SAS/STAT® 15.1 User's Guide
The SURVEYPHREG Procedure

This document is an individual chapter from *SAS/STAT® 15.1 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2018. *SAS/STAT® 15.1 User's Guide*. Cary, NC: SAS Institute Inc.

SAS/STAT® 15.1 User's Guide

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

November 2018

Chapter 119
The SURVEYPHREG Procedure

Contents

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

$$\lambda_i(t) = \lambda(t; \mathbf{Z}_i(t)) = \lambda_0(t)\ \exp(\mathbf{Z}_i'(t)\boldsymbol{\beta})$$

where $\lambda_0(t)$ is an arbitrary and unspecified baseline hazard function, $\mathbf{Z}_i(t)$ is the vector of explanatory variables for the ith population unit at time t, and $\boldsymbol{\beta}$ is the vector of unknown regression parameters.

The finite population regression parameter $\boldsymbol{\beta}_N$ 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 $\hat{\boldsymbol{\beta}}$ of the proportional hazards regression parameters $\boldsymbol{\beta}_N$ for the finite population by maximizing the partial pseudo-log-likelihood $l_\pi(\boldsymbol{\beta}; \mathbf{Z}_i(t), t_i)$ based on observed covariates $\mathbf{Z}_i(t)$ and observed survival time t_i. The procedure also produces an estimate of the sampling variance $\mathrm{V}(\hat{\boldsymbol{\beta}}|\mathcal{F}_N)$, which assumes that the values of the finite population $\mathcal{F}_N$ 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 9887 introduces PROC SURVEYPHREG with an example. The section "Syntax: SURVEYPHREG Procedure" on page 9891 describes the syntax of the procedure. The section "Details: SURVEYPHREG Procedure" on page 9923 summarizes the statistical techniques employed in PROC SURVEYPHREG. The section "Examples: SURVEYPHREG Procedure" on page 9963 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 Figure 119.1. 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, Figure 119.1 shows that 10% of the sampled units are censored and an estimated 10.05% of the population units are censored.

Figure 119.1 Summary Statistics

The SURVEYPHREG Procedure

Model Information	
Data Set	WORK.LIBRARYSURVEY
Dependent Variable	lenBorrow
Censoring Variable	Returned
Censoring Value(s)	0
Weight Variable	SamplingWeight
Stratum Variable	Branch
Ties Handling	BRESLOW

Number of Observations Read	100
Number of Observations Used	100
Sum of Weights Read	11616.79
Sum of Weights Used	11616.79

Design Summary	
Number of Strata	10

Figure 119.1 *continued*

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
100	90	10	10.00

Summary of the Weighted Number of Event and Censored Values

Total	Event	Censored	Percent Censored
11616.79	10449.22	1167.57	10.05

Variance Estimation	
Method	Taylor Series

Parameter estimates and their standard errors are shown in Figure 119.2. 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 Figure 119.2 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 119.2 Weighted Estimates and Their Standard Errors

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
Age	90	0.061593	0.008366	7.36	<.0001	1.064

Syntax: SURVEYPHREG Procedure

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

PROC SURVEYPHREG < *options* > ;
BY *variables* ;
CLASS *variable* < **(***options***)** > < *. . . variable* < **(***options***)** > > < */ options* > ;
CLUSTER *variables* ;
DOMAIN *variable* < *(* '*formatted-level-value*' *. . .* '*formatted-level-value*' *)* > < *variable* < *(* '*formatted-level-value*' *. . .* '*formatted-level-value*' *)* >∗*variable* < *(* '*formatted-level-value*' *. . .* '*formatted-level-value*' *)* > > ;
ESTIMATE < '*label*' > *estimate-specification* < */ options* > ;
FREQ *variable* ;
HAZARDRATIO < '*label*' > *variable* < */options* > ;
LSMEANS < *model-effects* > < */ options* > ;
LSMESTIMATE *model-effect lsmestimate-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 following sections describe the PROC SURVEYPHREG statement and then describe the other statements 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 119.1 summarizes the *options* available in the PROC SURVEYPHREG statement.

Table 119.1 PROC SURVEYPHREG Statement Options

Option	Description
ATRISK	Displays a table that contains the sum of weights for the number of units and the sum of weights for the corresponding number of events in the risk sets
DATA=	Names the input SAS data set
MISSING	Treats missing values as a valid category
NAMELEN=	Specifies the length of effect names
NOMCAR	Uses missing observations specified as *not missing completely at random*
NOPRINT	Suppresses all displayed output
ORDER=	Specifies the sort order of CLASS variables
RATE=	Specifies the sampling rate
TOTAL=	Specifies the total number of primary sampling units
VARMETHOD=	Specifies the variance estimation method

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

ATRISK
displays a table that contains the sum of weights for the number of units at risk at each distinct event time and the sum of weights for the corresponding number of events in the risk sets. For example, the risk set information in Figure 119.3 is displayed if the ATRISK option is specified in the example in the section "Getting Started: SURVEYPHREG Procedure" on page 9887.

Figure 119.3 Risk Set Information

The SURVEYPHREG Procedure

Risk Set Sum of Weights		
lenBorrow	At Risk	Event
1	11616.79	5440.11
2	6176.68	1177.71
3	4998.97	926.55
4	4072.42	1411.07
5	2661.35	461.89
6	2199.46	565.01
7	1634.45	236.58
8	1397.87	230.3

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 9935.

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

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 9935 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. Other replication methods do not use the NOMCAR option.

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.

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:

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

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

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

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

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

specifies the sampling rate, which PROC SURVEYPHREG uses to compute a finite population correction for Taylor series or bootstrap variance estimation. This option is ignored for BRR and jackknife variance estimation.

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.

You can specify the sampling rate in either of the following ways:

value specifies a nonnegative number to use for a nonstratified design or for a stratified design that has the same sampling rate in each stratum.

SAS-data-set specifies a *SAS-data-set* that contains the stratification variables and the sampling rates for a stratified design that has different sampling rates in the strata. You must provide the sampling rates in the data set variable named _RATE_.

The sampling rates 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%.

For more information, see the section "Population Totals and Sampling Rates" on page 9934.

If you do not specify the RATE= or TOTAL= option, then the Taylor series or bootstrap 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 population. PROC SURVEYPHREG uses the *value* to compute a finite population correction for Taylor series or bootstrap variance estimation. This option is ignored for BRR and jackknife variance estimation.

You can specify the total number of PSUs in either of the following ways:

value specifies a positive number to use for a nonstratified design or for a stratified design that has the same population total in each stratum.

SAS-data-set specifies a *SAS-data-set* that contains the stratification variables and the population totals for a stratified design that has different population totals in the strata. You must provide the stratum totals in the data set variable named _TOTAL_.

The stratum totals must be positive numbers.

For more information, see the section "Population Totals and Sampling Rates" on page 9934.

If you do not specify the TOTAL= or RATE= option, then the Taylor series or bootstrap 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=*method* **<(***method-options***)>**
specifies the variance estimation *method*. PROC SURVEYPHREG provides the Taylor series method and balanced repeated replication (BRR), jackknife, and bootstrap replication (resampling) methods.

Table 119.2 summarizes the available *methods* and *method-options*.

Table 119.2 Variance Estimation Options

method	**Variance Estimation Method**	*method-options*
BOOTSTRAP	Bootstrap	CENTER=FULLSAMPLE | REPLICATES DETAILS MH=*number* | *SAS-data-set* OUTWEIGHTS=*SAS-data-set* REPS=*number* SEED=*number*
BRR	Balanced repeated replication	CENTER=FULLSAMPLE | REPLICATES DETAILS FAY <=*value*> HADAMARD=*SAS-data-set* OUTWEIGHTS=*SAS-data-set* PRINTH REPS=*number*
JACKKNIFE	Jackknife	CENTER=FULLSAMPLE | REPLICATES DETAILS OUTJKCOEFS=*SAS-data-set* OUTWEIGHTS=*SAS-data-set*
TAYLOR	Taylor series linearization	None

By default, VARMETHOD=JACKKNIFE if you also specify a REPWEIGHTS statement; otherwise, VARMETHOD=TAYLOR by default.

You can specify the following *methods*:

BOOTSTRAP <(*method-options***)>**
requests variance estimation by the bootstrap method. The bootstrap method requires at least two primary sampling units (PSUs) in each stratum for stratified designs unless you provide replicate weights by using a REPWEIGHTS statement. For more information, see the section "Bootstrap Method" on page 9941.

You can specify the following *method-options*:

CENTER=FULLSAMPLE | REPLICATES
defines how to compute the deviations for the bootstrap method. You can specify the following values:

FULLSAMPLE computes the deviations of the replicate estimates from the full sample estimate.

REPLICATES computes the deviations of the replicate estimates from the average of the replicate estimates.

For more information, see the section "Bootstrap Method" on page 9941. By default, CENTER=FULLSAMPLE.

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.

MH=*value* | **(***values***)** | *SAS-data-set*

specifies the number of PSUs to select for the bootstrap replicate samples. You can provide bootstrap stratum sample sizes m_h by specifying a list of *values* or a *SAS-data-set*. Alternatively, you can provide a single bootstrap sample size *value* to use for all strata or for a nonstratified design. For more information, see the section "Bootstrap Method" on page 9941.

Each bootstrap sample size m_h must be a positive integer and must be less than n_h, which is the total number of PSUs in stratum h. By default, $m_h = n_h - 1$ for a stratified design. For a nonstratified design, the bootstrap sample size *value* must be less than n (the total number of PSUs in the sample). By default, $m = n - 1$ for a nonstratified design.

You can provide the bootstrap sample size by specifying one of the following forms:

MH=*value*

specifies a single bootstrap sample size *value* to use for all strata or for a nonstratified design.

MH=(*values***)**

specifies a list of stratum bootstrap sample size *values*. You can separate the values with blanks or commas, and you must enclose the list of values in parentheses. The number of values must not be less than the number of strata in the DATA= input data set.

The order of the stratum sample size values must match the order of the stratum levels in the DATA= input data set. Each stratum sample size value must be a positive integer and must be less than the total number of PSUs in the corresponding stratum.

MH=*SAS-data-set*

names a *SAS-data-set* that contains the stratum bootstrap sample sizes. You must provide the sample sizes in a data set variable named _NSIZE_ or SampleSize.

The *SAS-data-set* must contain all stratification variables that you specify in the STRATA statement. It must also contain all stratum levels that appear in the DATA= input data set. The order of the stratum levels in the *SAS-data-set* must match the order of the levels in the DATA= data set. If formats are associated with the STRATA variables, the formats must be consistent in the two data sets.

Each value of the _NSIZE_ or SampleSize variable must be a positive integer and must be less than the total number of PSUs in the corresponding stratum.

OUTWEIGHTS=*SAS-data-set*
names a *SAS-data-set* in which to store the replicate weights that PROC SURVEYPHREG creates for bootstrap variance estimation. For information about replicate weights, see the section "Bootstrap Method" on page 9941. For information about the contents of the OUTWEIGHTS= data set, see the section "Replicate Weights Output Data Set" on page 9957.

This *method-option* is not available when you provide replicate weights in a REPWEIGHTS statement.

REPS=*number*
specifies the *number* of replicates for bootstrap variance estimation, where *number* must be an integer greater than 1. Increasing the number of replicates improves the estimation precision but also increases the computation time. By default, REPS=250.

SEED=*number*
specifies the initial seed for random number generation, where *number* must be a positive integer.

If you do not specify this option or if you specify a *number* that is negative or 0, PROC SURVEYPHREG uses the time of day from the computer's clock to obtain an initial seed.

The seed that is used is displayed in the "Variance Estimation" table.

To reproduce the same bootstrap replicate weights and the same analysis in a subsequent execution of PROC SURVEYPHREG, you can specify the same initial seed that was used in the original analysis.

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 9939 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 9939 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 9940 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.

HADAMARD=*SAS-data-set*

H=*SAS-data-set*
names a SAS data set that contains the Hadamard matrix for BRR replicate construction. If you do not provide a Hadamard matrix with the HADAMARD= *method-option*, PROC SURVEYPHREG generates an appropriate Hadamard matrix for replicate construction. See the sections "Balanced Repeated Replication (BRR) Method" on page 9939 and "Hadamard Matrix" on page 9940 for details.

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*.

In the HADAMARD= input data set, each variable corresponds to a column of the Hadamard matrix, and each observation corresponds to a row of the matrix. You can use any variable names in the HADAMARD= data set. All values in the data set must equal either 1 or –1. You must ensure that the matrix you provide is indeed a Hadamard matrix—that is, $\mathbf{A}'\mathbf{A} = R\mathbf{I}$, where $\mathbf{A}$ is the Hadamard matrix of dimension R and $\mathbf{I}$ is an identity matrix. PROC SURVEYPHREG does not check the validity of the Hadamard matrix that you provide.

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 equal to 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 PROC SURVEYPHREG uses to construct replicates for BRR variance estimation.

OUTWEIGHTS=*SAS-data-set*
names an output SAS data set to store the replicate weights that PROC SURVEYPHREG creates for BRR variance estimation. For more information about replicate weights, see the section "Balanced Repeated Replication (BRR) Method" on page 9939. For more information about the contents of the OUTWEIGHTS= data set, see the section "Replicate Weights Output Data Set" on page 9957.

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

PRINTH

displays the Hadamard matrix that is used to construct replicates for BRR variance estimation. When you provide the Hadamard matrix in the HADAMARD= *method-option*, PROC SURVEYPHREG displays only the rows and columns that are actually used to construct replicates. For more information, see the sections "Balanced Repeated Replication (BRR) Method" on page 9939 and "Hadamard Matrix" on page 9940.

The PRINTH *method-option* is not available when you provide replicate weights by using a REPWEIGHTS statement, because PROC SURVEYPHREG 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 by using the HADAMARD= *method-option*, the number of replicates should be greater than the number of strata and should be a multiple of 4. For more information, see the section "Balanced Repeated Replication (BRR) Method" on page 9939. 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 to be larger than the REPS= value that you specify.

If you provide a Hadamard matrix by using the HADAMARD= *method-option*, the value of REPS= must not be greater 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 9942 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 | REPLICATES

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 9942 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.

OUTJKCOEFS=*SAS-data-set*
names an output SAS data set that contains jackknife coefficients. See the section "Jackknife Coefficients Output Data Set" on page 9957 for more details about the contents of the OUTJKCOEFS= data set.

OUTWEIGHTS=*SAS-data-set*
names an output SAS data set that contains replicate weights. See the section "Jackknife Method" on page 9942 for more information about replicate weights. See the section "Replicate Weights Output Data Set" on page 9957 for more details about the contents of the OUTWEIGHTS= data set.

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

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 9938 for more information.

BY Statement

BY *variables* **;**

You can specify a BY statement in 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 in 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.

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

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. If the PARAM= option is not specified together with any individual CLASS variable, then by default, PARAM=GLM.

Otherwise, the default is PARAM=EFFECT. 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
EFFECT	Effect coding
GLM	Less-than-full-rank reference cell coding (this *keyword* can be used only in a global option)
ORDINAL **THERMOMETER**	Cumulative parameterization for an ordinal CLASS variable
POLYNOMIAL **POLY**	Polynomial coding
REFERENCE **REF**	Reference cell coding
ORTHEFFECT	Orthogonalizes PARAM=EFFECT coding
ORTHORDINAL **ORTHOTHERM**	Orthogonalizes PARAM=ORDINAL coding
ORTHPOLY	Orthogonalizes PARAM=POLYNOMIAL coding
ORTHREF	Orthogonalizes PARAM=REFERENCE coding

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 a PARAM= option is specified as a variable option for some variables, then any variables for which PARAM= is not specified use either the EFFECT parameterization if the global PARAM= option is not specified, or the full-rank parameterization indicated in the global PARAM= option if specified. If the global PARAM=GLM option is specified and PARAM= is also specified for some variables, GLM parameterization is used for all variables.

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. For more information, see the section "Other Parameterizations" on page 397 in Chapter 19, "Shared Concepts and Topics."

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*:

FIRST designates the first ordered level as reference.

LAST designates the last ordered level as reference.

By default, REF=LAST.

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 a STRATA statement is specified, 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. For more information, see the section "Specifying the Sample Design" on page 9933.

If you provide replicate weights for replication variance estimation by specifying 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. For more information, 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 *variable* < *('formatted-level-value' . . . 'formatted-level-value')* > < *variable* < *('formatted-level-value' . . . 'formatted-level-value')* >∗*variable* < *('formatted-level-value' . . . 'formatted-level-value')* > > **;**

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 can yield inappropriate estimates of variance for domain statistics.

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 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*.

By default, the SURVEYPHREG procedure performs analyses for all levels of domains that are formed by the variables in the DOMAIN statement. Optionally, you can specify domain analyses for particular levels of a DOMAIN variable by listing quoted *formatted-level-values* in parentheses after the variable name. You must enclose each *formatted-level-value* in single or double quotation marks. You can specify one or more levels of each variable; when you specify more than one level, separate the levels by a space or a comma. The following example requests domain analysis only for females within each race category:

```
domain Race*Gender(''Female'');
```

The following example requests domain analyses only for white and Asian races, and separate domain analyses for both genders:

```
domain Race('White','Asian') Gender;
```

If a domain variable appears more than once in any domain cross-classification but the specified levels for that domain variable are not the same, then PROC SURVEYPHREG includes all specified levels of that variable in the domain cross-classification.

In the following example, two different levels for Race are specified in two DOMAIN statements:

```
domain Race('White')*Gender;
domain Race('Asian')*Gender;
```

Thus, the preceding specification is equivalent to the following:

```
domain Race('Asian' 'White')*Gender;
```

However, if a domain variable appears more than once in cross-classifications but the levels for that domain variable are not specified in all cross-classifications, then PROC SURVEYPHREG includes only the specified levels.

In the following example, a level for Gender is specified in the first DOMAIN statement but no levels for Gender are specified in the second DOMAIN statement:

```
domain Race('White')*Gender('Female');
domain Race('Asian')*Gender;
```

Thus, the preceding specification is equivalent to the following:

```
domain Race('White' 'Asian')*Gender('Female');
```

ESTIMATE Statement

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

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

Table 119.3 summarizes the *options* available in the ESTIMATE statement.

Table 119.3 ESTIMATE Statement Options

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

Table 119.3 *continued*

Option	Description
Generalized Linear Modeling	
CATEGORY=	Specifies how to construct estimable functions for multinomial data
EXP	Exponentiates and displays estimates
ILINK	Computes and displays estimates and standard errors on the inverse linked scale

For more information about the syntax of the ESTIMATE statement, see the section "ESTIMATE Statement" on page 451 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.

HAZARDRATIO Statement

HAZARDRATIO <'*label*'> *variable* <*/ options*> **;**

The HAZARDRATIO statement enables you to request hazard ratios for any variable in the model at customized settings. For example, if the model contains the interaction of a CLASS variable A and a continuous variable X, the following specification displays a table of hazard ratios that compares the hazards of each pair of levels of A at X=3:

```
hazardratio A / at (X=3);
```

The HAZARDRATIO statement identifies the variable whose hazard ratios are to be evaluated. If the variable is a continuous variable, the hazard ratio compares the hazards for a specified change (by default, an increase of 1 unit) in the variable. For a CLASS variable, a hazard ratio compares the hazards of two levels of the variable. You can specify more than one HAZARDRATIO statement, and an optional label (specified as a quoted string) helps identify the output.

Table 119.4 summarizes the *options* available in the HAZARDRATIO statement.

Table 119.4 HAZARDRATIO Statement Options

Option	Description
ALPHA=	Specifies the alpha level
AT	Specifies the variables that interact with the variable of interest
DIFF=	Specifies which differences to compute
E	Displays the coefficients for the log-hazard ratio
UNITS=	Specifies the units of change

You can specify the following *options* in the HAZARDRATIO statement:

ALPHA=*number*

specifies the alpha level of the interval estimates for the hazard ratios, where *number* must be between 0 and 1. The default is the value of the ALPHA= option in the PROC SURVEYPHREG statement, or 0.05 if that option is not specified.

AT (*variable***=ALL | REF |** *list* **< . . .** *variable***=ALL | REF |** *list* **>)**

specifies the variables that interact with the variable of interest and the corresponding values of the interacting variables. If the interacting variable is continuous and a numeric *list* is specified after the equal sign, hazard ratios are computed for each value in the list. If the interacting variable is a CLASS variable, you can specify, after the equal sign, a list of quoted strings that correspond to various levels of the CLASS variable, or you can specify the keyword ALL or REF. Hazard ratios are computed at each value of the list if a *list* is specified, or at each level of the interacting variable if ALL is specified, or at the reference level of the interacting variable if REF is specified.

If this option is not specified, PROC SURVEYPHREG finds all the variables that interact with the variable of interest. If an interacting variable is a CLASS variable, *variable*=ALL is the default; if the interacting variable is continuous, *variable*=*m* is the default, where *m* is the weighted average of the observed values of the continuous variable.

Suppose the model contains two interactions: an interaction A*B of CLASS variables A and B, and another interaction A*X of A with a continuous variable X. If 3.5 is the weighted average of the values of X, the following two HAZARDRATIO statements are equivalent:

```
hazardratio A;
hazardratio A / at (B=ALL X=3.5);
```

DIFF=*diff-request*

specifies which differences to consider for the level comparisons of a CLASS variable. This option is ignored in the estimation of hazard ratios for a continuous variable. You can specify the following *diff-requests*:

DISTINCT

DISTINCTPAIRS

requests all comparisons of only the distinct combinations of pairs.

PAIRWISE
requests all possible pairwise comparisons of levels.

REFERENCE
REF
requests comparisons between the reference level and all other levels of the CLASS variable.

For example, let A be a CLASS variable that has three levels (A1, A2, and A3), where A3 is specified as the reference level. The following table depicts the hazard ratios that are displayed for the three alternatives of the DIFF= option.

	Hazard Ratios Displayed					
DIFF=*option*	**A1 vs A2**	**A2 vs A1**	**A1 vs A3**	**A3 vs A1**	**A2 vs A3**	**A3 vs A2**
DISTINCT	√		√		√	
PAIRWISE	√	√	√	√	√	√
REF			√		√	

By default, DIFF=DISTINCT.

E
displays the vector $\mathbf{l}$ of linear coefficients such that $\mathbf{l}'\boldsymbol{\beta}$ is the log-hazard ratio, where $\boldsymbol{\beta}$ is the vector of regression coefficients.

UNITS=*value*
specifies the units of change in the continuous explanatory variable for which the customized hazard ratio is estimated. This option is ignored in the computation of the hazard ratios for a CLASS variable. By default, UNITS=1.

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 119.5 summarizes the *options* available in the LSMEANS statement.

Table 119.5 LSMEANS Statement Options

Option	Description
Construction and Computation of LS-Means	
AT	Modifies the covariate value in computing LS-means
BYLEVEL	Computes separate margins
DIFF	Computes differences of LS-means
OM=	Specifies the weighting scheme for LS-means computation as determined by the input data set

Table 119.5 *continued*

Option	Description
SINGULAR=	Tunes estimability checking
Degrees of Freedom and *p*-Values	
ADJUST=	Determines the method of multiple-comparison adjustment of LS-means differences
ALPHA=α	Determines the confidence level ($1 - \alpha$)
STEPDOWN	Adjusts multiple-comparison *p*-values further in a step-down fashion
Statistical Output	
CL	Constructs confidence limits for means and mean differences
CORR	Displays the correlation matrix of LS-means
COV	Displays the covariance matrix of LS-means
E	Prints the **L** matrix
LINES	Uses connecting lines to indicate nonsignificantly different subsets of LS-means
LINESTABLE	Displays the results of the LINES option as a table
MEANS	Prints the LS-means
PLOTS=	Produces graphs of means and mean comparisons
SEED=	Specifies the seed for computations that depend on random numbers
Generalized Linear Modeling	
EXP	Exponentiates and displays estimates of LS-means or LS-means differences
ILINK	Computes and displays estimates and standard errors of LS-means (but not differences) on the inverse linked scale
ODDSRATIO	Reports (simple) differences of least squares means in terms of odds ratios if permitted by the link function

For more information about the syntax of the LSMEANS statement, see the section "LSMEANS Statement" on page 467 in Chapter 19, "Shared Concepts and Topics."

LSMESTIMATE Statement

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

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

Table 119.6 summarizes the *options* available in the LSMESTIMATE statement.

Table 119.6 LSMESTIMATE Statement Options

Option	Description
Construction and Computation of LS-Means	
AT	Modifies covariate values in computing LS-means
BYLEVEL	Computes separate margins
DIVISOR=	Specifies a list of values to divide the coefficients
OM=	Specifies the weighting scheme for LS-means computation as determined by a data set
SINGULAR=	Tunes estimability checking
Degrees of Freedom and *p*-Values	
ADJUST=	Determines the method of multiple-comparison adjustment of LS-means differences
ALPHA=α	Determines the confidence level $(1 - \alpha)$
LOWER	Performs one-sided, lower-tailed inference
STEPDOWN	Adjusts multiple-comparison *p*-values further in a step-down fashion
TESTVALUE=	Specifies values under the null hypothesis for tests
UPPER	Performs one-sided, upper-tailed inference
Statistical Output	
CL	Constructs confidence limits for means and mean differences
CORR	Displays the correlation matrix of LS-means
COV	Displays the covariance matrix of LS-means
E	Prints the **L** matrix
ELSM	Prints the **K** matrix
JOINT	Produces a joint *F* or chi-square test for the LS-means and LS-means differences
PLOTS=	Produces graphs of means and mean comparisons
SEED=	Specifies the seed for computations that depend on random numbers
Generalized Linear Modeling	
CATEGORY=	Specifies how to construct estimable functions for multinomial data
EXP	Exponentiates and displays LS-means estimates
ILINK	Computes and displays estimates and standard errors of LS-means (but not differences) on the inverse linked scale

For more information about the syntax of the LSMESTIMATE statement, see the section "LSMESTIMATE Statement" on page 487 in Chapter 19, "Shared Concepts and Topics."

MODEL Statement

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

MODEL **(***t1***,** *t2***)** < ∗ *censor***(***list***)** > **=** *effects* < / *options* > **;**

The MODEL statement identifies the variables to be used as the failure time variables, the optional censoring variable, and the explanatory effects, including covariates, main effects, and interactions. For more information about explanatory effects, see the section "Specification of Effects" on page 4020 in Chapter 50, "The GLM Procedure." 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. Specify either of two forms of MODEL syntax: the first form allows one time variable, and the second form allows two time variables for the counting process style of input. For more information on the counting process style of input, see the section "Counting Process Style of Input" on page 9929.

For the first form of the MODEL statement, the name of the failure time variable (*response*) precedes the equal sign. This variable 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 (*effects*) are the explanatory variables (sometimes called independent variables or covariates) for the model.

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

The censoring variable must be numeric. The failure time variable must contain nonnegative values. Any observation that has a negative failure time is excluded from the analysis, as is any observation that has a missing value for any of the variables listed in the MODEL statement. For more information, see the section "Missing Values" on page 9935. Failure time variables that have a SAS date format are not recommended because the dates might be translated into negative numbers and consequently the corresponding observation would be discarded.

Table 119.7 summarizes the *options* available in the MODEL statement, which can be specified after a slash (/).

Table 119.7 MODEL Statement Options

Option	Description
ALPHA=	Specifies α for the $100(1-\alpha)\%$ confidence limits
CLPARM	Computes confidence limits for regression parameters
COVB	Displays covariance matrix
DF=	Specifies the denominator degrees of freedom
ENTRYTIME=	Specifies the delayed entry time variable
FIRTH	Specifies Firth's penalized likelihood method
HESS	Displays the Hessian matrix
INVHESS	Displays the inverse of the Hessian matrix
RISKLIMITS	Computes confidence limits for the exponentials of the regression parameters

Table 119.7 *continued*

Option	Description
SERATIO=	Computes the ratio of two standard errors for the regression coefficients
SINGULAR=	Specifies tolerance for testing singularity
TIES=	Specifies the method of handling ties in failure times
VADJUST=	Specifies a variance adjustment factor
VARRATIO=	Computes the ratio of two variances for the regression coefficients

ALPHA=α

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 α produces $100(1-\alpha)\%$ confidence limits. The default of ALPHA=0.05 produces 95% confidence limits.

The ALPHA= option has no effect unless you also specify the CLPARM or RISKLIMITS option.

CLPARM

produces confidence limits for regression parameters of Cox proportional hazards models. You can specify the confidence coefficient by using the ALPHA= option. Classification main effects that use parameterizations other than REF, EFFECT, or GLM are ignored. For more information, see the section "Confidence Intervals" on page 9951.

COVB

displays the estimated covariance matrix of the parameter estimates.

DF=*value* | *keyword* **<(***value***)>**

specifies the denominator degrees of freedom for hypothesis tests, specifies the degrees of freedom for confidence limits, and requests adjustments to the Wald test statistics. If you specify a *value*, it must be a nonnegative number.

In the description that follows, d denotes the usual degrees of freedom computed from the survey data by using the number of strata, clusters, or replicate weights. For more information, see the section "Degrees of Freedom" on page 9945.

By default, DF=PARMADJ when you use the Taylor series linearized variance estimator, and DF=DESIGN when you use the replication variance estimator. Alternatively, you can specify a nonnegative *value* for the degrees of freedom, or you can specify one of the following *keywords*:

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 that are used. Some replicate samples might not be usable, in the sense that they cannot be used for variance estimation because of factors such as inestimability or nonconvergence. These replicate samples are not accounted for in the denominator degrees of freedom unless you specify DF=ALLREPS. For more information, see the section "Degrees of Freedom" on page 9945.

DESIGN
computes the denominator degrees of freedom as d. When you specify DF=DESIGN, the corresponding Wald F statistics do not account for the number of parameters in the model. This option is useful if you do not want to apply the adjustment described in Korn and Graubard (1999, p. 93). For more information, see the section "Testing the Global Null Hypothesis" on page 9948.

DESIGN (*value*)
computes the denominator degrees of freedom as *value*. When you specify DF=DESIGN (*value*), the corresponding Wald F statistics do not account for the number of parameters in the model. This option is useful if you do not want to apply the adjustment described in Korn and Graubard (1999, p. 93) and you want to specify the denominator degrees of freedom. You might want to specify a denominator degrees of freedom other than d for reasons such as missing values or domain estimation for relatively small domains. For more information, see the section "Testing the Global Null Hypothesis" on page 9948.

DESIGNADJ
computes the denominator degrees of freedom as d. When you specify DF=DESIGNADJ, the corresponding Wald F statistics account for the number of parameters in the model. This option is useful if you are fitting a model that has many parameters relative to d but you want to use d as the denominator degrees of freedom. For more information, see the section "Testing the Global Null Hypothesis" on page 9948.

NONE
specifies the denominator degrees of freedom to be infinite. This option is useful if you want to compute chi-square tests and normal confidence intervals. For more information, see the section "Testing the Global Null Hypothesis" on page 9948.

PARMADJ
computes the denominator degrees of freedom as d minus the number of nonsingular parameters plus 1. When you specify DF=PARMADJ, the corresponding Wald F statistics account for the number of parameters in the model. This option is useful if you are fitting a model that has many parameters relative to d. For more information, see the section "Testing the Global Null Hypothesis" on page 9948.

PARMADJ (*value*)
computes the denominator degrees of freedom as *value*. When you specify DF=PARMADJ (*value*), the corresponding Wald F statistics account for the number of parameters in the model. This option is useful if you are fitting a model with that has parameters relative to d and you want to specify the denominator degrees of freedom. You might want to specify the denominator degrees of freedom for reasons such as missing values or domain estimation for relatively small domains. For more information, see the section "Testing the Global Null Hypothesis" on page 9948.

ENTRYTIME=*variable*

ENTRY=*variable*
specifies the name of the variable that represents the left-truncation time. This option has no effect when the counting process style of input is specified. For more information, see the section "Left-Truncation of Failure Times" on page 9930.

FIRTH
performs Firth's penalized maximum likelihood estimation to reduce bias in the parameter estimates (Heinze and Schemper 2001; Firth 1993). This method is useful when the likelihood is monotone—that is, the likelihood converges to a finite value, but at least one estimate diverges to infinity. This option is available only for the Breslow likelihood. When you specify this option, the likelihood ratio statistics are computed using the unadjusted likelihoods, and only the Wald test for the overall null hypothesis is available. For more information, see the section "Firth's Modification for Maximum Likelihood Estimation" on page 9931.

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. For more information, see the section "Hazard Ratios" on page 9951. You can specify the confidence coefficient by using the ALPHA= option. You must take great care 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 | SRSWOR | SRSWR
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 IND, MODEL, and either SRSWR or SRSWOR standard error ratios. If you specify the RATE= or the TOTAL= option in the PROC SURVEYPHREG statement, then SRSWOR standard error ratios are computed; otherwise, SRSWR standard error ratios are computed.

IND
computes the standard errors in the denominator by ignoring stratification and clustering. For more information, see the section "Variance Ratios and Standard Error Ratios" on page 9947.

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. For more information, see the section "Variance Ratios and Standard Error Ratios" on page 9947.

SRSWOR
computes the standard errors in the denominator as the square root of the diagonals of a scaled inverse Hessian matrix evaluated at the estimated regression parameters. If you specify the RATE= or the TOTAL= option in the PROC SURVEYPHREG statement, then the scaling factor also includes the sampling fractions. For more information, see the section "Variance Ratios and Standard Error Ratios" on page 9947.

SRSWR
computes the standard errors in the denominator as the square root of the diagonals of a scaled inverse Hessian matrix evaluated at the estimated regression parameters. For more information, see the section "Variance Ratios and Standard Error Ratios" on page 9947.

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. You can specify the following *methods*:

BRESLOW
uses the approximate partial likelihood of Breslow (1974).

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. By default, TIES=BRESLOW, which is the most efficient method when there are no ties.

VADJUST=DF | PARMADJ | NONE | AVGREPSS
specifies variance adjustment factors. You can specify the following keywords:

DF

PARMADJ
requests the degrees-of-freedom adjustment $(n-1)/(n-p)$ in the computation of the matrix $\mathbf{G}$ for the Taylor series linearization variance estimation.

NONE
excludes the degrees-of-freedom adjustment $(n-1)/(n-p)$ from the computation of the matrix $\mathbf{G}$ for the Taylor series linearization variance estimation. By default, VADJUST=NONE.

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. VADJUST=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. For more information, see the section "Variance Adjustment Factors" on page 9946.

VARRATIO=ALL | MODEL | IND | SRSWOR | SRSWR
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 IND, MODEL, and either SRSWR or SRSWOR variance ratios. If you specify the RATE= or the TOTAL= option in the PROC SURVEYPHREG statement, then SRSWOR variance ratios are computed; otherwise, SRSWR variance ratios are computed.

IND
computes the variances in the denominator by ignoring stratification and clustering. For more information, see the section "Variance Ratios and Standard Error Ratios" on page 9947.

MODEL
computes the variances in the denominator as the diagonals of the inverse Hessian matrix evaluated at the estimated regression parameters. For more information, see the section "Variance Ratios and Standard Error Ratios" on page 9947.

SRSWOR
computes the variances in the denominator as the diagonals of a scaled inverse Hessian matrix evaluated at the estimated regression parameters. If you specify the RATE= or the TOTAL= option in the PROC SURVEYPHREG statement, then the scaling factor also includes the sampling fractions. For more information, see the section "Variance Ratios and Standard Error Ratios" on page 9947.

SRSWR
computes the variances in the denominator as the diagonals of a scaled inverse Hessian matrix evaluated at the estimated regression parameters. For more information, see the section "Variance Ratios and Standard Error Ratios" on page 9947.

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 499 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 ($\mathbf{z}'_j\hat{\boldsymbol{\beta}}$) 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*n* convention. See the section "OUT= Data Set for the OUTPUT Statement" on page 9956 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 τ_j (or at the right endpoint of the at-risk interval when a counting-process specification is used in the MODEL statement, as described in the section "Counting Process Style of Input" on page 9929).

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 τ_j can be interpreted as the difference over $[0, \tau_j]$ in the observed number of events minus the expected number of events given by the model.

RESSCH

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, $\sqrt{\mathbf{z}_j' \hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}|F)\mathbf{z}_j}$.

WTATRISK

specifies the weighted number of subjects at risk at the observation time τ_j (or at the right endpoint of the at-risk interval when a counting-process specification is used in the MODEL statement, as described in the section "Counting Process Style of Input" on page 9929).

XBETA

specifies the estimate of the linear predictor, $\mathbf{z}_j' \hat{\boldsymbol{\beta}}$.

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. You cannot use programming statements in PROC SURVEYPHREG to modify the observed values of the explanatory variables that are available in the input data set.

The following DATA step statements are available in PROC SURVEYPHREG:

ABORT;
ARRAY *arrayname* < **[** *dimensions* **]** > < **$** > < *variables-and-constants* >**;**
CALL *name* < **(***expression* < **,** *expression . . .* > **)** >**;**
DELETE;
DO < *variable* **=** *expression* < **TO** *expression* > < **BY** *expression* > >
    < **,** *expression* < **TO** *expression* > < **BY** *expression* > > . . .
    < **WHILE** *expression* > < **UNTIL** *expression* >**;**
END;
GOTO *statement-label***;**
IF *expression***;**
IF *expression* **THEN** *program-statement***;**
                ELSE *program-statement***;**
variable **=** *expression***;**
variable **+** *expression***;**
LINK *statement-label***;**
PUT < *variable* > < **=** > . . . **;**
RETURN;
SELECT < **(***expression***)** >**;**
STOP;
SUBSTR(*variable***,** *index***,** *length***)=** *expression***;**
WHEN (*expression***)***program-statement***;**
        OTHERWISE *program-statement***;**

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.

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

BP1	blood pressure at T1
T2	time 2
BP2	blood pressure at T2
WT	design weight
PSU	identification of primary sampling units

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

```
proc 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 replication variance estimation, which you request with the VARMETHOD=BOOTSTRAP, VARMETHOD=BRR, or VARMETHOD=JACKKNIFE option in the PROC SURVEYPHREG statement. If you do not provide a REPWEIGHTS statement for VARMETHOD=BRR or VARMETHOD=JACKKNIFE, then PROC SURVEYPHREG constructs replicate weights for the analysis. For more information, see the sections "Balanced Repeated Replication (BRR) Method" on page 9939 and "Jackknife Method" on page 9942. For VARMETHOD=BOOTSTRAP, you must specify the REPWEIGHTS statement to provide replicate weights. For more information, see the section "Bootstrap Method" on page 9941.

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 (/):

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=*value*

specifies a single jackknife coefficient for all replicates. The coefficient *value* must be a nonnegative number.

JKCOEFS=(*values***)**

specifies jackknife coefficients for VARMETHOD=JACKKNIFE, where each coefficient corresponds to an individual replicate identified by a REPWEIGHTS variable. You can separate *values* with blanks or commas. The coefficient *values* must be nonnegative numbers. The number of *values* 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=*SAS-data-set*

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 9942 for details about jackknife coefficients.

REPCOEFS=*replication-coefficient-specification*

specifies replicate coefficients for replication methods. When you specify VARMETHOD=JACKKNIFE, the default value for the replicate coefficient is $(R-1)/R$, where R is the total number of replicates. When you specify VARMETHOD=BOOTSTRAP or VARMETHOD=BRR, the default value for the replicate coefficient is $1/R$.

For VARMETHOD=BOOTSTRAP or VARMETHOD=JACKKNIFE, you can specify one of the following three *replication-coefficient-specifications*:

REPCOEFS=*value*

specifies a single replicate coefficient for all replicates, where *value* must be a nonnegative number.

REPCOEFS=(*values***)**

specifies a list of replicate coefficients, where each coefficient corresponds to an individual replicate that is identified by a REPWEIGHTS variable. You can separate *values* with blanks or commas. The coefficient *values* must be nonnegative numbers. The number of *values* must equal the number of replicate weight variables specified in the REPWEIGHTS statement. List these values in the same order in which you list the corresponding replicate weight variables in the REPWEIGHTS statement.

REPCOEFS=*SAS-data-set*

names a *SAS-data-set* that contains the replicate coefficients. You must provide the replicate coefficients in a variable named Coefficient or RepCoefficient in the *SAS-data-set*. Each coefficient value must be a nonnegative number. The observations in the *SAS-data-set* should correspond to the replicates that are identified by the variables that are specified in the REPWEIGHTS statement. Arrange the coefficients or observations in the *SAS-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 *SAS-data-set* must not be less than the number of variables specified in the REPWEIGHTS statement.

For more information about replication coefficients, see the section "Replicate Weights Method" on page 9944.

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.

This statement uses the same *options* as the LSMEANS statement, which are summarized in Table 19.23 in Chapter 19, "Shared Concepts and Topics." For more information about the syntax of the SLICE statement, see the section "SLICE Statement" on page 516 in Chapter 19, "Shared Concepts and Topics."

STORE Statement

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

The STORE statement saves 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 using the PLM procedure. For more information about the syntax of the STORE statement, see the section "STORE Statement" on page 520 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 9933 for more information.

If you provide replicate weights for replication variance estimation in 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 89, "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. For more information about constructing Type I, II, and III estimable functions, see Chapter 15, "The Four Types of Estimable Functions."

Table 119.8 summarizes the *options* available in the TEST statement.

Table 119.8 TEST Statement Options

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

For more information about the syntax of the TEST statement, see the section "TEST Statement" on page 521 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 9935 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

$$\lambda_i(t) = \lambda(t; \mathbf{Z}_i(t)) = \lambda_0(t)\ \exp(\mathbf{Z}_i'(t)\boldsymbol{\beta})$$

where $\lambda_0(t)$ is an arbitrary and unspecified baseline hazard function, $\mathbf{Z}_i(t)$ is the vector of explanatory variables for the ith unit at time t, and $\boldsymbol{\beta}$ is the vector of unknown regression parameters that are associated with the explanatory variables. The vector $\boldsymbol{\beta}$ 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 $\boldsymbol{\beta}$. Let $\boldsymbol{\beta}_N$ be the desired estimator. Assuming a working model with uncorrelated responses, $\boldsymbol{\beta}_N$ is obtained by maximizing the partial log likelihood,

$$l(\boldsymbol{\beta}) = \sum_{i \in U} \log\left\{ \mathrm{L}(\boldsymbol{\beta}; \mathbf{Z}_i(t), t_i) \right\}$$

with respect to $\boldsymbol{\beta}$, where $\mathrm{L}(\boldsymbol{\beta}; \mathbf{Z}_i(t), t_i)$ is Cox's partial likelihood function.

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

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

See "Partial Likelihood Function for the Cox Model" on page 9928 for more details.

A sample-based estimator $\hat{\boldsymbol{\beta}}$ for the finite population quantity $\boldsymbol{\beta}_N$ can be obtained by maximizing the partial pseudo-log-likelihood $l_\pi(\boldsymbol{\beta}; \mathbf{Z}_i(t), t_i)$ with respect to $\boldsymbol{\beta}$. The design-based variance for $\hat{\boldsymbol{\beta}}$ is obtained by assuming the set of finite population values $\mathcal{F}_N$ 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

$$(\mathbf{w}, \mathbf{t}, \boldsymbol{\Delta}, \mathbf{Z}) = (w_{hij}, t_{hij}, \Delta_{hij}, \mathbf{z}_{hij})$$

where

- $\mathbf{w}$ denotes the vector of sampling weights
- $\mathbf{t}$ denotes the event time variable
- $\mathbf{\Delta}$ denotes the event indicator
- $\mathbf{Z}$ 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 or bootstrap variance estimation. You can specify the stratum sampling rates with the RATE= option. Or if you specify population totals with the TOTAL= option, PROC SURVEYPHREG 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 9934 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 Variable Usage

The following DATA step creates an artificial data set, Test, to be used in this section. Test contains six variables: T contains the failure times; Status is the censoring indicator variable, whose value is 1 for an uncensored failure time and 0 for a censored failure time; A is an auxiliary variable whose value ranges from 1.0 to 3.9; MirrorT is an exact copy of T; W is the observation weight; and S is the strata indicator.

```
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
 12    0    3   10   1    4    1   3   20   2
 11    1    3   10   1    8    1   1   20   2
  6    1    3   10   1    7    1   3   20   2
  7    1    3   10   1   12    1   3   20   2
  9    1    2   10   1   15    1   2   20   2
  3    1    2   10   1   14    0   3   20   2
  6    1    1   10   1   13    1   2   20   2
;
```

Time Variable on the Right Side of the MODEL Statement

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

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

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

```
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.

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 119.4. 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 119.4 shows the parameters that correspond to A and their respective regression coefficients estimates.

Figure 119.4 Design Variable and Regression Coefficient Estimates

The SURVEYPHREG Procedure

Class Level Information

Class	Levels	Values
A	3	1 2 3

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
A 1	30	-1.162184	0.644483	-1.80	0.0814	0.313
A 2	30	-0.616962	0.513355	-1.20	0.2388	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 119.5 and are identical to those in Figure 119.4. The `if A=3 then A=2` programming statement has no effect on the explanatory variable for A, which have already been determined.

Figure 119.5 Design Variable and Regression Coefficient Estimates

The SURVEYPHREG Procedure

Class Level Information		
Class	**Levels**	**Values**
A	3	1 2 3

Analysis of Maximum Likelihood Estimates						
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Hazard Ratio**
A 1	30	-1.162184	0.644483	-1.80	0.0814	0.313
A 2	30	-0.616962	0.513355	-1.20	0.2388	0.540
A 3	30	0	.	.	.	1.000

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.

Figure 119.6 Single Time-Dependent Variable X*A

The SURVEYPHREG Procedure

Analysis of Maximum Likelihood Estimates						
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Hazard Ratio**
A 1	31	0.158010	1.182556	0.13	0.8946	1.171
A 2	31	0.008993	0.652504	0.01	0.9891	1.009
A 3	31	0	.	.	.	1.000
X	31	0.092679	0.071328	1.30	0.2034	1.097

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 119.7.

Figure 119.7 Simple Test of Proportional Hazards Assumption

The SURVEYPHREG Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
A 1	31	-0.007655	1.221122	-0.01	0.9950	0.992
A 2	31	-0.881383	1.743507	-0.51	0.6168	0.414
A 3	31	0	.	.	.	1.000
X1	31	-0.155220	0.164334	-0.94	0.3522	0.856
X2	31	0.011554	0.188932	0.06	0.9516	1.012

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 $\mathcal{D}_k$ of individuals that fail at $t_{(k)}$. Let w_{hij} be the weight associated with the jth observation unit in the ith cluster in stratum h. Using this notation, the pseudo-likelihood functions used in PROC SURVEYPHREG to estimate $\boldsymbol{\beta}_N$ are described in the following sections.

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

The Breslow likelihood is expressed as

$$L_{\text{Breslow}}(\boldsymbol{\beta}) = \prod_{k=1}^{K} \frac{\exp\left(\boldsymbol{\beta}' \sum_{\mathcal{D}_k} w_{hij}\mathbf{Z}_{hij}(t)\right)}{\left\{\sum_{\mathcal{R}_k} w_{hij} \exp(\boldsymbol{\beta}'\mathbf{Z}_{hij}(t))\right\}^{\sum_{\mathcal{D}_k} w_{hij}}}$$

The Efron likelihood is expressed as

$$L_{\text{Efron}}(\boldsymbol{\beta}) = \prod_{k=1}^{K} \frac{\exp\left(\boldsymbol{\beta}' \sum_{\mathcal{D}_k} w_{hij}\mathbf{Z}_{hij}(t)\right)}{\{\phi(\boldsymbol{\beta}, \mathbf{Z}, \mathbf{w}, k)\}^{\frac{1}{d_k}\sum_{\mathcal{D}_k} w_{hij}}}$$

where $\phi(\boldsymbol{\beta}, \mathbf{Z}, \mathbf{w}, k)$ is

$$\phi(\boldsymbol{\beta}, \mathbf{Z}, \mathbf{w}, k) = \prod_{l=1}^{d_k} \left\{\sum_{\mathcal{R}_k} w_{hij} \exp\left(\boldsymbol{\beta}'\mathbf{Z}_{hij}(t)\right) - \frac{l-1}{d_k} \sum_{\mathcal{D}_k} w_{hij} \exp\left(\boldsymbol{\beta}'\mathbf{Z}_{hij}(t)\right)\right\}$$

Counting Process Style of Input

In the counting process formulation, data for each subject are identified by a triple $\{N, Y, \mathbf{Z}\}$ of counting, at-risk, and covariate processes. $N(t)$ indicates the sum of weights for all events that the subject experiences over the time interval $(0, t]$, $Y(t)$ indicates whether the subject is at risk at time t (1 if at risk and 0 otherwise), and $\mathbf{Z}(t)$ is a vector of explanatory variables for the subject at time t. The sample path of N is a step function with jumps at the event times, and $N(0) = 0$. Unless $\mathbf{Z}(t)$ changes continuously with time, the data for each subject can be represented by multiple observations, each of which identifies by a semiclosed time interval $(t_1, t_2]$, the values of the explanatory variables over that interval, and the event status at t_2. The subject remains at risk during the interval $(t_1, t_2]$, and an event might occur at t_2. Values of the explanatory variables for the subject remain unchanged in the interval. This style of data input was originated by Therneau (1994).

For example, suppose a patient (ID=1) with an analysis weight of 10 has a tumor recurrence at weeks 3, 10, and 15 and is followed up until week 23. Consider three fixed explanatory variables Trt (treatment), Number (initial tumor number), and Size (initial tumor size), one weight variable Weight (analysis weight), one patient identification variable ID, and one time-dependent covariate Z that represents a hormone level. The value of Z might change during the follow-up period. The data for this patient are represented by the following four observations:

ID	Weight	T1	T2	Status	Trt	Number	Size	Z
1	10	0	3	1	1	1	3	12.3
1	10	3	10	1	1	1	3	14.7
1	10	10	15	1	1	1	3	13.8
1	10	15	23	0	1	1	3	15.5

Here (T1,T2] contains the at-risk intervals. The variable Status indicates whether a recurrence has occurred at T2: a value of 1 indicates a tumor recurrence, and a value of 0 indicates non-recurrence. Assume the patients are selected independently. Because there are multiple observation rows for every patient, you should use the CLUSTER statement to identify each individual patient. The CLUSTER statement computes the variability between the patients. The following statements fit a multiplicative hazards model with baseline covariates Trt, Number, and Size, and a time-varying covariate Z. For more information, see the section "The Multiplicative Hazards Model" on page 9930.

```
proc surveyphreg;
   weight Weight;
   cluster ID;
   model (T1,T2) * Status(0) = Trt Number Size Z;
run;
```

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

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

Left-Truncation of Failure Times

Left-truncation occurs when individuals are not observed at the natural time origin of the phenomenon under study but come under observation at some known later time (called the left-truncation time). The risk set just prior to an event time does not include individuals whose left-truncation times exceed that event time. Thus, any contribution to the likelihood must be conditional on the truncation limit having been exceeded.

You use the ENTRY= option to specify the variable that represents the left-truncation time. Suppose T1 and T2 represent the left-truncation time and the survival time, respectively. To account for left-truncation, you specify the following statements:

```
proc surveyphreg;
   model T2*Dead(0)=X1-X10/entry=T1;
   title 'The ENTRY= option is Specified';
run;
```

Equivalently, you can use the counting process style of input for left-truncation:

```
proc surveyphreg;
   model (T1,T2)*Dead(0)=X1-X10;
   title 'Counting Process Style of Input';
run;
```

The Multiplicative Hazards Model

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

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

where

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

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

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

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

where w_i is the weight for subject i, $\Delta N_i(t) = 1$ if $N_i(t) - N_i(t-) = 1$, and $\Delta N_i(t) = 0$ otherwise.

A CLUSTER statement is necessary in order to appropriately estimate variances from multiplicative hazards models. If your design has a primary sampling unit (PSU), then use the PSU identification in the CLUSTER statement. Otherwise, use the subject identification in the CLUSTER statement.

Firth's Modification for Maximum Likelihood Estimation

In fitting a Cox model, the phenomenon of monotone likelihood is observed if the likelihood converges to a finite value while at least one parameter diverges (Heinze and Schemper 2001).

Firth (1993) recommended using the penalized likelihood $L^*(\boldsymbol{\beta}) = L(\boldsymbol{\beta})|\mathcal{I}(\boldsymbol{\beta})|^{0.5}$ to reduce the first-order bias in estimating the canonical parameters of an exponential family model, where $L(\boldsymbol{\beta})$ and $\mathcal{I}(\boldsymbol{\beta})$ are the unpenalized likelihood and information matrix, respectively.

Heinze (1999) and Heinze and Schemper (2001) applied the idea of Firth (1993) by maximizing the penalized partial log likelihood

$$l^*(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + 0.5 \log(|\mathcal{I}(\boldsymbol{\beta})|)$$

to obtain estimates of regression parameters when a monotone likelihood is observed.

The score function $\mathbf{U}(\boldsymbol{\beta})$ is replaced by the penalized score function, $\mathbf{U}^*(\boldsymbol{\beta}) \equiv (U^*(\beta_1), \ldots, U^*(\beta_p))'$, where

$$U^*(\beta_r) = U(\beta_r) + 0.5 \mathrm{tr}\left\{ \mathcal{I}^{-1}(\boldsymbol{\beta}) \frac{\partial \mathcal{I}(\boldsymbol{\beta})}{\partial \beta_r} \right\} \qquad r = 1, \ldots, p$$

The Firth estimate is obtained iteratively as

$$\boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} + \mathcal{I}^{-1}(\boldsymbol{\beta}^{(s)}) \mathbf{U}^*(\boldsymbol{\beta}^{(s)})$$

Although the estimated regression parameters, $\hat{\boldsymbol{\beta}}$, are obtained by maximizing the penalized partial likelihood, the Taylor series linearized variance estimator uses the score residuals and the information matrix from the unpenalized likelihood that are evaluated at $\hat{\boldsymbol{\beta}}$. For more information, see the section "Taylor Series Linearization" on page 9938.

The replication variance estimation methods use the replicated version of the penalized score function to obtain replicate estimates, $\hat{\boldsymbol{\beta}}^{(r)}$, for the regression parameters. The replicate estimates are then used in the replication variance estimation, as described in the sections "Balanced Repeated Replication (BRR) Method" on page 9939, "Bootstrap Method" on page 9941, "Jackknife Method" on page 9942, and "Replicate Weights Method" on page 9944.

Explicit Formulas for the Score Function, Fisher Information, and Partial Derivatives for the Information Matrix

Using the notation in the sections "Notation and Estimation" on page 9923 and "Partial Likelihood Function for the Cox Model" on page 9928, the Breslow unpenalized log partial likelihood is given by

$$l(\boldsymbol{\beta}) = \log(L_{\text{Breslow}}(\boldsymbol{\beta})) = \sum_{k=1}^{K}\left\{\boldsymbol{\beta}' \sum_{hij\in\mathcal{D}_k} \tilde{w}_{hij}\mathbf{Z}_{hij}(t_k) - \Big(\sum_{hij\in\mathcal{D}_k} \tilde{w}_{hij}\Big)\log \sum_{hij\in\mathcal{R}_k} \tilde{w}_{hij}\exp\big(\boldsymbol{\beta}'\mathbf{Z}_{hij}(t_k)\big)\right\}$$

where $\tilde{w}_{hij} = n\big(\sum_{hij\in\mathcal{A}} w_{hij}\big)^{-1} w_{hij}$; n is the number of observation units; and w_{hij} is the weight for unit j in PSU i and stratum h.

Denote

$$\mathbf{S}_k^{(a)}(\boldsymbol{\beta}) = \sum_{hij\in\mathcal{R}_k} \tilde{w}_{hij}\exp\big(\boldsymbol{\beta}'\mathbf{Z}_{hij}(t_k)\big)[\mathbf{Z}_{hij}(t_k)]^{\otimes a}$$

where $a = 0, 1, 2$.

Then the score function is given by

$$\begin{aligned}\mathbf{U}(\boldsymbol{\beta}) &\equiv (U(\beta_1),\ldots,U(\beta_p))' \\ &= \frac{\partial l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} \\ &= \sum_{k=1}^{K}\left\{\sum_{hij\in\mathcal{D}_k}\tilde{w}_{hij}\mathbf{Z}_{hij}(t_k) - \sum_{hij\in\mathcal{D}_k}\tilde{w}_{hij}\frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^0(\boldsymbol{\beta})}\right\}\end{aligned}$$

and the Fisher information matrix is given by

$$\begin{aligned}\mathcal{I}(\boldsymbol{\beta}) &= -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^2} \\ &= \sum_{k=1}^{K}\sum_{hij\in\mathcal{D}_k}\tilde{w}_{hij}\left\{\frac{\mathbf{S}_k^{(2)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} - \left[\frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{\mathbf{S}_k^{(0)}(\boldsymbol{\beta})}\right]\left[\frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{\mathbf{S}_k^{(0)}(\boldsymbol{\beta})}\right]'\right\}\end{aligned}$$

Denote

$$\mathbf{Q}_{kr}^{(a)}(\boldsymbol{\beta}) = \sum_{hij\in\mathcal{R}_k}\tilde{w}_{hij}\exp\big(\boldsymbol{\beta}'\mathbf{Z}_{hij}(t_k)\big)Z_{hij,r}(t_k)[\mathbf{Z}_{hij}(t_k)]^{\otimes a}$$

where $a = 0, 1, 2$; $r = 1,\ldots,p$; and $\mathbf{Z}_{hij}(t) = (Z_{hij,1}(t),\ldots,Z_{hij,p}(t))$. Then

$$\begin{aligned}\frac{\partial\mathcal{I}(\boldsymbol{\beta})}{\partial\beta_r} = \sum_{k=1}^{K}\sum_{hij\in\mathcal{D}_k}\tilde{w}_{hij}\Bigg\{ &\left[\frac{\mathbf{Q}_{kr}^{(2)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} - \frac{\mathbf{Q}_{kr}^{(0)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})}\frac{\mathbf{S}_k^{(2)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})}\right] \\ &-\left[\frac{\mathbf{Q}_{kr}^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} - \frac{\mathbf{Q}_{kr}^{(0)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})}\frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})}\right]\left[\frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})}\right]' \\ &-\left[\frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})}\right]\left[\frac{\mathbf{Q}_{kr}^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} - \frac{\mathbf{Q}_{kr}^{(0)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})}\frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})}\right]' \Bigg\}\end{aligned}$$

where $r = 1,\ldots,p$.

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, WEIGHT, and REPWEIGHTS statements, and with the RATE= or TOTAL= option in the PROC SURVEYPHREG statement.

If you provide replicate weights for replication 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 9937. 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 replication 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 replication 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 9935 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 or bootstrap variance estimation, you can specify either the sampling rate or the population total by using the RATE= or TOTAL= option, respectively, 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 or bootstrap 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 or bootstrap variance estimation does not include a finite population correction. For fairly small sampling fractions, this correction is often ignored. For more information, see Cochran (1977) and Kish (1965).

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=*value* option, *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.

Replicate Weights

If you have replicate weights available for your survey data, use the REPWEIGHTS statement to name the variables that contain the replicate weights. Replicate weights must be positive numbers. If an observation has a replicate weight that is nonpositive or missing, then PROC SURVEYPHREG does not perform any analyses. For more information, see the section "Missing Values" on page 9935.

If you have replicate coefficients available for your survey data, use the REPCOEFS= option in the REPWEIGHTS statement to specify the replicate coefficients.

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 *complete respondent*, and an observation whose response to one or more survey items are missing is called an *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 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 WEIGHT variable, then PROC SURVEYPHREG excludes that observation from the analysis.

REPWEIGHTS Variables

If you provide replicate weights in a REPWEIGHTS statement for replication variance estimation, all REPWEIGHTS variable values must be nonmissing. Similarly, if you provide jackknife coefficients in the JKCOEFS= option in the REPWEIGHTS statement, all values of the 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 CLASS, STRATA, CLUSTER, or DOMAIN variable, unless you specify the MISSING option in the PROC SURVEYPHREG statement. If you specify the MISSING option, the procedure treats missing values as a valid (nonmissing) category for all categorical variables, which include STRATA variables, CLUSTER variables, CLASS variables, and DOMAIN variables.

Analysis Variables

By default, 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 MODEL statement, unless you specify the MISSING or NOMCAR option in the PROC SURVEYPHREG statement. When the procedure excludes observations with missing values from analyses, it displays the total frequency of observations used in the 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 9935 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 that have missing values in the WEIGHT, FREQ, CLUSTER, STRATA, or DOMAIN variables. Observations that have missing values of the WEIGHT and FREQ variables are always excluded from the analysis. Observations that have 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 by specifying VARMETHOD=BOOTSTRAP, VARMETHOD=BRR, or VARMETHOD=JACKKNIFE, do not use the NOMCAR option.

Degrees of Freedom

PROC SURVEYPHREG uses the degrees of freedom of the variance estimator to obtain t confidence limits and Wald-type F tests. The procedure computes the degrees of freedom based on the variance estimation method, the sample design, and the number of estimable parameters. For more information, see the section "Degrees of Freedom" on page 9945. This section describes how missing values can affect the computation of the degrees of freedom.

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 that are included in the analysis. These numbers do not count observations that are excluded from the analysis because they have 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 the degrees of freedom. If you specify the NOMCAR option for Taylor series variance estimation, observations that have missing values for variables in the regression model are included in computing the 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 9923 for definitions of the notation used in this section. Let

$$S^{(r)}(\boldsymbol{\beta},t) = \sum_{A} w_{hij}\, y_{hij}(t) \exp\left(\boldsymbol{\beta}'\mathbf{Z}_{hij}(t)\right) \mathbf{Z}_{hij}^{\otimes r}(t)$$

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

$$\mathbf{a}^{\otimes r} = \begin{cases} \mathbf{a}\mathbf{a}' & , \quad r = 1 \\ I_{\dim(\mathbf{a})} & , \quad r = 0 \end{cases}$$

and let $I_{\dim(\mathbf{a})}$ be the identity matrix of appropriate dimension.

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

$$\begin{aligned} \mathbf{L}_{hij}(\boldsymbol{\beta}) &= \Delta_{hij}\left\{\mathbf{Z}_{hij}(t_{hij}) - \bar{\mathbf{Z}}(\boldsymbol{\beta},t_{hij})\right\} \\ &\quad - \sum_{(h\prime,i\prime,j\prime)\in A} \Delta_{h\prime i\prime j\prime} \frac{w_{h\prime i\prime j\prime} Y_{hij}(t_{h\prime i\prime j\prime}) \exp\left(\boldsymbol{\beta}'\mathbf{Z}_{hij}(t_{h\prime i\prime j\prime})\right)}{S^{(0)}(\boldsymbol{\beta},t_{h\prime i\prime j\prime})} \left\{\mathbf{Z}_{hij}(t_{h\prime i\prime j\prime}) - \bar{\mathbf{Z}}(\boldsymbol{\beta},t_{h\prime i\prime j\prime})\right\} \end{aligned}$$

For TIES=EFRON, the computation of the score residuals is modified to comply with the Efron partial likelihood. See the section "Residuals" on page 9951 for more information.

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

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

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

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

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

$$\begin{aligned} \mathbf{e}_{hij} &= w_{hij}\mathbf{L}_{hij}(\hat{\boldsymbol{\beta}}) \\ \mathbf{e}_{hi+} &= \sum_{j=1}^{m_{hi}} \mathbf{e}_{hij} \\ \bar{\mathbf{e}}_{h\cdot\cdot} &= \frac{1}{n_h} \sum_{i=1}^{n_h} \mathbf{e}_{hi+} \end{aligned}$$

The factor $(n-1)/(n-p)$ in the computation of the matrix $\mathbf{G}$ 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 rth ($r = 1, 2, ..., R$) replicate is drawn from the full sample according to the rth 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 rth 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 rth 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 $\hat{\boldsymbol{\beta}}$ be the estimated proportional hazards regression coefficients from the full sample, and let $\hat{\boldsymbol{\beta}}_r$ be the estimated proportional hazards regression coefficients from the rth replicate by using replicate weights. PROC SURVEYPHREG estimates the covariance matrix of $\hat{\boldsymbol{\beta}}$ by

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

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 $\hat{\boldsymbol{\beta}}$ by

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

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

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

If one or more components of $\hat{\boldsymbol{\beta}}_r$ 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 $\hat{\boldsymbol{\beta}}_r$ might not be available for a replicate sample even if $\hat{\boldsymbol{\beta}}$ is defined for the full sample. Let R_a be the number of replicates where $\hat{\boldsymbol{\beta}}_r$ is available, and let $R - R_a$ be the number of replicates where $\hat{\boldsymbol{\beta}}_r$ is not available. Without loss of generality, assume that $\hat{\boldsymbol{\beta}}_r$ is available only for the first R_a replicates; then the BRR variance estimator is

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

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, ϵ ($0 \leq \epsilon < 1$), and instead of deleting one PSU per stratum, it multiplies the original weight by the coefficient ϵ. 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 ϵ 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 ($\hat{\boldsymbol{\beta}}_r$) might be undefined even when the full sample estimator ($\hat{\boldsymbol{\beta}}$) is defined. Fay's BRR method is especially useful for this situation because it uses all the sampled units in every replicate.

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

$$\widehat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = \frac{1}{R(1-\epsilon)^2}\sum_{r=1}^{R}\left(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}\right)'$$

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 $\mathbf{A}$ of dimension R is a square matrix that has all elements equal to 1 or –1 such that $\mathbf{A}'\mathbf{A} = R\mathbf{I}$, where $\mathbf{I}$ 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 $k = 8$:

$$\begin{array}{rrrrrrrr}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\
1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\
1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\
1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\
1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\
1 & -1 & -1 & 1 & -1 & 1 & 1 & -1
\end{array}$$

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 9939 for details about how the Hadamard matrix is used to construct replicates for BRR variance estimation.

Bootstrap Method

The naive bootstrap variance estimator that is suitable for infinite population is not consistent when applied to complex surveys. Bootstrap replicate samples for complex surveys are created by using a simple random sample with replacement of primary sampling units (PSUs) within each stratum. PSUs in different strata are sampled independently. The original sampling weights are then adjusted in each replicate to reflect the full sample. These adjusted weights are also called bootstrap replicate weights. McCarthy and Snowden (1985), Rao and Wu (1988), Sitter (1992b), and Sitter (1992a) provide several adjusted bootstrap variance estimators that are consistent for complex surveys. For more information about bootstrap variance estimation for complex surveys, see Mashreghi, Haziza, and Léger (2016), Beaumont and Patak (2012), Lohr (2010, Section 9.3.3), Fuller (2009, Section 4.5), Wolter (2007, Chapter 5), and Shao and Tu (1995, Section 6.2.4).

If you do not provide replicate weights by using the REPWEIGHTS statement, then the BOOTSTRAP option in the PROC SURVEYPHREG statement creates bootstrap replicate weights for you. This bootstrap method is similar to the method of Rao, Wu, and Yue (1992) and is also known as bootstrap weights method (Mashreghi, Haziza, and Léger 2016).

Each replicate is obtained by selecting a simple random sample with replacement of m_h PSUs from stratum h. The rth bootstrap replicate weight for observation unit j in PSU i and stratum h is given by

$$w_{hij}^{(r)} = w_{hij}\left\{1 - \sqrt{(1-f_h)m_h/(n_h-1)} + \sqrt{(1-f_h)m_h/(n_h-1)}(n_h/m_h)k_{hi}^{(r)}\right\}$$

where $k_{hi}^{(r)}$ is the number of times PSU i is selected in replicate sample r, and f_h is the sampling fraction in stratum h.

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

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

with $\sum_{h=1}^{H} n_h - H$ degrees of freedom, where n_h is the number of PSUs in stratum h, and H is the number of strata.

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

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

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

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

If one or more components of $\hat{\boldsymbol{\beta}}_r$ 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 $\hat{\boldsymbol{\beta}}_r$ might not be available for a replicate sample even if $\hat{\boldsymbol{\beta}}$ is defined for the full sample. Let R_a be the number of replicates where $\hat{\boldsymbol{\beta}}_r$ is available, and let $R - R_a$ be the number of replicates where $\hat{\boldsymbol{\beta}}_r$ is not available. Without loss of generality, assume that $\hat{\boldsymbol{\beta}}_r$ is available only for the first R_a replicates; then the bootstrap variance estimator is

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

with degrees of freedom equal to the minimum of $\sum_{h=1}^{H} n_h - H$ and R_a, where n_h is the number of PSUs in stratum h, and H is the number of strata.

Although PROC SURVEYPHREG creates bootstrap weights only from the bootstrap weights method (Rao, Wu, and Yue 1992), bootstrap weights that are generated from any bootstrap methods can be used in the REPWEIGHTS statement. If the bootstrap replicate weights are available to you for a survey, then you can use the REPWEIGHTS statement to name the variables that contain the bootstrap replicate weights and specify the VARMETHOD=BOOTSTRAP option in the PROC SURVEYPHREG statement. The SURVEYPHREG procedure uses $1/R$ as the default bootstrap replicate coefficient when you specify the VARMETHOD=BOOTSTRAP option, where R is the total number of replicates. Alternatively, you can specify different replicate coefficients by using the REPCOEFS= option in the REPWEIGHTS statement.

For more information, see the section "Replicate Weights Method" on page 9944.

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* α_r. The modified weights are called replicate weights.

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

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

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

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

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 $\hat{\boldsymbol{\beta}}$ by

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

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

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

If one or more components of $\hat{\boldsymbol{\beta}}_r$ 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 $\hat{\boldsymbol{\beta}}_r$ might not be available for a replicate sample even if $\hat{\boldsymbol{\beta}}$ is defined for the full sample. Let R_a be the number of replicates where $\hat{\boldsymbol{\beta}}_r$ are available, and let $R - R_a$ be the number of replicates where $\hat{\boldsymbol{\beta}}_r$ are not available. Without loss of generality, assume that $\hat{\boldsymbol{\beta}}_r$ is available only for the first R_a replicates; then the jackknife variance estimator is

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

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 9946 for details.

Replicate Weights Method

The replicate weights variance estimation method is a general-purpose variance estimation method that uses the replicate weights and replicate coefficients that you provide by using the REPWEIGHTS statement and the REPCOEFS= option, respectively.

If you provide your own replicate weights in a REPWEIGHTS statement but do not specify replicate coefficients in a REPCOEFS= option, then the default replicate coefficient depends on the VARMETHOD= option in the PROC SURVEYPHREG statement as shown in the following table:

Value of VARMETHOD=	Default Replicate Coefficient
None specified	$(R-1)/R$
BOOTSTRAP	$1/R$
BRR	$1/R$
JACKKNIFE	$(R-1)/R$

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

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

with R degrees of freedom, where R is the number of replicates.

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

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

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

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

If one or more components of $\hat{\boldsymbol{\beta}}_r$ 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 $\hat{\boldsymbol{\beta}}_r$ might not be available for a replicate sample even if $\hat{\boldsymbol{\beta}}$ is defined for the full sample. Let R_a be the number of replicates where $\hat{\boldsymbol{\beta}}_r$ are available, and let $R - R_a$ be the number of replicates where $\hat{\boldsymbol{\beta}}_r$ are not available. Without loss of generality, assume that $\hat{\boldsymbol{\beta}}_r$ are available only for the first R_a replicates; then the jackknife variance estimator is

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

with R_a degrees of freedom. Alternatively, you can use the VADJUST=AVGREPSS option in the MODEL statement to use the average sum of squares for the invalid replicate samples. For more information, see "Variance Adjustment Factors" on page 9946.

Degrees of Freedom

PROC SURVEYPHREG uses the degrees of freedom of the variance estimator to obtain *t* confidence limits and Wald-type *F* tests. The procedure computes the degrees of freedom based on the variance estimation method, the sample design, and the number of estimable parameters. Alternatively, you can specify the degrees of freedom by using the DF= option in the MODEL statement. This section describes how PROC SURVEYPHREG computes different values of the degrees of freedom based on the variance estimation method and the sample design. For more information about how degrees of freedom depend on the number of estimable parameters and the DF= option in the MODEL statement, see the section "Hypothesis Tests, Confidence Intervals, and Residuals" on page 9948.

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 that are included in the analysis. These numbers do not count observations that are excluded from the analysis because they have missing values. For more information, see the section "Missing Values" on page 9935. If you specify the MISSING option in the CLASS statement, missing values are treated as valid nonmissing levels and are included in computing the degrees of freedom. If you specify the NOMCAR option for Taylor series variance estimation, observations that have missing values of the analysis variables are included in computing the degrees of freedom.

If you provide replicate weights by using the REPWEIGHTS statement, the degrees of freedom is equal to the number of replicates used, which is the number of REPWEIGHTS variables that provide replicate estimates. Alternatively, you can specify DF=ALLREPS in the MODEL statement to specify that *df* equals the number of replicates. For bootstrap variance estimation, the number of replicates is typically much larger than the number of independent variance units. Use the DF= option in the MODEL statement to specify that *df* equals the number of independent variance units from your survey design.

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

For bootstrap variance estimation (when you do not use the REPWEIGHTS statement), PROC SURVEYPHREG calculates the degrees of freedom as the number of clusters minus the number of strata. If you do not specify the CLUSTER statement, then the procedure treats each observation as a cluster. If you do not specify the STRATA statement, then the procedure assumes that all observations are in the same stratum. For bootstrap 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 they cannot be used for parameter estimation because of factors such as nonconvergence or inestimability, then *df* equals the lesser of the number of clusters minus the number of strata and the number of replicate samples that produces parameter estimates. Alternatively, you can specify DF=ALLREPS in the MODEL statement to specify that *df* equals the number of clusters minus the number of strata.

For jackknife variance estimation (when you do not use the REPWEIGHTS statement), PROC SURVEYPHREG calculates the degrees of freedom as the number of clusters minus the number of strata. If you do not specify the CLUSTER statement, then the procedure treats each observation as a cluster. If you do not specify the STRATA statement, 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 they cannot be used for parameter estimation because of factors such as nonconvergence or inestimability, then *df* equals the number of clusters (or observations if no CLUSTER statement is specified) minus the number of strata (or 1 if no STRATA statement is specified) minus the number of replicate samples that are not used. Alternatively, you can specify DF=ALLREPS in the MODEL statement to specify that *df* equals the number of clusters minus the number of strata.

Variance Adjustment Factors

PROC SURVEYPHREG provides options for adjusting the default variance estimators. VADJUST=NONE and VADJUST=DF are available for the Taylor series linearization variance estimator. 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 $(n-1)/(n-p)$ 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 $\mathbf{G}$ for the Taylor series linearization variance estimation. If you do not want to use this adjustment, then specify 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

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

as the jackknife variance estimator. An alternative estimator is

$$\begin{aligned}\widehat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) &= \sum_{r=1}^{R_a} \alpha_r \left(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}\right)' + (R - R_a)\left\{\frac{1}{R_a}\sum_{r=1}^{R_a} \alpha_r \left(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}\right)'\right\} \\ &= \frac{R}{R_a}\sum_{r=1}^{R_a} \alpha_r (\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}})'\end{aligned}$$

which uses the average replicate sum of squares for the $R - R_a$ unusable replicate samples. If you specify the VADJUST=AVGREPSS option, PROC SURVEYPHREG uses the second variance estimator for the jackknife replication method. Note that you can specify the 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 the following variance ratios and standard error ratios:

- If you specify the 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 jth diagonal element of the observed information matrix $\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})$.

- If you specify the VARRATIO=IND option, then the procedure computes the variance ratio of the estimated regression parameter $\hat{\beta}_j$ as $\frac{\hat{V}(\hat{\beta}_j)}{\hat{V}_{\mathrm{IND}}(\hat{\beta}_j)}$, where $\hat{V}(\hat{\beta}_j)$, the estimated variance of $\hat{\beta}_j$, uses the complete design information and $\hat{V}_{\mathrm{IND}}(\hat{\beta}_j)$ is the jth diagonal element of $\hat{V}_{\mathrm{IND}}(\hat{\boldsymbol{\beta}})$. $\hat{V}_{\mathrm{IND}}(\hat{\boldsymbol{\beta}})$ is the sandwich variance estimator, which ignores the strata and the clusters and is computed as

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

  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).

- If you specify the VARRATIO=SRSWR option, then the procedure computes the variance ratio of the estimated regression parameter $\hat{\beta}_j$ as $\frac{\hat{V}(\hat{\beta}_j)}{\hat{V}_{\mathrm{SRSWR}}(\hat{\beta}_j)}$, where $\hat{V}(\hat{\beta}_j)$, the estimated variance of $\hat{\beta}_j$, uses the complete design information and $\hat{V}_{\mathrm{SRSWR}}(\hat{\beta}_j)$ is the jth diagonal element of $\hat{V}_{\mathrm{SRSWR}}(\hat{\boldsymbol{\beta}})$. $\hat{V}_{\mathrm{SRSWR}}(\hat{\boldsymbol{\beta}})$ is given by

$$\hat{V}_{\mathrm{SRSWR}}(\hat{\boldsymbol{\beta}}) = \mathcal{I}^{-1}(\hat{\boldsymbol{\beta}}) \sum_h \sum_i \sum_j w_{hij}/n$$

  where w_{hij} is the observation weight for unit (h, i, j) 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). The matrix $[\hat{V}_{\mathrm{SRSWR}}(\hat{\boldsymbol{\beta}})]^{-1}\hat{V}(\hat{\boldsymbol{\beta}})$ is often called the generalized design effect matrix (Rao, Scott, and Skinner 1998).

- If you specify the VARRATIO=SRSWOR option, then the procedure computes the variance ratio of the estimated regression parameter $\hat{\beta}_j$ as $\frac{\hat{V}(\hat{\beta}_j)}{\hat{V}_{\mathrm{SRSWOR}}(\hat{\beta}_j)}$, where $\hat{V}(\hat{\beta}_j)$, the estimated variance of $\hat{\beta}_j$, uses the complete design information and $\hat{V}_{\mathrm{SRSWOR}}(\hat{\beta}_j)$ is the jth diagonal element of $\hat{V}_{\mathrm{SRSWOR}}(\hat{\boldsymbol{\beta}})$. $\hat{V}_{\mathrm{SRSWOR}}(\hat{\boldsymbol{\beta}})$ is given by

$$\hat{V}_{\mathrm{SRSWR}}(\hat{\boldsymbol{\beta}}) = (1-f)\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}}) \sum_h \sum_i \sum_j w_{hij}/n$$

  where w_{hij} is the observation weight for unit (h, i, j), 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 or bootstrap 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 are available to test the global null hypothesis $H_0: \boldsymbol{\beta} = \mathbf{0}$. Let d be the usual degrees of freedom computed from the survey data by using the number of strata, clusters, or replicate weights;

and let p be the number of estimable parameters in the null hypothesis H_0. For more information about computing d, see the section "Degrees of Freedom" on page 9945.

The unadjusted likelihood ratio test statistic is expressed as

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

where $\mathrm{L}(\cdot)$ denotes the partial pseudo-likelihood that is described in the section "Partial Likelihood Function for the Cox Model" on page 9928 and $\hat{\boldsymbol{\beta}}$ denotes the estimated regression parameters. The p-value for the unadjusted test is computed by using a chi-square distribution with p degrees of freedom.

The unadjusted likelihood ratio statistic is sensitive to the scaling of the weights. PROC SURVEYPHREG computes an adjusted likelihood ratio test statistic that is invariant to the scaling of the weights. The adjusted test is similar to the second-order adjusted Rao-Scott chi-square tests. For more information, see Rao, Scott, and Skinner (1998), and Lumley and Scott (2013). The adjusted likelihood ratio test statistic is expressed as

$$\chi^2_{\mathrm{RS2}} = \frac{(n/\hat{N})\chi^2_{\mathrm{LR}}}{\bar{\delta}_{\cdot}(1+\hat{a}^2)}$$

where $\delta_1, \delta_2, \ldots, \delta_r$ are the positive eigenvalues from the generalized design effect matrix ("Variance Ratios and Standard Error Ratios" on page 9947), $\bar{\delta}_{\cdot} = 1/r \sum_{i=1}^{r} \delta_i$ is the mean of the positive eigenvalues, $\hat{a}^2 = (r-1)^{-1} \sum_{i=1}^{r} (\delta_i - \bar{\delta}_{\cdot})^2 / \bar{\delta}_{\cdot}^2$ is the squared coefficient of variations of the positive eigenvalues, n is the number of observation units, and $\hat{N} = \sum_{hij} w_{hij}$ is the sum of the weights over all observation units. The p-value for the adjusted test is computed by using a chi-square distribution with $r/(1+\hat{a}^2)$ degrees of freedom.

The usual assumptions that are required for a likelihood ratio test do not hold for the pseudo-likelihood that is used by PROC SURVEYPHREG (Rao, Scott, and Skinner 1998), leading to other methods of testing the global null hypothesis, such as the Wald test discussed in the following paragraph.

The Wald test uses the variance estimator that accounts for complex sampling such as stratification, clustering, and unequal weighting. Let $Q = \hat{\boldsymbol{\beta}}' \left[\widehat{\mathbf{V}}(\hat{\boldsymbol{\beta}})\right]^{-1} \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the estimated regression parameters and $\widehat{\mathbf{V}}(\hat{\boldsymbol{\beta}})$ is the estimated covariance matrix for $\hat{\boldsymbol{\beta}}$. You can request the Wald tests that are described in the following table by using the DF= option in the MODEL statement.

Value of DF=	Test Request	Test Statistic	Numerator Degrees of Freedom	Denominator Degrees of Freedom
NONE	Chi-square	Q	p	∞
v	Customized F	vQ/pd	p	v
DESIGN	Unadjusted F	Q/p	p	d
DESIGN (v)	Unadjusted F	Q/p	p	v
PARMADJ	Adjusted F	$(d-p+1)Q/pd$	p	$d-p+1$
PARMADJ (v)	Adjusted F	$(v-p+1)Q/pv$	p	$v-p+1$
DESIGNADJ	Adjusted F	$(d-p+1)Q/pd$	p	d

Model Fit Statistics

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

- –2 log likelihood:

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

  where $L(.)$ 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 9928, and $\hat{\boldsymbol{\beta}}$ is the maximum pseudo-log-likelihood estimate of the proportional hazards regression coefficients.

- Akaike's information criterion (AIC):

$$\text{AIC} = -2 \text{ Log L} + 2p$$

The AIC statistic provides a different way of adjusting the log-likelihood statistic for the number of estimable parameters in the model.

Neither of these criteria is adjusted for the complex sample design, and both criteria are sensitive to the scale of the weights.

Contrasts

For a testable hypothesis $H_0: \mathbf{L}\boldsymbol{\beta} = 0$, you can request different Wald tests by using the DF= option in the MODEL statement.

Let

$$Q = (\mathbf{L}^*\hat{\boldsymbol{\beta}})'(\mathbf{L}^{*\prime}\widehat{\mathbf{V}}\mathbf{L}^*)^{-1}(\mathbf{L}^*\hat{\boldsymbol{\beta}})$$

where $\mathbf{L}$ is a contrast vector or matrix that you specify, $\boldsymbol{\beta}$ is the vector of regression parameters, $\hat{\boldsymbol{\beta}}$ is the estimated regression coefficients, $\widehat{\mathbf{V}}$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$, and $\mathbf{L}^*$ is a matrix such that the following are true:

- $\mathbf{L}^*$ 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 that is computed by using the $\mathbf{L}^*$ matrix is equivalent to the Wald F statistic computed by using the $\mathbf{L}$ matrix.

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 such an $\mathbf{L}^*$ matrix cannot be constructed for a given set of linear contrasts, in which case the contrasts are not testable. Let r be the rank of $\mathbf{L}$. The following table describes the Wald tests available in PROC SURVEYPHREG.

Value of DF=	Test Request	Test Statistic	Numerator Degrees of Freedom	Denominator Degrees of Freedom
NONE	Chi-square	Q	r	∞
v	Customized F	vQ/rd	r	v
DESIGN	Unadjusted F	Q/r	r	d
DESIGN (v)	Unadjusted F	Q/r	r	v
PARMADJ	Adjusted F	$(d-r+1)Q/rd$	r	$d-r+1$
PARMADJ (v)	Adjusted F	$(v-r+1)Q/rv$	r	$v-r+1$
DESIGNADJ	Adjusted F	Q/r	r	d

Confidence Intervals

By default, the SURVEYPHREG procedure computes t confidence limits for the estimated regression coefficients. Alternatively, you can specify DF=NONE in the MODEL statement to request standard normal confidence intervals. The t confidence interval for a linear combination $\mathbf{l}'\boldsymbol{\beta}$ of the regression coefficients is computed as

$$\left(\mathbf{l}'\hat{\boldsymbol{\beta}} \pm t_{df,\alpha/2}\sqrt{\mathbf{l}'\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})\mathbf{l}}\right)$$

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 9945 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 $\beta_j = \mathbf{e}_j'\boldsymbol{\beta}$ is defined as $\exp(\beta_j)$, where $\mathbf{e}_j$ denotes the jth unit vector. In general, a log-hazard ratio can be written as $\mathbf{l}'\boldsymbol{\beta}$, a linear combination of the regression coefficients, and the hazard ratio $\exp(\mathbf{l}'\boldsymbol{\beta})$ is obtained by replacing e_j with $\mathbf{l}$.

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

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

where $t_{df,\alpha/2}$ is the $100(1-\alpha/2)$ percentile point of the t distribution with df degrees of freedom. For more information about hazard ratios see the section "Hazard Ratios" on page 9954, and for more information about df see the section "Degrees of Freedom" on page 9945. 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 9923 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

$$S^{(r)}(\boldsymbol{\beta}, t) = \sum_{A} w_{hij}\, y_{hij}(t) \exp\left(\boldsymbol{\beta}' \mathbf{Z}_{hij}(t)\right) \mathbf{Z}_{hij}^{\otimes r}(t)$$

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

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

Further let

$$\begin{aligned} d\Lambda_0(\boldsymbol{\beta}, t) &= \sum_{A} \frac{w_{hij}\, dn_{hij}(t)}{S^{(0)}(\boldsymbol{\beta}, t)} \\ dM_{hij}(\boldsymbol{\beta}, t) &= dn_{hij}(t) - y_{hij}(t) \exp\left(\boldsymbol{\beta}' \mathbf{Z}_{hij}(t)\right) d\Lambda_0(\boldsymbol{\beta}, t) \end{aligned}$$

The martingale residual at t is defined as

$$\begin{aligned} \hat{M}_{hij}(t) &= \int_0^t dM_{hij}(\hat{\boldsymbol{\beta}}, \tau) \\ &= n_{hij}(t) - \int_0^t y_{hij}(\tau) \exp\left(\hat{\boldsymbol{\beta}}' \mathbf{Z}_{hij}(\tau)\right) d\Lambda_0(\hat{\boldsymbol{\beta}}, \tau) \end{aligned}$$

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 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

$$\hat{M}_{(h,i,j)} = \Delta_{(h,i,j)} - \mathrm{e}^{\hat{\boldsymbol{\beta}}' \mathbf{Z}_{(h,i,j)}} \int_0^{t_{(h,i,j)}} d\Lambda_0(\hat{\boldsymbol{\beta}}, s)$$

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

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

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

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

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

$$\hat{\mathbf{U}}_{hij}(t_{hij,k}) = \mathbf{Z}_{hij}(t_{hij,k}) - \bar{\mathbf{Z}}(\hat{\boldsymbol{\beta}}, t_{hij,k})$$

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

$$\mathbf{L}_{hij}(\boldsymbol{\beta}, t) = \int_0^t [\mathbf{Z}_{hij}(\tau) - \bar{\mathbf{Z}}(\boldsymbol{\beta}, \tau)] dM_{hij}(\boldsymbol{\beta}, \tau)$$

The vector $\hat{\mathbf{L}}_{hij} \equiv \mathbf{L}_{hij}(\hat{\boldsymbol{\beta}}, \infty)$ 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

$$\begin{aligned}
S^{(r)}(\boldsymbol{\beta}, l, t) &= \sum_A w_{hij}\, y_{hij}(t) \left\{ 1 - \frac{l-1}{d(t)} \delta_{hij}(t) \right\} \exp\left(\boldsymbol{\beta}' \mathbf{Z}_{hij}(t)\right) \mathbf{Z}_{hij}^{\otimes r}(t) \\
\bar{\mathbf{Z}}(\boldsymbol{\beta}, l, t) &= \frac{S^{(1)}(\boldsymbol{\beta}, l, t)}{S^{(0)}(\boldsymbol{\beta}, l, t)} \\
d\Lambda_0(\boldsymbol{\beta}, l, t) &= \sum_A \frac{w_{hij}\, dn_{hij}(t)}{S^{(0)}(\boldsymbol{\beta}, l, t)} \\
dM_{hij}(\boldsymbol{\beta}, l, t) &= dn_{hij}(t) - y_{hij}(t) \left(1 - \delta_{hij}(t) \frac{l-1}{d(t)} \right) \exp\left(\boldsymbol{\beta}' \mathbf{Z}_{hij}(t)\right) d\Lambda_0(\boldsymbol{\beta}, l, t)
\end{aligned}$$

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

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

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

$$\hat{\mathbf{U}}_{hij}(t_{hij,k}) = \mathbf{Z}_{hij}(t_{hij,k}) - \frac{1}{d(t_{hij,k})} \sum_{l=1}^{d(t_{hij,k})} \bar{\mathbf{Z}}(\hat{\boldsymbol{\beta}}, l, t_{hij,k})$$

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

$$\mathbf{L}_{hij}(\boldsymbol{\beta}, t) = \int_0^t \frac{1}{d(\tau)} \sum_{l=1}^{d(\tau)} \Big(\mathbf{Z}_{hij}(\tau) - \bar{\mathbf{Z}}(\boldsymbol{\beta}, l, \tau) \Big) dM_{hij}(\boldsymbol{\beta}, l, \tau)$$

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

Hazard Ratios

Consider a dichotomous risk factor variable X that takes the value 1 if the risk factor is present and 0 if the risk factor is absent. The log-hazard function is given by

$$\log[\lambda(t|X)] = \log[\lambda_0(t)] + \beta_1 X$$

where $\lambda_0(t)$ is the baseline hazard function.

The hazard ratio ψ is defined as the ratio of the hazard for those with the risk factor ($X = 1$) to the hazard for those without the risk factor ($X = 0$). The log of the hazard ratio is given by

$$\log(\psi) \equiv \log[\psi(X = 1, X = 0)] = \log[\lambda(t|X = 1)] - \log[\lambda(t|X = 0)] = \beta_1$$

In general, the hazard ratio can be computed by exponentiating the difference of the log-hazard between any two population profiles. This is the approach taken by the HAZARDRATIO statement, so the computations are available regardless of parameterization, interactions, and nestings. However, as shown in the preceding equation for $\log(\psi)$, hazard ratios of main effects can be computed as functions of the parameter estimates. The remainder of this section is concerned with this methodology.

The parameter β_1, which is associated with X, represents the change in the log-hazard rate from X=0 to X=1. So the hazard ratio is obtained by simply exponentiating the value of the parameter that is associated with the risk factor. The hazard ratio indicates how the hazard changes as you change X from 0 to 1. For example, $\psi = 2$ means that the hazard when X=1 is twice the hazard when X=0.

Suppose the values of the dichotomous risk factor are coded as constants a and b instead of 0 and 1. The hazard when X=a becomes $\lambda(t)\exp(a\beta_1)$, and the hazard when X=b becomes $\lambda(t)\exp(b\beta_1)$. The hazard ratio that corresponds to an increase in X from a to b is

$$\psi = \exp[(b-a)\beta_1] = [\exp(\beta_1)]^{b-a} \equiv [\exp(\beta_1)]^c$$

Note that for any a and b such that $c = b - a = 1$, $\psi = \exp(\beta_1)$. So the hazard ratio can be interpreted as the change in the hazard for any increase of one unit in the corresponding risk factor. However, the change in hazard for some amount other than one unit is often of greater interest. For example, a change of one pound in body weight might be too small to be considered important, whereas a change of 10 pounds might be more meaningful. The hazard ratio for a change in X from a to b is estimated by raising the hazard ratio estimate for a unit change in X to the power of $c = b - a$ as shown previously.

For a polytomous risk factor, the computation of hazard ratios depends on how the risk factor is parameterized. For illustration, suppose that Cell is a risk factor that has four categories: Adeno, Large, Small, and Squamous.

For the effect parameterization scheme (PARAM=EFFECT) with Squamous as the reference group, the design variables for Cell are as follows:

Cell	Design Variables for Cell X_1	X_2	X_3
Adeno	1	0	0
Large	0	1	0
Small	0	0	1
Squamous	–1	–1	–1

The log-hazard for Adeno is

$$\begin{aligned}\log[\lambda(t|\text{Adeno})] &= \log[\lambda_0(t)] + \beta_1(X_1 = 1) + \beta_2(X_2 = 0) + \beta_3(X_3 = 0) \\ &= \lambda_0(t) + \beta_1\end{aligned}$$

The log-hazard for Squamous is

$$\begin{aligned}\log[\lambda(t|\text{Squamous})] &= \log[\lambda_0(t)] + \beta_1(X_1 = -1) + \beta_2(X_2 = -1) + \beta_3(X_3 = -1)) \\ &= \log[\lambda_0(t)] - \beta_1 - \beta_2 - \beta_3\end{aligned}$$

Therefore, the log-hazard ratio of Adeno versus Squamous

$$\begin{aligned}\log[\psi(\text{Adeno}, \text{Squamous})] &= \log[\lambda(t|\text{Adeno})] - \log[\lambda(t|\text{Squamous})] \\ &= 2\beta_1 + \beta_2 + \beta_3\end{aligned}$$

For the reference cell parameterization scheme (PARAM=REF) with Squamous as the reference cell, the design variables for Cell are as follows:

Cell	Design Variables for Cell X_1	X_2	X_3
Adeno	1	0	0
Large	0	1	0
Small	0	0	1
Squamous	0	0	0

The log-hazard ratio of Adeno versus Squamous is given by

$$\begin{aligned}&\log(\psi(\text{Adeno}, \text{Squamous})) \\ &= \log[\lambda(t|\text{Adeno})] - \log[\lambda(t|\text{Squamous})] \\ &= (\log[\lambda_0(t)] + \beta_1(X_1 = 1) + \beta_2(X_2 = 0) + \beta_3(X_3 = 0)) - \\ &\quad (\log[\lambda_0(t)] + \beta_1(X_1 = 0) + \beta_2(X_2 = 0) + \beta_3(X_3 = 0)) \\ &= \beta_1\end{aligned}$$

For the GLM parameterization scheme (PARAM=GLM), the design variables for Cell are as follows:

Cell	Design Variables for Cell			
	X_1	X_2	X_3	X_4
Adeno	1	0	0	0
Large	0	1	0	0
Small	0	0	1	0
Squamous	0	0	0	1

The log-hazard ratio of Adeno versus Squamous is

$$\begin{aligned}
&\log(\psi(\text{Adeno, Squamous})) \\
&\quad = \log[\lambda(t|\text{Adeno})] - \log[\lambda(t|\text{Squamous})] \\
&\quad = \log[\lambda_0(t)] + \beta_1(X_1 = 1) + \beta_2(X_2 = 0) + \beta_3(X_3 = 0) + \beta_4(X_4 = 0)) - \\
&\qquad (\log(\lambda_0(t)) + \beta_1(X_1 = 0) + \beta_2(X_2 = 0) + \beta_3(X_3 = 0) + \beta_4(X_4 = 1)) \\
&\quad = \beta_1 - \beta_4
\end{aligned}$$

The fact that the log-hazard ratio ($\log(\psi) = \mathbf{l}^T\boldsymbol{\beta}$) is a linear function of the parameters enables the HAZARDRATIO statement to estimate the hazard ratio of the main effect even in the presence of interactions and nest effects. The log-hazard ratio is estimated by $\mathbf{l}^T\hat{\boldsymbol{\beta}}$, and the variance of the estimated log-hazard ratio is computed by $\mathbf{l}^T\hat{V}(\hat{\boldsymbol{\beta}})\mathbf{l}$, where $\hat{V}(\hat{\boldsymbol{\beta}})$ is the estimated variance for $\hat{\boldsymbol{\beta}}$.

To customize hazard ratios for specific units of change for a continuous risk factor, you can use the UNITS= option in a HAZARDRATIO statement to specify a list of relevant units for each explanatory variable in the model. Estimates of these customized hazard ratios are given in a separate table. Let (L_j, U_j) be a confidence interval for $\log(\psi)$. The corresponding lower and upper confidence limits for the customized hazard ratio $\exp(c\beta_j)$ are $\exp(cL_j)$ and $\exp(cU_j)$, respectively, (for $c > 0$), or $\exp(cU_j)$ and $\exp(cL_j)$, respectively, (for $c < 0$).

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 9961 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 9951 for details about how these statistics are calculated.

Replicate Weights Output Data Set

If you specify the OUTWEIGHTS= *method-option* for VARMETHOD=BOOTSTRAP, 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 9935 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 the 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 NLOPTIONS 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 9945; 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.

Risk Set Sum of Weights

The "Risk Set Sum of Weights" table is displayed if you specify the ATRISK option in the PROC SURVEYPHREG statement. The table displays, for each event time, the sum of weights of units at risk and the sum of weights of units that experience the event. The ODS name of the "Risk Set Sum of Weights" table is RiskSetInfo.

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 VARMETHOD=BOOTSTRAP, 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
- Bootstrap Seed, for VARMETHOD=BOOTSTRAP

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, VARMETHOD=BOOTSTRAP, or 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 SURVEYPHRG uses default names to identify the replicate weights. For more information, see the section "Replicate Weights Output Data Set" on page 9957.

Define AXTOL as the maximum absolute difference of the regression parameters between the last two iterations,

$$\text{AXTOL} = \max_j |\beta_j^{(t)} - \beta_j^{(t-1)}|$$

where $\beta_j^{(t)}$ is the jth component of $\boldsymbol{\beta}^{(t)}$ at iteration t.

The convergence status for a replicate sample is 1 if at least one convergence criterion for maximum likelihood estimation is satisfied, 2 if at least one convergence criterion for maximum likelihood estimation is satisfied but AXTOL is greater than 0.1, and 0 if no maximum likelihood convergence criteria are satisfied and thus the maximum likelihood estimates are not available in the replicate sample. 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.

Hazard Ratios for *label*

The "Hazard Ratios for *label*" table is displayed if you specify the HAZARDRATIO statement. The table displays the estimate and confidence limits for each hazard ratio. The ODS name of the "Hazard Ratios for *label*" table is HazardRatios.

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 119.9 lists the table names, along with the corresponding analysis options.

Table 119.9 ODS Tables Produced by PROC SURVEYPHREG

ODS Table Name	Description	Statement / Option
CensoredSummary	Summary of event and censored observations	Default
ClassLevelInfo	CLASS variable levels	CLASS
ConvergenceStatus	Convergence status	NLOPTIONS / PALL
CovB	Covariance of parameter estimates	MODEL / COVB
DesignSummary	Design summary	STRATA or CLUSTER
FitStatistics	Model fit statistics	Default

Table 119.9 *continued*

ODS Table Name	Description	Statement / Option
GlobalTests	Tests of the global null hypothesis	Default
HadamardMatrix	Hadamard matrix	PROC / VARMETHOD=BRR(PRINTH)
HazardRatios	Hazard ratios and confidence limits	HAZARDRATIO
Hessian	Observed Hessian matrix	MODEL / HESSIAN
InvHessian	Inverse Hessian matrix	MODEL / INVHESS
IterHist	Iteration history	NLOPTIONS / PHISTORY or PHISTPARMS
ModelInfo	Model information	Default
NObs	Number of observations	Default
ParameterEstimates	Maximum likelihood estimates	Default
ParameterEstimatesStart	Initial parameter values	NLOPTIONS / PALL
RepEstimates	Maximum likelihood estimates for replicate samples	PROC / VARMETHOD=BRR(DETAILS), VARMETHOD=JACKKNIFE(DETAILS) or VARMETHOD=BOOTSTRAP(DETAILS)
RiskSetInfo	Risk set information	PROC / ATRISK
StrataInfo	Stratum information	STRATA / LIST
VarianceEstimation	Variance estimation	Default
WeightedCensoredSummary	Summary of weighted number of event and censored observations	WEIGHT

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 623 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 622 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."

Examples: SURVEYPHREG Procedure

Example 119.1: 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. For more information about the data set and other alternative analyses, see Example 89.11 in Chapter 89, "The PHREG Procedure."

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:

```
proc format;
   value type 0='Juvenile' 1='Adult';
   value Rx  1='Laser' 0='Others';
run;

data Blind;
   format Treatment Rx.
          DiabeticType Type.;
   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
```

```
   ... more lines ...

1705  8.00 0 0 1 1705  8.00 0 0 0 1717 51.60 0 1 1 1717 42.33 1 1 0
1727 49.97 0 1 1 1727  2.90 1 1 0 1746 45.90 0 0 1 1746  1.43 1 0 0
1749 41.93 0 1 1 1749 41.93 0 1 0
;
```

The following 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. The HAZARDRATIO statement requests estimated hazard ratios for the treatments.

```
proc surveyphreg data=Blind;
   class Treatment DiabeticType / param=ref;
   model Time*Status(0) = Treatment DiabeticType Treatment*DiabeticType;
   cluster ID;
   hazardratio Treatment;
run;
```

Output 119.1.1 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 adjusted likelihood ratio test and the Wald test (Output 119.1.2), indicating that the survival time is highly dependent on Treatment and DiabeticType. In this example, the adjusted likelihood ratio statistic has an approximate chi-square distribution with 2.7 degrees of freedom, and the Wald statistic has an approximate F distribution with 3 numerator degrees of freedom and 194 denominator degrees of freedom. The denominator degrees of freedom is calculated as the number of clusters (197) minus the number of estimable parameters (3). Although the adjusted and the unadjusted likelihood ratio tests are very similar for this data set, the unadjusted likelihood ratio test does not account for clustering. It is recommended that you use the adjusted likelihood ratio test for clustered data.

Output 119.1.1 Summary Information

The SURVEYPHREG Procedure

Number of Observations Read	394
Number of Observations Used	394

Design Summary	
Number of Clusters	197

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
394	155	239	60.66

Variance Estimation	
Method	Taylor Series

Output 119.1.2 Global Test Results

Testing Global Null Hypothesis: BETA=0

Test	Test Statistic	Num DF	Den DF	p-Value
Likelihood Ratio (Unadj.)	28.4556	3	Infty	<.0001
Likelihood Ratio (Adj.)	28.1668	2.703	Infty	<.0001
Wald	11.4455	3	194	<.0001

Output 119.1.3 displays parameter estimates, standard errors, t statistics, denominator degrees of freedom, p-values, and hazard ratios. In this example data set, Treatment and the Treatment × DiabeticType interaction are significant with p-values 0.023 and 0.006, respectively. Because the model contains the Treatment × DiabeticType interaction, the exponential of the estimated regression coefficient is not the hazard ratio. The HAZARDRATIO statement requests the hazard ratios.

Output 119.1.3 Parameter Estimates

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
Treatment Laser	196	-0.424672	0.185438	-2.29	0.0231	0.654
DiabeticType Adult	196	0.340841	0.196076	1.74	0.0837	1.406
Treatment*DiabeticTy Laser Adult	196	-0.845665	0.304303	-2.78	0.0060	0.429

Output 119.1.4 displays the estimated hazard ratios and corresponding 95% Wald confidence intervals. For both types of diabetes, the 95% confidence interval for the hazard ratio lies below 1, indicating the effectiveness of the treatment for both types of diabetes.

Output 119.1.4 Hazard Ratios

Hazard Ratios for Treatment

Description	Point Estimate	95% Wald Confidence Limits	
Treatment Laser vs Others At DiabeticType=Adult	0.281	0.174	0.453
Treatment Laser vs Others At DiabeticType=Juvenile	0.654	0.454	0.943

Example 119.2: 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. For more information about the survey and the data sets, see the Center for Disease Control and Prevention's website (https://www.cdc.gov/).

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.

```
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  009  0.02754    3580   2  50  0  157
   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.

```
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 119.2.1. 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 19,085,105 ($= 146.8085 \times 130{,}000$) observation units in the population. Note that the estimated number of observation units in the population 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 119.2.1 Model Information, Data Summary, Design Summary, and Information about Variance Estimation

The SURVEYPHREG Procedure

Model Information	
Data Set	WORK.CANCER
Dependent Variable	Age
Censoring Variable	Cancer
Censoring Value(s)	0
Weight Variable	AnalysisWt
Stratum Variable	Strata
Cluster Variable	PSU
Ties Handling	EFRON

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

Design Summary	
Number of Strata	10
Number of Clusters	596

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
1017	119	898	88.30

Summary of the Weighted Number of Event and Censored Values

Total	Event	Censored	Percent Censored
146.8085	17.01185	129.7966	88.41

Variance Estimation	
Method	Taylor Series
Missing Values	Excluded

The "Iteration History" table in Output 119.2.2 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.

Output 119.2.2 Iteration History

Maximum Likelihood Iteration History

Iteration	Restarts	Function Calls	Active Constraints	Objective Function	Objective Function Change	Max Abs Gradient Element	Ridge	Ratio Between Actual and Predicted Change
1	0	4	0	-63.34004	1.6501	21.9620	0	0.916
2	0	6	0	-63.29819	0.0418	0.2005	0	1.052
3	0	8	0	-63.29776	0.000430	0.00293	0	1.012
4	0	10	0	-63.29776	1.528E-7	1.102E-6	0	1.000

Estimates for proportional hazards regression coefficients and their standard errors are shown in Output 119.2.3. 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 584 denominator degrees of freedom.

Output 119.2.3 Parameter Estimates

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
BodyWeight	586	0.011920	0.003150	3.78	0.0002	1.012
Smoke -1	586	-1.174048	0.738358	-1.59	0.1124	0.309
Smoke 1	586	-1.006515	0.577955	-1.74	0.0821	0.365
Smoke 2	586	-0.674183	0.557587	-1.21	0.2271	0.510
Smoke 3	586	0	.	.	.	1.000

Type III Tests of Model Effects

Effect	Num DF	Den DF	F Value	Pr > F
BodyWeight	1	586	14.32	0.0002
Smoke	3	584	1.49	0.2162

Estimate

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Exponentiated
Row 1	-0.7532	0.3864	586	-1.95	0.0518	0.4709

Example 119.3: Domain Analysis

This example uses a data set from the NHANES I Epidemiologic Followup Study (NHEFS); see Example 119.2 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)
- 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)

```
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):

```
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 119.3.1 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 119.3.1 Summary Statistics for the Entire Population

The SURVEYPHREG Procedure

Number of Observations Read	1891
Number of Observations Used	1668
Sum of Weights Read	1.0298E8
Sum of Weights Used	89439590

Design Summary	
Number of Strata	22
Number of Clusters	55

Output 119.3.1 *continued*

Summary of the Number of Event and Censored Values			
Total	Event	Censored	Percent Censored
1668	717	951	57.01

Summary of the Weighted Number of Event and Censored Values			
Total	Event	Censored	Percent Censored
89439590	27650348	61789242	69.08

Variance Estimation	
Method	Taylor Series
Missing Values	NOMCAR

Output 119.3.2 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 119.3.2 Inference for the Entire Population

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
POVARIND 1	33	0.384961	0.077586	4.96	<.0001	1.470
POVARIND 2	33	0	.	.	.	1.000

Output 119.3.3 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 119.3.3 Summary Statistics for the Male Subpopulation

The SURVEYPHREG Procedure

Domain Analysis for domain GENDER=1

Number of Observations Read	1891
Number of Observations Used	813
Sum of Weights Read	48887067
Sum of Weights Used	42629905

Output 119.3.3 *continued*

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
813	404	409	50.31

Summary of the Weighted Number of Event and Censored Values

Total	Event	Censored	Percent Censored
42629905	15126321	27503584	64.52

Output 119.3.4 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 119.3.4 Inference for the Male Subpopulation

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
POVARIND 1	33	0.424922	0.156583	2.71	0.0105	1.529
POVARIND 2	33	0	.	.	.	1.000

Output 119.3.5 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 119.3.5 Summary Statistics for the Female Subpopulation

The SURVEYPHREG Procedure

Domain Analysis for domain GENDER=2

Number of Observations Read	1891
Number of Observations Used	855
Sum of Weights Read	54091604
Sum of Weights Used	46809685

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
855	313	542	63.39

Summary of the Weighted Number of Event and Censored Values

Total	Event	Censored	Percent Censored
46809685	12524027	34285658	73.24

Output 119.3.6 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 119.3.6 Inference for the Female Subpopulation

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
POVARIND 1	33	0.434579	0.115766	3.75	0.0007	1.544
POVARIND 2	33	0	.	.	.	1.000

Example 119.4: Variance Estimation by Using Replicate Weights

Consider the data set LibrarySurvey from "Getting Started: SURVEYPHREG Procedure" on page 9887. 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.

```
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 2 4 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 119.4.1 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 9887. 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 119.4.1 Summary Statistics for Overall Analysis

The SURVEYPHREG Procedure

Number of Observations Read	100
Number of Observations Used	100
Sum of Weights Read	11616.79
Sum of Weights Used	11616.79

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
100	90	10	10.00

Summary of the Weighted Number of Event and Censored Values

Total	Event	Censored	Percent Censored
11616.79	10449.22	1167.57	10.05

Variance Estimation

Method	Jackknife
Replicate Weights	WORK.LIBRARYREPWEIGHTS
Number of Replicates	46

Output 119.4.2 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 9887, 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.

Output 119.4.2 Inferences Based on Survey Design for Overall Analysis

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
Age	46	0.061593	0.009159	6.73	<.0001	1.064

Example 119.5: 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 119.3. 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.

```
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 119.5.1 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 119.3: Domain Analysis" on page 9970.

Output 119.5.1 Data Summary, Censored Summary, and Information about Variance Estimation

The SURVEYPHREG Procedure

Number of Observations Read	1891
Number of Observations Used	1668
Sum of Weights Read	1.0298E8
Sum of Weights Used	89439590

Output 119.5.1 *continued*

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
1668	717	951	57.01

Summary of the Weighted Number of Event and Censored Values

Total	Event	Censored	Percent Censored
89439590	27650348	61789242	69.08

Variance Estimation	
Method	Taylor Series
Missing Values	NOMCAR

Output 119.5.2 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 119.5.2 Parameter Estimates

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
GENDER 1	33	1.605505	0.709081	2.26	0.0303	4.980
GENDER 2	33	0	.	.	.	1.000
x	33	-0.015648	0.010079	-1.55	0.1301	0.984

References

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

Beaumont, J.-F., and Patak, Z. (2012). "On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling." *International Statistical Review* 80:127–148.

Binder, D. A. (1983). "On the Variances of Asymptotically Normal Estimators from Complex Surveys." *International Statistical Review* 51:279–292.

Binder, D. A. (1990). "Fitting Cox's Proportional Hazards Models from Survey Data." In *Proceedings of the Survey Research Methods Section*, 342–347. Alexandria, VA: American Statistical Association.

Binder, D. A. (1992). "Fitting Cox's Proportional Hazards Models from Survey Data." *Biometrika* 79:139–147.

Boudreau, C., and Lawless, J. F. (2006). "Survival Analysis Based on the Proportional Hazards Model and Survey Data." *Canadian Journal of Statistics* 34:203–216.

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

Brick, J. M., and Kalton, G. (1996). "Handling Missing Data in Survey Research." *Statistical Methods in Medical Research* 5:215–238.

Chambers, R. L., and Skinner, C. J. (2003). *Analysis of Survey Data*. Chichester, UK: John Wiley & Sons.

Cochran, W. G. (1977). *Sampling Techniques*. 3rd ed. New York: John Wiley & Sons.

Cox, D. R. (1972). "Regression Models and Life-Tables." *Journal of the Royal Statistical Society, Series B* 34:187–220. With discussion.

Cox, D. R. (1975). "Partial Likelihood." *Biometrika* 62:269–276.

Dippo, C. S., Fay, R. E., and Morganstein, D. H. (1984). "Computing Variances from Complex Samples with Replicate Weights." In *Proceedings of the Survey Research Methods Section*, 489–494. Alexandria, VA: American Statistical Association.

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

Fay, R. E. (1984). "Some Properties of Estimates of Variance Based on Replication Methods." In *Proceedings of the Survey Research Methods Section*, 495–500. Alexandria, VA: American Statistical Association.

Fay, R. E. (1989). "Theory and Application of Replicate Weighting for Variance Calculations." In *Proceedings of the Survey Research Methods Section*, 212–217. Alexandria, VA: American Statistical Association.

Firth, D. (1993). "Bias Reduction of Maximum Likelihood Estimates." *Biometrika* 80:27–38.

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

Fuller, W. A. (1975). "Regression Analysis for Sample Survey." *Sankhyā, Series C* 37:117–132.

Fuller, W. A. (2009). *Sampling Statistics*. Hoboken, NJ: John Wiley & Sons.

Fuller, W. A., Kennedy, W. J., Schnell, D., Sullivan, G., and Park, H. J. (1989). *PC CARP*. Ames: Iowa State University Statistical Laboratory.

Godambe, V. P., and Thompson, M. E. (1986). "Parameters of Superpopulation and Survey Population: Their Relationships and Estimation." *International Statistical Review* 54:127–138.

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

Heinze, G. (1999). *The Application of Firth's Procedure to Cox and Logistic Regression*. Technical Report 10, updated January 2001, Department of Medical Computer Sciences, Section of Clinical Biometrics, University of Vienna.

Heinze, G., and Schemper, M. (2001). "A Solution to the Problem of Monotone Likelihood in Cox Regression." *Biometrics* 51:114–119.

Judkins, D. R. (1990). "Fay's Method for Variance Estimation." *Journal of Official Statistics* 6:223–239.

Kalbfleisch, J. D., and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd ed. Hoboken, NJ: John Wiley & Sons.

Kalton, G., and Kasprzyk, D. (1986). "The Treatment of Missing Survey Data." *Survey Methodology* 12:1–16.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

Kish, L., and Frankel, M. R. (1974). "Inference from Complex Samples." *Journal of the Royal Statistical Society, Series B* 36:1–37.

Korn, E. L., and Graubard, B. I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons.

Lawless, J. F. (2003). *Statistical Model and Methods for Lifetime Data*. 2nd ed. New York: John Wiley & Sons.

Lee, E. W., Wei, L. J., and Amato, D. A. (1992). "Cox-Type Regression Analysis for Large Numbers of Small Groups of Correlated Failure Time Observations." In *Survival Analysis: State of the Art*, edited by J. P. Klein and P. K. Goel, 237–247. Dordrecht, Netherlands: Kluwer Academic.

Lin, D. Y. (2000). "On Fitting Cox's Proportional Hazards Models to Survey Data." *Biometrika* 87:37–47.

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

Lohr, S. L. (2010). *Sampling: Design and Analysis*. 2nd ed. Boston: Brooks/Cole.

Lumley, T., and Scott, A. (2013). "Partial Likelihood Ratio Tests for the Cox Model under Complex Sampling." *Statistics in Medicine* 32:110–123.

Mashreghi, Z., Haziza, D., and Léger, C. (2016). "A Survey of Bootstrap Methods in Finite Population Sampling." *Statistics Surveys* 10:1–52.

McCarthy, P. J., and Snowden, C. B. (1985). *The Bootstrap and Finite Population Sampling*. Vital and Health Statistics, Series 2, no. 95. DHHS Publication Number 85–1369. Washington, DC: US Government Printing Office. https://www.cdc.gov/nchs/data/series/sr_02/sr02_095.pdf.

Pfeffermann, D. (1993). "The Role of Sampling Weights When Modeling Survey Data." *International Statistical Review* 61:317–337.

Rao, J. N. K., Scott, A. J., and Skinner, C. J. (1998). "Quasi-score Tests with Survey Data." *Statistica Sinica* 8:1059–1070.

Rao, J. N. K., and Shao, J. (1996). "On Balanced Half-Sample Variance Estimation in Stratified Random Sampling." *Journal of the American Statistical Association* 91:343–348.

Rao, J. N. K., and Shao, J. (1999). "Modified Balanced Repeated Replication for Complex Survey Data." *Biometrika* 86:403–415.

Rao, J. N. K., and Wu, C. F. J. (1988). "Resampling Inference with Complex Survey Data." *Journal of the American Statistical Association* 83:231–241.

Rao, J. N. K., Wu, C. F. J., and Yue, K. (1992). "Some Recent Work on Resampling Methods for Complex Surveys." *Survey Methodology* 18:209–217.

Rust, K. (1985). "Variance Estimation for Complex Estimators in Sample Surveys." *Journal of Official Statistics* 1:381–397.

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Chichester, UK: John Wiley & Sons.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

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

Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.

Sitter, R. R. (1992a). "Comparing Three Bootstrap Methods for Survey Data." *Canadian Journal of Statistics* 20:135–154.

Sitter, R. R. (1992b). "A Resampling Procedure for Complex Survey Data." *Journal of the American Statistical Association* 87:755–765.

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

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

Wolter, K. M. (2007). *Introduction to Variance Estimation*. 2nd ed. New York: Springer.

Subject Index

Syntax Index