 with a standard error of 0.009. The denominator degrees of freedom (46) for the \( t \) test is equal to the number of replicates used. Note that the estimated proportional hazards regression coefficient is the same as the estimated proportional hazards regression coefficient in the example in “Getting Started: SURVEYPHREG Procedure” on page 8241, but the standard error and the denominator degrees of freedom are different. This is not surprising because these two examples use the same estimator to estimate the regression coefficients but different estimators to estimate the variance.

Example 100.5: A Test of the Proportional Hazards Assumption by Using the Programming Statements

You can use programming statements in PROC SURVEYPHREG to create time-dependent covariates to test the proportional hazards assumption for complex survey data. Consider the data set mortality from Example 100.3. The data set contains 1,891 observations from the 1992 NHANES I Epidemiologic Followup study (NHEFS) vital and tracing status.

Suppose you want to fit a proportional hazards model to this data and construct a test for the proportional hazards assumption on gender. The following statements request a proportional hazards regression of age on gender and \( x \), where the time-dependent covariate \( x \) is created using the programming statements. The explanatory variable \( x \) assumes the value of the time variable age for the male subgroup. The variable vitalstatus is the censor indicator, and a value of 1, 4, 5, or 6 indicates a censored observation. The WEIGHT statement specifies the sampling weight, and the CLASS statement specifies that gender is a classification variable.

```sas
proc surveyphreg data = mortality nomcar;
   class gender;
   strata varstrata;
   cluster varpsu;
   weight sweight;
   model age*vitalstatus(1 4 5 6) = gender x;
   x = age*(gender=1);
run;
```
Output 100.5.1 displays some summary information. The “Number of Observations,” “Censored Summary,” and “Weighted Censored Summary” tables are exactly the same as in the example discussed in “Example 100.3: Domain Analysis” on page 8303.

**Output 100.5.1** Data Summary, Censored Summary, and Information about Variance Estimation

<table>
<thead>
<tr>
<th>The SURVEYPHREG Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observations Read</td>
</tr>
<tr>
<td>Number of Observations Used</td>
</tr>
<tr>
<td>Sum of Weights Read</td>
</tr>
<tr>
<td>Sum of Weights Used</td>
</tr>
</tbody>
</table>

Summary of the Number of Event and Censored Values

<table>
<thead>
<tr>
<th>Total</th>
<th>Event</th>
<th>Censored</th>
<th>Percent Censored</th>
</tr>
</thead>
<tbody>
<tr>
<td>1891</td>
<td>717</td>
<td>1174</td>
<td>62.08</td>
</tr>
</tbody>
</table>

Summary of the Weighted Number of Event and Censored Values

<table>
<thead>
<tr>
<th>Total</th>
<th>Event</th>
<th>Censored</th>
<th>Percent Censored</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0298E8</td>
<td>27650348</td>
<td>75328323</td>
<td>73.15</td>
</tr>
</tbody>
</table>

Variance Estimation

<table>
<thead>
<tr>
<th>Method</th>
<th>Taylor Series</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Values</td>
<td>NOMCAR</td>
</tr>
</tbody>
</table>

Output 100.5.2 displays the estimated regression coefficients and their standard errors. The variable gender has two levels, and only one level is estimable. By default, PROC SURVEYPHREG estimates the first level (GENDER 1) and assigns a zero value for the second level. The estimated regression coefficient is 1.61 with a standard error of 0.71. The estimated regression coefficient for $x$ is –0.02 with a standard error of 0.01. The $t$ statistic for $x$ is –1.55 with a $p$-value of 0.13 on 33 degrees of freedom. This test suggests that an interaction between the time variable age and gender is not significant. Therefore, there is little evidence of an exponential trend over time in the hazard ratio for gender.

**Output 100.5.2** Parameter Estimates

<table>
<thead>
<tr>
<th>Analysis of Maximum Likelihood Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>GENDER 1</td>
</tr>
<tr>
<td>GENDER 2</td>
</tr>
<tr>
<td>$x$</td>
</tr>
</tbody>
</table>

References


Subject Index

Akaike’s information criterion
  SURVEYPHREG procedure, 8287

alpha level
  hazard ratio estimates (SURVEYPHREG), 8259

balanced repeated replication
  SURVEYPHREG procedure, 8280
  variance estimation (SURVEYPHREG), 8280

Breslow method
  likelihood (SURVEYPHREG), 8261

BRR
  SURVEYPHREG procedure, 8280
  variance estimation (SURVEYPHREG), 8280

censored
  survival times (SURVEYPHREG), 8269
  values summary
  SURVEYPHREG procedure, 8293

censoring
  variable (SURVEYPHREG), 8263

CLASS variables
  programming statements (SURVEYPHREG), 8263

CLUSTER variables
  programming statements (SURVEYPHREG), 8263

covariance matrix
  SURVEYPHREG procedure, 8259

Cox regression analysis
  semiparametric model (SURVEYPHREG), 8240

degrees of freedom
  SURVEYPHREG procedure, 8284

deviance residuals
  SURVEYPHREG procedure, 8263, 8290

domain analysis
  SURVEYPHREG procedure, 8286
  programming statements (SURVEYPHREG), 8263

domains
  SURVEYPHREG procedure, 8254

donor stratum
  SURVEYPHREG procedure, 8283

Efron method

likelihood (SURVEYPHREG), 8261
  event values summary
  SURVEYPHREG procedure, 8293

Fay coefficient
  SURVEYPHREG procedure, 8282

Fay’s BRR method
  variance estimation (SURVEYPHREG), 8282

finite population correction
  SURVEYPHREG procedure, 8246

frequency variable
  programming statements (SURVEYPHREG), 8263
  values (SURVEYPHREG), 8256

global null hypothesis
  SURVEYPHREG procedure, 8294

Hadamard matrix
  BRR variance estimation (SURVEYPHREG), 8282

hazard function
  baseline (SURVEYPHREG), 8269
  hazard ratios
    Wald’s confidence limits (SURVEYPHREG), 8260

Hessian matrix
  SURVEYPHREG procedure, 8260

inverse Hessian matrix
  SURVEYPHREG procedure, 8260

jackknife
  SURVEYPHREG procedure, 8283
  coefficients
    SURVEYPHREG procedure, 8283
  variance estimation
    SURVEYPHREG procedure, 8283

Lee-Wei-Amato model
  SURVEYPHREG procedure, 8297

likelihood ratio test
  SURVEYPHREG procedure, 8287, 8294

linear predictor
  SURVEYPHREG procedure, 8262, 8263

linearization method
  SURVEYPHREG procedure, 8279

local influence
  score residuals (SURVEYPHREG), 8263, 8290
martingale residuals
SURVEYPHREG procedure, 8263
missing values
SURVEYPHREG procedure, 8264, 8277
model
fit criteria (SURVEYPHREG), 8287
model information
SURVEYPHREG procedure, 8292, 8293
number of observations
SURVEYPHREG procedure, 8293
number of replicates
SURVEYPHREG procedure, 8280, 8282, 8283
number of subjects at risk
SURVEYPHREG procedure, 8262
options summary
ESTIMATE statement, 8255
parameter estimates
SURVEYPHREG procedure, 8294
partial likelihood
SURVEYPHREG procedure, 8268, 8274
primary sampling units (PSUs)
SURVEYPHREG procedure, 8254
programming statements
SURVEYPHREG procedure, 8263, 8265
proportional hazards model
SURVEYPHREG procedure, 8240
replicate weights
SURVEYPHREG procedure, 8265, 8279
replication methods
SURVEYPHREG procedure, 8279
residuals
deviance (SURVEYPHREG), 8263, 8290
martingale (SURVEYPHREG), 8263
Schoenfeld (SURVEYPHREG), 8263, 8290, 8291
score (SURVEYPHREG), 8263, 8290
response variable
SURVEYPHREG procedure, 8263
risk set
SURVEYPHREG procedure, 8274
sample design
SURVEYPHREG procedure, 8275
sampling rates
SURVEYPHREG procedure, 8246, 8276
sampling weights
SURVEYPHREG procedure, 8268, 8276
Schoenfeld residuals
SURVEYPHREG procedure, 8263, 8290, 8291
score residuals
SURVEYPHREG procedure, 8263, 8290
semiparametric model
SURVEYPHREG procedure, 8240
singularity criterion
SURVEYPHREG procedure, 8261
standard error
SURVEYPHREG procedure, 8262, 8263, 8294
standard error ratio
SURVEYPHREG procedure, 8285
STRATA variables
programming statements (SURVEYPHREG), 8263
stratification
SURVEYPHREG procedure, 8267, 8275
subdomain analysis, see domain analysis
subpopulation analysis, see domain analysis
subgroup analysis, see domain analysis
survey data analysis
SURVEYPHREG procedure, 8240
survey sampling
data analysis (SURVEYPHREG), 8240
SURVEYPHREG procedure, 8240
Akaike’s information criterion, 8287
alpha level, 8259
balanced repeated replication, 8280
Breslow likelihood, 8261
BRR, 8280
BRR variance estimation, 8280
censored values summary, 8293
clustering, 8254, 8275
continuous time scale, 8261
covariance matrix, 8259
Cox regression analysis, 8240
DATA step statements, 8263
degrees of freedom, 8284
design summary table, 8293
displayed output, 8292
domain analysis, 8286
domain variable, 8254
domains, 8254
donor stratum, 8283
Efron likelihood, 8261
event values summary, 8293
Fay coefficient, 8282
Fay’s BRR variance estimation, 8282
finite population correction, 8246
global null hypothesis, 8294
Hadamard matrix (BRR variance estimation), 8282
hazard ratio confidence intervals, 8259, 8260
Hessian matrix, 8260
hypothesis tests and confidence intervals, 8287
inverse Hessian matrix, 8260
jackknife, 8283
jackknife coefficients, 8283
jackknife variance estimation, 8283
Lee-Wei-Amato model, 8297
likelihood ratio test, 8287, 8294
linear predictor, 8262, 8263
linearization method, 8279
missing values, 8264, 8277
model fit statistics, 8287
model information, 8293
number of observations, 8293
number of replicates, 8280, 8282, 8283
number of subjects at risk, 8262
ODS graph names, 8296
ODS graphics, 8296
ODS table names, 8295
output data sets, 8291
OUTPUT statistics, 8262, 8263
parameter estimates, 8294
parameter estimates confidence intervals, 8259
partial likelihood, 8268, 8274
population totals, 8247, 8276
primary sampling units (PSUs), 8254
programming statements, 8263, 8265
proportional hazards model, 8240
replicate weights, 8265, 8279
replication methods, 8279
residuals, 8263, 8289–8291
risk set, 8274
sample design, 8275
sampling rates, 8246, 8276
sampling weights, 8268, 8276
singularity criterion, 8261
standard error, 8262, 8263, 8294
standard error ratio, 8285
stratification, 8267, 8275
survival distribution function, 8270
survival times, 8269
survivor function, 8269, 8270
Taylor series linearized variance estimation, 8251
Taylor series variance estimation

SURVEYPHREG procedure
ordering of effects, 8246
survival distribution function
SURVEYPHREG procedure, 8270
survival times
SURVEYPHREG procedure, 8269
survivor function
definition (SURVEYPHREG), 8270
SURVEYPHREG procedure, 8269
Syntax Index

ALPHA= option
   MODEL statement (SURVEYPHREG), 8259

BY statement
   SURVEYPHREG procedure, 8251

CENTER= option
   VARMETHOD=BRR (PROC SURVEYPHREG statement), 8248
   VARMETHOD=JK (PROC SURVEYPHREG statement), 8250

CLASS statement
   SURVEYPHREG procedure, 8251

CLPARM option
   MODEL statement (SURVEYPHREG), 8259

CLUSTER statement
   SURVEYPHREG procedure, 8254

COVB option
   MODEL statement (SURVEYPHREG), 8259

DATA= option
   PROC SURVEYPHREG statement, 8245

DESCENDING option
   CLASS statement (SURVEYPHREG), 8252

DETAILS option
   VARMETHOD=BRR (PROC SURVEYPHREG statement), 8248
   VARMETHOD=JK (PROC SURVEYPHREG statement), 8250

DF= option
   MODEL statement (SURVEYPHREG), 8259
   REPWEIGHTS statement (SURVEYPHREG), 8265

DF=ALLREPS
   DF= (SURVEYPHREG), 8260

DF=NONE
   DF= (SURVEYPHREG), 8260

DF=PARMADJ
   DF= (SURVEYPHREG), 8260

DOMAIN statement
   SURVEYPHREG procedure, 8254

ESTIMATE statement
   SURVEYPHREG procedure, 8255

FAY= option
   VARMETHOD=BRR (PROC SURVEYPHREG statement), 8248

FREQ statement

SURVEYPHREG procedure, 8256

HADAMARD= option
   VARMETHOD=BRR (PROC SURVEYPHREG statement), 8248

HESS option
   MODEL statement (SURVEYPHREG), 8260

INVHESS option
   MODEL statement (SURVEYPHREG), 8260

JKCOEFS= option
   REPWEIGHTS statement (SURVEYPHREG), 8265

keyword= option
   OUTPUT statement (SURVEYPHREG), 8262

LIST option
   STRATA statement (SURVEYPHREG), 8267

LSMEANS statement
   SURVEYPHREG procedure, 8256

LSMEANtimate statement
   SURVEYPHREG procedure, 8257

MISSING option
   CLASS statement (SURVEYPHREG), 8252
   PROC SURVEYPHREG statement, 8252

MODEL statement
   SURVEYPHREG procedure, 8258

NLOPTIONS statement
   SURVEYPHREG procedure, 8262

NOMCAR option
   PROC SURVEYPHREG statement, 8252

NOPRINT option
   PROC SURVEYPHREG statement, 8252

ORDER= option
   CLASS statement (SURVEYPHREG), 8252
   PROC SURVEYPHREG statement, 8256

OUT= option
   OUTPUT statement (SURVEYPHREG), 8262

OUTJKCOEFS= option
   VARMETHOD=JK (PROC SURVEYPHREG statement), 8251

OUTPUT statement
   SURVEYPHREG procedure, 8262

OUTWEIGHTS= option
VARMETHOD=BRR (PROC SURVEYPHREG statement), 8249
VARMETHOD=JK (PROC SURVEYPHREG statement), 8250

PARAM= option
CLASS statement (SURVEYPHREG), 8252
PRINTH option
VARMETHOD=BRR (PROC SURVEYPHREG statement), 8249

PROC SURVEYPHREG statement, see SURVEYPHREG procedure

RATE= option
PROC SURVEYPHREG statement, 8246
REF= option
CLASS statement (SURVEYPHREG), 8253
REPS= option
VARMETHOD=BRR (PROC SURVEYPHREG statement), 8250

REPWEIGHTS statement
SURVEYPHREG procedure, 8265
RISKLIMITS= option
MODEL statement (SURVEYPHREG), 8260

SERATIO= option
MODEL statement (SURVEYPHREG), 8260
SINGULAR= option
MODEL statement (SURVEYPHREG), 8261
SLICE statement
SURVEYPHREG procedure, 8266
SURVEYPHREG procedure, PROC SURVEYPHREG statement
DATA= option, 8245
MISSING option, 8245
STORE statement
SURVEYPHREG procedure, 8266
STRATA statement
SURVEYPHREG procedure, 8267
SURVEYPHREG procedure
DF=ALLREPS, 8260
DF=NONE, 8260
DF=PARMADIJ, 8260
NLOPTIONS statement, 8262
SURVEYPHREG procedure, BY statement, 8251
SURVEYPHREG procedure, CLASS statement, 8251
MISSING option, 8252
ORDER= option, 8252
PARAM= option, 8252
SURVEYPHREG procedure, CLUSTER statement, 8254
SURVEYPHREG procedure, DOMAIN statement, 8254
SURVEYPHREG procedure, FREQ statement, 8256
SURVEYPHREG procedure, MODEL statement, 8258
ALPHA= option, 8259
CLPARM option, 8259
COVB option, 8259
DF= option, 8259
HESS option, 8260
IN VHess option, 8260
RISK LIMITS= option, 8260
SERATIO= option, 8260
SINGULAR= option, 8261
TIES= option, 8261
V ADJUST= option, 8261
VARRATIO= option, 8261

SURVEYPHREG procedure, NLOPTIONS statement, 8262
SURVEYPHREG procedure, OUTPUT statement, 8262
keyword= option, 8262
OUT= option, 8262

SURVEYPHREG procedure, PROC SURVEYPHREG statement, 8245
CENTER= option (V ARMETHOD=BRR), 8248
CENTER= option (V ARMETHOD=JK), 8250
DETAILS option (V ARMETHOD=BRR), 8248
DETAILS option (V ARMETHOD=JK), 8250
FAY= option (V ARMETHOD=BRR), 8248
HADAMARD= option (V ARMETHOD=BRR), 8249
NOMCAR option, 8245
NOPRINT option, 8245
OUTJKCOEFS= option (V ARMETHOD=JK), 8251

OUTWEIGHTS= option (V ARMETHOD=BRR), 8249
OUTWEIGHTS= option (V ARMETHOD=JK), 8250

PRIN TH option (V ARMETHOD=BRR), 8249
RATE= option, 8246
REPS= option (V ARMETHOD=BRR), 8250
TOTAL= option, 8247
V ARMETHOD= option, 8247

SURVEYPHREG procedure, REPWEIGHTS statement, 8265
DF= option, 8265
JKCOEFS= option, 8265

SURVEYPHREG procedure, STRATA statement, 8267
LIST option, 8267

SURVEYPHREG procedure, WEIGHT statement, 8268

SURVEYPHREG procedure, CLASS statement
DESCENDING option, 8252
REF= option, 8253
TRUNCATE option, 8252
SURVEYPHREG procedure, ESTIMATE statement, 8255
SURVEYPHREG procedure, LSMEANS statement, 8256
SURVEYPHREG procedure, LSMESTIMATE statement, 8257
SURVEYPHREG procedure, PROC SURVEYPHREG statement
   ORDER= option, 8246
SURVEYPHREG procedure, SLICE statement, 8266
SURVEYPHREG procedure, STORE statement, 8266
SURVEYPHREG procedure, TEST statement, 8267

TEST statement
   SURVEYPHREG procedure, 8267
TIES= option
   MODEL statement (SURVEYPHREG), 8261
TOTAL= option
   PROC SURVEYPHREG statement, 8247
TRUNCATE option
   CLASS statement (SURVEYPHREG), 8253

VADJUST= option
   MODEL statement (SURVEYPHREG), 8261
VARMETHOD= option
   PROC SURVEYPHREG statement, 8247
VARRATIO= option
   MODEL statement (SURVEYPHREG), 8261

WEIGHT statement
   SURVEYPHREG procedure, 8268