SAS/STAT® 15.1 User's Guide
The RMSTREG 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 103
The RMSTREG Procedure

Contents

Overview: RMSTREG Procedure

The RMSTREG procedure analyzes time-to-event data by using regression with respect to the restricted mean survival time (RMST), using specialized methods such as those developed by Andersen, Hansen, and Klein (2004) and Tian, Zhao, and Wei (2014). The RMST is defined as the expected value of the time-to-event variable up to a prespecified time point τ.

For a prospective study in which the time to an event (or survival) is of interest, it is common practice to summarize the outcome in terms of the mean or median survival rate. However, if the last observation is censored, then the mean cannot be reliably estimated; and when not enough events occur, the median can be inestimable. The survival rate at a specified time is another common summary statistic, and although it can be less problematic than the mean or median, it does not provide an overall summary of the time-to-event outcome. The restricted mean survival time, sometimes called the restricted mean event time, is an alternative measure that is more often reliably estimable than the mean and median survival time in certain situations. Also, it provides a summary of the whole survival curve up to a time horizon, in contrast to the survival rate at a specified time (Royston and Parmar 2013; Uno et al. 2014; Trinquart et al. 2016).

Although it is straightforward to estimate the RMST in the absence of covariates by using the Kaplan-Meier curve up to time τ, estimating RMST when covariates are present is more complicated. It is possible to estimate RMST from a classical survival model, such as the Cox proportional hazards model (Zucker 1998), but the process is complex and difficult to extend. You can avoid this difficulty by using generalized linear modeling techniques to directly model the RMST. This has the double advantage of making inferences on the results straightforward and providing all the machinery of generalized linear model comparisons for studying RMST effects. This is the technique that PROC RMSTREG employs.

The RMSTREG procedure supports classification and interaction effects, along with certain generalized linear models links (Nelder and Wedderburn 1972), such as linear and log-linear functions. Models are fitted using methods based on estimating equations, such as pseudovalue regression (Andersen, Hansen, and Klein 2004) or inverse probability censoring weighting (IPCW) (Tian, Zhao, and Wei 2014). Compared to Cox regression and other classical methods, these methods enable you to model the RMST directly, facilitating model-based inference and prediction.

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

Comparison with Other Procedures

The RMSTREG procedure compares most closely to the PHREG and LIFEREG procedures in SAS/STAT software. The three procedures fit different models to time-to-event data. The PHREG procedure fits the Cox proportional hazards model and various extensions. The LIFEREG procedure fits accelerated failure time models. The RMSTREG procedure fits generalized linear models to the RMST. Table 103.1 summarizes the key features of these procedures.

Table 103.1 Survival Modeling Procedures

Procedure	Focus	Model Type	Estimation Method
PROC LIFEREG	Time-to-event	Accelerated failure time models	Likelihood
PROC PHREG	Hazard function	Proportional hazards models	Partial likelihood
PROC RMSTREG	Restricted mean survival time	Generalized linear models	Estimating equations

Getting Started: RMSTREG Procedure

This section presents an example that illustrates some of the basic features of the RMSTREG procedure by analyzing liver disease data from Lin, Wei, and Ying (1993).

The data represent 418 patients who have primary biliary cirrhosis (PBC), among whom 161 had died as of the date of data listing. A subset of the variables is saved in the SAS data set Liver. The data set contains the following variables:

- Time, follow-up time, in years
- Status, event indicator with the value 1 for death time and 0 for censored time
- Age, age in years, from birth to study registration
- Albumin, serum albumin level, in g/dl
- Bilirubin, serum bilirubin level, in mg/dl
- Edema, edema presence
- Protime, prothrombin time, in seconds

The following statements create the data set Liver:

```
data Liver;
   input Time Status Age Albumin Bilirubin Edema Protime @@;
   label Time="Follow-Up Time in Years";
   Time= Time / 365.25;
   datalines;
  400 1 58.7652 2.60 14.5 1.0 12.2 4500 0 56.4463 4.14  1.1 0.0 10.6
 1012 1 70.0726 3.48  1.4 0.5 12.0 1925 1 54.7406 2.54  1.8 0.5 10.3
 1504 0 38.1054 3.53  3.4 0.0 10.9 2503 1 66.2587 3.98  0.8 0.0 11.0
 1832 0 55.5346 4.09  1.0 0.0  9.7 2466 1 53.0568 4.00  0.3 0.0 11.0
 2400 1 42.5079 3.08  3.2 0.0 11.0   51 1 70.5599 2.74 12.6 1.0 11.5
 3762 1 53.7139 4.16  1.4 0.0 12.0  304 1 59.1376 3.52  3.6 0.0 13.6
 3577 0 45.6893 3.85  0.7 0.0 10.6 1217 1 56.2218 2.27  0.8 1.0 11.0
 3584 1 64.6461 3.87  0.8 0.0 11.0 3672 0 40.4435 3.66  0.7 0.0 10.8

   ... more lines ...

  989 0 35.0000 3.23  0.7 0.0 10.8  681 1 67.0000 2.96  1.2 0.0 10.9
```

```
 1103 0 39.0000 3.83  0.9 0.0 11.2 1055 0 57.0000 3.42  1.6 0.0  9.9
  691 0 58.0000 3.75  0.8 0.0 10.4  976 0 53.0000 3.29  0.7 0.0 10.6
;
```

The following statements fit a linear model for the RMST with the covariates Bilirubin, Age, and Edema:

```
proc rmstreg data=liver tau=10;
   class Edema;
   model Time*Status(0) = Age Bilirubin Edema / link=linear method=pv;
run;
```

The TAU= option in the PROC RMSTREG statement specifies the time limit that defines the RMST for this analysis. If you omit this option, the largest event time from the input data is used. Because the variable Edema is specified as a CLASS variable, it contributes one dummy variable to the regression model for each of its values.

In the MODEL statement, the response consists of the observed variable Time and an indicator variable Status, which specifies whether or not the Time value is censored. The values of Time are considered to be censored if the value of Status is 0; otherwise, they are considered to be event times.

An intercept term is included by default. Thus, the model matrix $\mathbf{X}$ consists of a column of 1s that represent the intercept term, three columns of 0s and 1s that correspond to the levels of the Edema variable, and two additional columns for the values of the variables Age and Bilirubin.

That is, the model matrix is

$$\mathbf{X} = \left[\begin{array}{c|ccc|cc} 1 & 1 & 0 & 0 & \text{Age} & \text{Bilirubin} \\ 1 & 0 & 1 & 0 & \text{Age} & \text{Bilirubin} \\ 1 & 0 & 0 & 1 & \text{Age} & \text{Bilirubin} \end{array}\right]$$

The LINK=LINEAR option fits a linear model. That is, the RMST at $\tau = 10$ for a specific subject (denoted by μ_i) is related to the linear predictor by

$$\log(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$$

The METHOD=PV option specifies pseudovalue regression to fit the model.

Figure 103.1 shows the "Model Information" table, which provides information about the specified linear model of the RMST and the input data set.

Figure 103.1 Model Information

The RMSTREG Procedure

Model Information	
Data Set	WORK.LIVER
Time Variable	Time
Censoring Variable	Status
Censoring Value(s)	0
Link Function	Linear
Estimation Method	Pseudo Value
Tau Value	10

Figure 103.2 displays a summary of the number of event and censored observations in the data set.

Figure 103.2 Event and Censoring Summary

Summary of the Number of Event and Censored Values

Total	Event	Censored
418	161	257

Figure 103.3 shows how the Edema variable is coded in the model matrix.

Figure 103.3 CLASS Variable Level Information

Class Level Information

Class	Value	Design Variables		
Edema	**0**	1	0	0
	0.5	0	1	0
	1	0	0	1

For each parameter in the model, PROC RMSTREG displays a table (Figure 103.4) that contains columns of the parameter name, the degrees of freedom associated with the parameter, the estimated parameter value, the standard error of the parameter estimate, the confidence intervals, and the Wald chi-square statistic and associated *p*-value for testing the significance of the parameter to the model. If a column of the model matrix that corresponds to a parameter is found to be linearly dependent, or *aliased*, with columns that correspond to parameters preceding it in the model, PROC RMSTREG assigns it zero degrees of freedom and displays a value of 0 for both the parameter estimate and its standard error.

Figure 103.4 Analysis of Parameter Estimates

Analysis of Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	9.0588	1.1201	6.8635	11.2540	65.41	<.0001
Age		1	-0.0686	0.0142	-0.0964	-0.0408	23.41	<.0001
Bilirubin		1	-0.3614	0.0383	-0.4366	-0.2862	88.82	<.0001
Edema	0	1	3.0259	0.7353	1.5849	4.4670	16.94	<.0001
Edema	0.5	1	1.7778	0.8658	0.0807	3.4748	4.22	0.0401
Edema	1	0	0.0000					

You generally assess the importance of the main effects in the model. By default, PROC RMSTREG performs a Type 3 analysis, the results of which are shown in Figure 103.5.

Figure 103.5 Type 3 Analysis of Effects

Type 3 Analysis of Effects			
Effect	**DF**	**Chi-Square**	**Pr > ChiSq**
Age	1	23.4052	<.0001
Bilirubin	1	88.8223	<.0001
Edema	2	19.4651	<.0001

The results indicate that all three variables are strong predictors of the RMST at $\tau = 10$. The Type 3 chi-square value for the Age variable, for example, is computed by comparing the model with Intercept, Age, Bilirubin, and Edema included and the model with Age excluded. The hypothesis that is tested in this case is that Age adds no predictability to the model, over and above Edema and Bilirubin. This hypothesis is strongly rejected.

If the RMST at $\tau = 5$ is of interest, then the following statements fit this model by using pseudovalue regression:

```
proc rmstreg data=liver tau=5;
   class Edema;
   model Time*Status(0) = Age Bilirubin Edema / link=linear method=pv;
run;
```

Figure 103.6 displays the parameter estimates and the Type 3 tests from the fitted model. The new results appear to be similar to the previous ones. Using values 5 and 10 for τ, respectively, makes little difference.

Figure 103.6 Parameter Estimates and Type 3 Tests for RMST at $\tau = 5$

The RMSTREG Procedure

Type 3 Analysis of Effects			
Effect	**DF**	**Chi-Square**	**Pr > ChiSq**
Age	1	18.9013	<.0001
Bilirubin	1	62.3617	<.0001
Edema	2	30.8671	<.0001

Analysis of Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**95% Confidence Limits**		**Chi-Square**	**Pr > ChiSq**
Intercept		1	4.0394	0.4848	3.0892	4.9896	69.42	<.0001
Age		1	-0.0220	0.0051	-0.0319	-0.0121	18.90	<.0001
Bilirubin		1	-0.1378	0.0174	-0.1720	-0.1036	62.36	<.0001
Edema	0	1	1.8901	0.3621	1.1805	2.5997	27.25	<.0001
Edema	0.5	1	1.2573	0.4205	0.4331	2.0815	8.94	0.0028
Edema	1	0	0.0000					

Syntax: RMSTREG Procedure

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

PROC RMSTREG < *options* > **;**
BY *variables* **;**
CLASS *variable* < **(***options***)** > . . . < *variable* < **(***options***)** > > < / *options* > **;**
ESTIMATE < '*label*' > *estimate-specification* < / *options* > **;**
FREQ | **FREQUENCY** *variable* **;**
LSMEANS < *model-effects* > < / *options* > **;**
LSMESTIMATE *model-effect* < '*label*' > *values* < *divisor*=*n* > < **,** . . . < '*label*' > *values* < *divisor*=*n* > >
< / *options* > **;**
MODEL *response* < ∗ *censor***(***list***)** > **=** < *effects* > < / *options* > **;**
SLICE *model-effect* < / *options* > **;**
STORE < **OUT=** > *item-store-name* < / **LABEL=**'*label*' > **;**
TEST < *model-effects* > < / *options* > **;**

The PROC RMSTREG and MODEL statements are required. The CLASS statement, if present, must precede the MODEL statement. The following sections describe the PROC RMSTREG statement and then describe the other statements in alphabetical order.

PROC RMSTREG Statement

PROC RMSTREG < *options* > **;**

The PROC RMSTREG statement invokes the RMSTREG procedure. Table 103.2 summarizes the *options* available in the PROC RMSTREG statement.

Table 103.2 PROC RMSTREG Statement Options

Option	Description
ALPHA=	Specifies the alpha level
DATA=	Specifies the input data set
ITHISTORY	Prints the iteration history of the model parameters
NAMELEN=	Specifies the length of effect names
NOPRINT	Suppresses all displayed output
OUTPV=	Outputs the pseudovalues used in pseudovalue regression to a SAS data set
SINGULAR=	Specifies the tolerance for testing the singularity
TAU=	Specifies the upper time limit that defines the RMST

You can specify the following *options*.

ALPHA=*number*
specifies the α level for $100(1-\alpha)\%$ confidence limits. The *number* must be between 0 and 1; the default value is 0.05, which results in 95% intervals. This value is used as the default level for

confidence limits that are computed by the MODEL statement. You can override this default by specifying the ALPHA= option in the MODEL statement.

DATA=*SAS-data-set*
specifies the SAS data set that contains the data to be analyzed. If you omit this option, PROC RMSTREG uses the most recently created SAS data set.

ITHISTORY
displays the iteration history of computing estimates on the basis of the estimating equations and the final evaluation of the gradient.

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

NOPRINT
suppresses all displayed output. This option temporarily disables the Output Delivery System (ODS). For more information, see Chapter 20, "Using the Output Delivery System."

OUTPV=*SAS-data-set*
names the output data set that contains the pseudovalues that are used in the pseudovalue regression analysis. To use this option, you must also specify METHOD=PV in the MODEL statement.

SINGULAR=*number*

EPSILON=*number*
specifies the tolerance for testing the singularity of the $\mathbf{X}'\mathbf{X}$ matrix that is formed from the design matrix $\mathbf{X}$ and for testing the singularity of the Hessian matrix upon convergence of the optimization algorithm. Appropriately, the test requires that a pivot be at least *number* times the original diagonal value. By default, *number* is 10^7 times the machine epsilon. On most machines, the default *number* is approximately 10^{-9}.

TAU=*value*
specifies the upper time limit of the RMST. The default *value* is the largest event time, where *value* must be positive.

BY Statement

BY *variables* **;**

You can specify a BY statement in PROC RMSTREG 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 RMSTREG procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables by using the DATASETS procedure (in Base SAS software).

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

CLASS Statement

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

The CLASS statement names the classification variables to be used as explanatory variables in the analysis. Response variables do not need to be specified in the CLASS statement.

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 them after a slash (/). *Global-options* are applied to all the variables that are specified in the CLASS statement. If you specify more than one CLASS statement, the *global-options* that are specified in any one CLASS statement apply to all CLASS statements. However, individual CLASS variable *options* override the *global-options*. You can specify the following values for either an *option* or a *global-option*:

CPREFIX=*n*

specifies that, at most, the first *n* characters of a CLASS variable name be used in creating names for the corresponding design variables. The default is $32 - \min(32, \max(2, f))$, where f is the formatted length of the CLASS variable.

DESCENDING

DESC

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

LPREFIX=*n*

specifies that, at most, the first *n* characters of a CLASS variable label be used in creating labels for the corresponding design variables. The default is $256 - \min(256, \max(2, f))$, where f is the formatted length of the CLASS variable.

MISSING

treats missing values (., ._, .A, . . . , .Z for numeric variables and blanks for character variables) as valid values of 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 this 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 RMSTREG 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. You can specify any of the *keywords* shown in the following table. The default is PARAM=GLM. 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.

Class Variable Default Parameterization

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

Class Variable Naming Convention

Parameter names for a CLASS predictor variable are constructed by concatenating the CLASS variable name with the CLASS levels. However, for the POLYNOMIAL and orthogonal parameterizations, parameter names are formed by concatenating the CLASS variable name and keywords that reflect the parameterization. For examples and more information, see the section "Other Parameterizations" on page 397 in Chapter 19, "Shared Concepts and Topics."

Class Variable Parameterization with Unbalanced Designs

PROC RMSTREG initially parameterizes the CLASS variables by looking at the levels of the variables across the complete data set. If you have an *unbalanced* replication of levels across variables or BY groups, then the design matrix and the parameter interpretation might be different from what you expect. For example, suppose you have a model that has one CLASS variable A with three levels (1, 2, and 3) and another CLASS

variable B with two levels (1 and 2). If the third level of A occurs only with the first level of B, if you use the EFFECT parameterization, and if your model contains the effect A(B) and an intercept, then the design for A within the second level of B is not a differential effect. In particular, the design looks like the following:

		Design Matrix			
		A(B=1)		A(B=2)	
B	A	A1	A2	A1	A2
1	1	1	0	0	0
1	2	0	1	0	0
1	3	–1	–1	0	0
2	1	0	0	1	0
2	2	0	0	0	1

PROC RMSTREG detects linear dependency among the last two design variables and sets the parameter for A2(B=2) to zero, resulting in an interpretation of these parameters as if they were reference- or dummy-coded. The REFERENCE or GLM parameterization might be more appropriate for such problems.

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 103.3 summarizes the *options* available in the ESTIMATE statement.

Table 103.3 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

Table 103.3 *continued*

Option	Description
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 **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
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* **;**

FREQUENCY *variable* **;**

The *variable* in the FREQ statement identifies a variable in the input data set that contains the frequency of occurrence of each observation. PROC RMSTREG treats each observation as if it appeared *n* times, where *n* is the value of the FREQ variable for the observation. If the frequency value is not an integer, it is truncated to an integer. If it is less than 1 or missing, the observation is not used. The frequencies must be the same for all observations within each subject.

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

Table 103.4 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
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 103.5 summarizes the *options* available in the LSMESTIMATE statement.

Table 103.5 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

Table 103.5 *continued*

Option	Description
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* > **;**

The MODEL statement specifies the response variables and the effects (explanatory variables). If you omit the explanatory variables, PROC RMSTREG fits an intercept-only model. An intercept term is included in the model by default. You can omit the intercept by specifying the NOINT option.

The name of the failure time variable precedes the equal sign. This name 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 one of these values, the corresponding failure time is considered to be censored. Following the equal sign are the explanatory effects (sometimes called independent variables or covariates) for the model.

The censoring variable must be numeric, and the failure time variables 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. Failure time variables in a SAS date format are not recommended, because the dates might be translated into negative numbers and consequently the corresponding observation would be discarded.

The *effects* in the MODEL statement consist of an explanatory variable or combination of variables. Explanatory variables can be continuous or classification variables. Classification variables can be character or numeric. Explanatory variables that represent nominal (classification) data must be declared in a CLASS statement. Interactions between variables can also be included as effects. Columns of the design matrix are automatically generated for classification variables and interactions. The syntax for specifying effects is the same as in the GLM procedure. For more information, see the section "Specification of Effects" on page 4020 in Chapter 50, "The GLM Procedure."

Table 103.6 summarizes the *options* available in the MODEL statement.

Table 103.6 MODEL Statement Options

Option	Description
ALPHA=	Sets the confidence coefficient
CONVERGE=	Sets the convergence criterion
CONVH=	Sets the relative Hessian convergence criterion
CORRB	Displays the parameter estimate correlation matrix
COVB	Displays the parameter estimate covariance matrix

Table 103.6 *continued*

Option	Description
LINK=	Specifies the link function
MAXITER=	Sets the maximum allowable number of iterations for all iterative computation processes
METHOD=	Specifies the estimation method for model fitting
NOFIT	Suppresses model fitting
NOINT	Excludes the intercept term from the model

You can specify the following *options* after a slash (/).

ALPHA=*number*
sets the confidence coefficient for parameter confidence intervals to 1–*number*. The value of *number* must be between 0 and 1. The default value of *number* is 0.05.

CONVERGE=*number*
sets the convergence criterion for parameter estimation. If the maximum absolute difference between regression parameter estimates is less than the value of *number* in two successive iterations, convergence is declared. If the absolute value of a regression parameter estimate is greater than 0.08, then the absolute difference normalized by the regression parameter value is used instead of the absolute difference. The default value of *number* is 0.0001.

CONVH=*number*
sets the relative Hessian convergence criterion. The value of *number* must be between 0 and 1. After convergence is determined by using the change in parameter criterion that is specified by the CONVERGE= option, the quantity $tc = \mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$ is computed and compared to *number*, where $\mathbf{g}$ is the gradient vector and $\mathbf{H}$ is the Hessian matrix for the model parameters. If tc is greater than *number*, a warning is issued that the relative Hessian convergence criterion has been exceeded. This criterion detects the occasional case where the change in parameter convergence criterion is satisfied but a global solution has not been attained. By default, CONVH=1E–4.

CORRB
displays the parameter estimate correlation matrix.

COVB
displays the parameter estimate covariance matrix.

LINK=*keyword*
specifies the link function in the model. You can specify the *keywords* shown in Table 103.7.

Table 103.7 Built-In Link Functions of the RMSTREG Procedure

LINK=	Link Function	$g(\mu) =$
IDENTITY | ID | LINEAR	Identity	μ
LOG	Log	$\log(\mu)$

By default, LINK=LOG.

MAXITER=*number*

MAXIT=*number*

sets the maximum allowable number of iterations for all iterative computation processes in PROC RMSTREG. By default, MAXITER=50.

METHOD=*method* **<(STRATA=***variable***)>**

specifies the estimation method to fit the specified model. You can specify the *methods* shown in Table 103.8.

Table 103.8 Estimation Methods of the RMSTREG Procedure

METHOD=	Estimation Method
PV	Pseudovalue regression (Andersen, Hansen, and Klein 2004)
IPCW	Inverse probability censoring weighting estimation (Tian, Zhao, and Wei 2014)

By default, METHOD=PV. For METHOD=IPCW, you can compute stratified weights and use them in the iterative estimation process by specifying the STRATA=*variable* option. For more information about pseudovalue regression, see the section "Pseudovalue Regression" on page 8636. For more information about the inverse probability censoring weighting techniques, see the section "Inverse Probability Censoring Weighting Estimation" on page 8637.

NOFIT

suppresses model fitting. If this option is specified along with other MODEL statement options, NOFIT takes precedence, and all other options are ignored.

NOINT

excludes the intercept term from the model. An intercept is included unless you specify this option.

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."

TEST Statement

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

The TEST statement enables you to perform Wald 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 103.9 summarizes the *options* that you can specify in the TEST statement.

Table 103.9 TEST Statement Options

Option	Description
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."

Details: RMSTREG Procedure

Concepts and Basic Estimators

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

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

Assume that τ is a prespecified time point of interest. Let R be the minimal of T and τ,

$$R = T \wedge \tau = \min(T, \tau)$$

The restricted mean survival time (RMST) is defined as the expected value of R:

$$\mathrm{RMST}(\tau) = E(R) = E[\min(T, \tau)]$$

It can be evaluated by the area under the survival function over $[0, \tau]$ as

$$\mathrm{RMST}(\tau) = \int_0^\tau S(u)du$$

Let $t_1 < t_2 < \cdots < t_D$ represent the distinct event times. For each $i = 1, \ldots, D$, let Y_i be the number of surviving units (the size of the risk set) just prior to t_i, and let d_i be the number of units that fail at t_i.

The Kaplan-Meier (product-limit) estimate of the survival function at t_i is the cumulative product

$$\hat{S}(t_i) = \prod_{j=1}^{i} \left(1 - \frac{d_j}{Y_j}\right)$$

The estimated mean survival time is

$$\hat{\mu} = \sum_{i=1}^{D} \hat{S}(t_{i-1})(t_i - t_{i-1})$$

where t_0 is defined to be zero.

The $\mathrm{RMST}(\tau)$ is estimated by

$$\widehat{\mathrm{RMST}}(\tau) = \sum_{i=1}^{N^*} \hat{S}(t_{i-1})(t_i - t_{i-1}) + \hat{S}(t_{N^*})(\tau - t_{N^*})$$

where N^* is the number of t_i values that are less than τ.

When the largest observed time is censored, the mean survival time is underestimated because the Kaplan-Meier estimate does not drop to zero (Klein and Moeschberger 2003). The RMST, on the other hand, can still be estimated unbiasedly as long as $\tau \leq t_D$.

Supported Models

Let D_i be the response variable for the ith observation. The quantity $\mathbf{x}_i$ is a column vector of covariates, or explanatory variables, for observation i that is known from the experimental setting and is considered to be fixed, or nonrandom.

The expected value of D_i, denoted by μ_i, is

$$\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ is an unknown parameter vector and another column can be added to $\mathbf{x}_i$ for an intercept effect.

Under the specification of generalized linear models (Nelder and Wedderburn 1972), μ_i is related to a linear predictor through a monotone and differentiable link function g:

$$g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$$

Assume that τ is a prespecified time point of interest. Let T_i be the time-to-event variable for the ith subject. The subject-specific RMST at τ is defined by $\mathrm{RMST}_i(\tau) = E[\min(T_i, \tau)]$ and can be conveniently modeled via a generalized linear model as

$$g[\mathrm{RMST}_i(\tau)] = \mathbf{x}_i'\boldsymbol{\beta}$$

Under the natural logarithm link $g(\cdot) = \log(\cdot)$, the model is

$$\log[\mathrm{RMST}_i(\tau)] = \mathbf{x}_i'\boldsymbol{\beta}$$

Under the identity or linear link, the model is

$$\mathrm{RMST}_i(\tau) = \mathbf{x}_i'\boldsymbol{\beta}$$

Specification of Effects

Each term in a model is called an effect. You specify effects in the MODEL statement by using a special notation that uses variable names and operators. There are two types of variables: *classification* (CLASS) variables and *continuous* variables. There are two primary types of operators: *crossing* and *nesting*. A third type, the *bar* operator, is used to simplify effect specification.

Variables that identify classification levels are called CLASS variables in SAS and are specified in a CLASS statement. These might also be called *categorical, qualitative, discrete,* or *nominal* variables. CLASS variables can be either character or numeric. The values of CLASS variables are called *levels*. For example, the CLASS variable Sex could have the levels "male" and "female."

In a model, an explanatory variable that is not declared in a CLASS statement is assumed to be continuous. Continuous variables must be numeric. For example, the heights and weights of subjects in an experiment are continuous variables.

The following list shows types of effects that are often useful in practice, where A, B, and C are classification variables and X1 and X2 are continuous variables:

- Regressor effects are specified by writing continuous variables by themselves: X1, X2.
- Polynomial effects are specified by using asterisks to join two or more continuous variables: X1*X2.
- Main effects are specified by writing classification variables by themselves: A, B, C.
- Crossed effects (interactions) are specified by using asterisks to join two or more classification variables: A*B, B*C, A*B*C.
- Nested effects are specified by following a main effect or crossed effect with a classification variable or list of classification variables that are enclosed in parentheses: B(A), C(B A), A*B(C). In this example, B(A) is "B nested within A."
- Combinations of continuous and classification variables can be specified in the same way by using the crossing and nesting operators.

The bar operator uses a vertical bar (|) to join two effects. The bar operator is shorthand notation for including the left-hand side, the right-hand side, and the cross between them as effects in the model. For example, the expression A | B is equivalent to A B A*B. The effects that are joined by the bar operator can be classification variables, continuous variables, or combinations of effects that are defined by using operators. Multiple bars are permitted. For example, A | B | C means A B C A*B A*C B*C A*B*C.

You can specify the maximum number of variables in any effect that results from bar evaluation by specifying the maximum number, preceded by an @ sign. For example, A | B | C@2 results in effects that involve no more than two variables: A B C A*B A*C B*C.

Parameterization Used in PROC RMSTREG

Design Matrix

The linear predictor part of a generalized linear model is

$$\eta = \mathbf{X}'\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ is an unknown parameter vector and $\mathbf{X}$ is a known design matrix. By default, all models automatically contain an intercept term; that is, the first column of $\mathbf{X}$ contains all 1s. Additional columns of $\mathbf{X}$ are generated for classification variables, regression variables, and any interaction terms included in the model. It is important to understand the ordering of classification variable parameters when you use the ESTIMATE statement. The ordering of these parameters is displayed in the "CLASS Level Information" table and in tables that display the parameter estimates of the fitted model.

When you specify an overparameterized model by using the PARAM=GLM option in the CLASS statement, some columns of $\mathbf{X}$ can be linearly dependent on other columns. For example, when you specify a model that consists of an intercept term and a classification variable, the column that corresponds to any one of the levels of the classification variable is linearly dependent on the other columns of $\mathbf{X}$. The columns of $\mathbf{X}'\mathbf{X}$ are checked in the order in which the model is specified for dependence on preceding columns. If a dependency is found, the parameter that corresponds to the dependent column is set to 0 along with its standard error to indicate that it is not estimated. The order in which the levels of a classification variable are checked for dependencies can be set by the ORDER= option in the CLASS statement. For full-rank parameterizations, the columns of the $\mathbf{X}$ matrix are designed to be linearly independent.

You can exclude the intercept term from the model by specifying the NOINT option in the MODEL statement.

Missing Level Combinations

All levels of interaction terms that involves classification variables might not be represented in the data. In that case, PROC RMSTREG does not include parameters in the model for the missing levels.

Type 3 Tests and Joint Tests

For models that use less-than-full-rank parameterization (as specified by the PARAM=GLM option in the CLASS statement), a Type 3 test of an effect of interest (main effect or interaction) is a test of the Type III estimable functions that are defined for that effect. When the model contains no missing cells, performing the Type 3 test of a main effect corresponds to testing the hypothesis of equal marginal means. For more

information about Type III estimable functions, see Chapter 50, "The GLM Procedure," and Chapter 15, "The Four Types of Estimable Functions." Also see Littell, Freund, and Spector (1991).

For models that use full-rank parameterization, all parameters are estimable when there are no missing cells, so it is unnecessary to define estimable functions. The standard test of an effect of interest in this case is the joint test that the values of the parameters associated with that effect are zero. For a model that uses effects parameterization (as specified by the PARAM=EFFECT option in the CLASS statement), performing the joint test for a main effect is equivalent to testing the equality of marginal means. For a model that uses reference parameterization (as specified by the PARAM=REF option in the CLASS statement), performing the joint test is equivalent to testing the equality of cell means at the reference level of the other model effects. For more information about the coding scheme and the associated interpretation of results, see Muller and Fetterman (2002, Chapter 14).

If there is no interaction term, the Type 3 test of an effect for a model that uses GLM parameterization is the same as the joint test of the effect for the model that uses full-rank parameterization. In this situation, the joint test is also called the Type 3 test. For a model that contains an interaction term and no missing cells, the Type 3 test of a component main effect under GLM parameterization is the same as the joint test of the component main effect under effect parameterization. Both test the equality of cell means. But this Type 3 test differs from the joint test under reference parameterization, which tests the equality of cell means at the reference level of the other component main effect. If some cells are missing, you can obtain meaningful tests only by testing a Type III estimable function, so in this case you should use GLM parameterization.

The results of a Type 3 test or a joint test do not depend on the order in which you specify the terms in the MODEL statement.

Confidence Intervals and Predicted Values

Wald Confidence Intervals

You can use PROC RMSTREG to produce Wald confidence intervals for the parameters. The $(1-\alpha)100\%$ Wald confidence interval for a parameter β is defined as

$$\hat{\beta} \pm z_{1-\alpha/2}\hat{\sigma}$$

where z_p is the $100p$th percentile of the standard normal distribution, $\hat{\beta}$ is the parameter estimate, and $\hat{\sigma}$ is the estimate of its standard error.

Predicted Values

A predicted value, or fitted value, of the mean μ_i that corresponds to the vector of covariates $\mathbf{x}_i$ is given by

$$\hat{\mu_i} = g^{-1}(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$$

where g is the link function, regardless of whether $\mathbf{x}_i$ corresponds to an observation. That is, even if the response variable is missing, the predicted value is still computed for valid $\mathbf{x}_i$. In the case where $\mathbf{x}_i$ does not correspond to a valid observation, $\mathbf{x}_i$ is not checked for estimability. You should check the estimability of $\mathbf{x}_i$ in this case in order to ensure the uniqueness of the predicted value of the mean.

Confidence Intervals on Predicted Values

Approximate confidence intervals for predicted values of the mean can be computed as follows. The variance of the linear predictor $\eta_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ is estimated by

$$\sigma_x^2 = \mathbf{x}_i' \boldsymbol{\Sigma} \mathbf{x}_i$$

where $\boldsymbol{\Sigma}$ is the estimated covariance of $\hat{\boldsymbol{\beta}}$.

Approximate $100(1-\alpha)\%$ confidence intervals are computed as

$$g^{-1}\left(\mathbf{x}_i'\hat{\beta} \pm z_{1-\alpha/2}\sigma_x\right)$$

where z_p is the $100p$th percentile of the standard normal distribution and g is the link function. If either endpoint in the argument is outside the valid range of arguments for the inverse link function, the corresponding confidence interval endpoint is set to missing.

Pseudovalue Regression

Pseudovalue regression is a generic method of fitting generalized linear models to time-to-event data (Andersen, Klein, and Rosthøj 2003). This section describes how the method works and how you can apply it to analyze models of the RMST.

Let $\mathbf{D}_1, \ldots, \mathbf{D}_n$ be independent and identically distributed quantities that might be random variables or vectors of variables. Let $\boldsymbol{\theta} = E[f(\mathbf{D}_i)]$ for some function $f(\cdot)$. Suppose $\hat{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\theta}$.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be independent and identically distributed samples of covariates, and define the conditional expectation of $f(\mathbf{D}_i)$ given by $\mathbf{x}_i$ as

$$\boldsymbol{\theta}_i = E[f(\mathbf{D}_i)|\mathbf{x}_i]$$

The ith pseudo-observation of $\boldsymbol{\theta}$ is computed as

$$\hat{\boldsymbol{\theta}}_i = n\hat{\boldsymbol{\theta}} - (n-1)\hat{\boldsymbol{\theta}}^{-i}$$

where $\hat{\boldsymbol{\theta}}^{-i}$ is the jackknife leave-one-out estimator for $\boldsymbol{\theta}$ based on $\{\mathbf{D}_j : j \neq i\}$.

The generalized linear model (Nelder and Wedderburn 1972) for $\boldsymbol{\theta}$ assumes

$$g(\boldsymbol{\theta}_i) = \mathbf{x}_i'\boldsymbol{\beta}$$

where $g(\cdot)$ is a suitable link function. Note that an additional column can be added to X_i for an intercept effect.

Using pseudo-observations, you can estimate the regression parameters $\boldsymbol{\beta}$ by solving the following estimating equations

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{U}_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\theta}_i}{\partial \boldsymbol{\beta}}\right)' \mathbf{V}_i^{-1}\left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right) = \mathbf{0}$$

where $\mathbf{V}_i$ is a working covariance matrix.

Let $\hat{\boldsymbol{\beta}}$ be a solution of the estimating equations. You can use a sandwich estimator to estimate the variance of $\hat{\boldsymbol{\beta}}$. It takes the form

$$\boldsymbol{\Sigma}_e = \mathbf{I}_0^{-1}\mathbf{I}_1\mathbf{I}_0^{-1}$$

$\mathbf{I}_0^{-1}$ is the model-based estimator of $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ and is given by

$$\mathbf{I}_0 = \sum_{i=1}^{n} \frac{\partial \boldsymbol{\theta}_i}{\partial \boldsymbol{\beta}}' \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\theta}_i}{\partial \boldsymbol{\beta}}$$

$\mathbf{I}_1^{-1}$ is the empirical estimator of $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ and is computed as

$$\mathbf{I}_1 = \sum_{i=1}^{n} \mathbf{U}_i(\hat{\boldsymbol{\beta}})'\mathbf{U}_i(\hat{\boldsymbol{\beta}})$$

Andersen, Hansen, and Klein (2004) proposed using pseudovalue regression to analyze the RMST models. Assume τ is a prespecified time point of interest. Let T_i be the time-to-event variable for the ith subject. The RMST models can be fitted using pseudovalue regression by letting

$$\boldsymbol{\theta}_i = \mathrm{RMST}_i(\tau) = E(T_i \wedge \tau | \mathbf{x}_i)$$

$$\mathbf{V}_i = \begin{cases} \theta_i = \mathrm{RMST}_i(\tau) & g(u) = \log(u) \\ 1 & g(u) = u \end{cases}$$

Because the nonparametric estimator $\widehat{\mathrm{RMST}}(\tau)$ is unbiased, it can be used in place of $\hat{\boldsymbol{\theta}}$ in the estimation process.

Inverse Probability Censoring Weighting Estimation

Suppose you can observe the quantities $(T, C, \mathbf{x})$, where T is the event time, C is the censoring time, and $\mathbf{x}$ is a p-dimensional vector of covariates. For the ith subject, $i = 1, \ldots, n$, let $U_i = T_i \wedge C_i$, $\Delta_i = I(T_i \leq C_i)$, and $\mathbf{x}_i$ be the observed time, event indicator, and covariate vector, respectively.

Assume that τ is a prespecified time point of interest and $P(T > \tau) > 0$. Let

$$\begin{aligned} R_i &= T_i \wedge \tau \\ \mathrm{RMST}_i(\tau) &= E(R_i|\mathbf{x}_i) \\ \tilde{\Delta}_i &= I(R_i \leq C_i) \\ w_i &= \frac{\tilde{\Delta}_i}{\hat{G}(R_i)} \end{aligned}$$

where $\hat{G}(t)$ is the Kaplan-Meier estimate (alternatively, the Breslow estimate) of the survival function of the censoring variable, which is calculated using $\{(U_i, 1 - \Delta_i) : i = 1, 2, \ldots, n\}$.

Suppose that the following relationship holds for the RMST,

$$g[\mathrm{RMST}_i(\tau)] = \mathbf{x}_i'\boldsymbol{\beta}$$

where $g(\cdot)$ is a smooth and strictly increasing function. Note that another column can be added to X_i for an intercept effect.

Under suitable regularity conditions, the regression coefficients $\boldsymbol{\beta}$ are estimated by solving the following score function (Tian, Zhao, and Wei 2014):

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \left(R_i - g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}) \right) \mathbf{x}_i = \mathbf{0}$$

Let

$$\hat{\boldsymbol{\Omega}} = \sum_{i=1}^{n} \mathbf{x}_i{}^{\otimes 2} \left(g^{-1}(\mathbf{x}_i'\hat{\boldsymbol{\beta}}) \right)$$

The sandwich variance estimate of $\hat{\boldsymbol{\beta}}$ is

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\Omega}}^{-1} \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\Omega}}^{-1}$$

where $\hat{\boldsymbol{\Sigma}}$ is the empirical variance-covariance matrix of $\mathbf{U}(\hat{\boldsymbol{\beta}})$ that is given by

$$\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{n} (\hat{\boldsymbol{\eta}}_i + \hat{\boldsymbol{\psi}}_i)^{\otimes 2}$$

where

$$\hat{\boldsymbol{\eta}}_i = w_i \left(R_i - g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}) \right) \mathbf{x}_i$$

$$\hat{\boldsymbol{\psi}}_i = \int_0^{\infty} \frac{\hat{\mathbf{q}}(u)}{\pi(u)} d\hat{M}_i^c(u)$$

$$\hat{\mathbf{q}}(u) = \sum_{i=1}^{n} w_i \left(R_i - g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}) \right) \mathbf{x}_i I(U_i \geq u)$$

$$\pi(u) = \sum_j I(U_j \geq u)$$

$$\hat{M}_i^c(t) = I(U_i \leq t, \Delta_i = 0) - \int_0^t I(U_i \geq u) d\hat{\Lambda}^c(u)$$

$$\hat{\Lambda}^c(t) = \int_0^t \frac{dN^c(u)}{\pi(u)}$$

$$N^c(u) = \sum_j I(U_j \leq u, \Delta_j = 0)$$

Estimation with Stratified Weights

Assuming that you have K strata, within each stratum the censoring distribution is homogeneous. For the ith subject, let $B_i \in (1, \ldots, K)$ be the stratum indicator. It is more appropriate to use stratum-specific weights in the estimation. For the kth stratum, you compute the Kaplan-Meier estimate $\hat{G}_k(t)$ for the censoring variable by using $\{(U_i, 1-\Delta_i) : B_i = k, i = 1, 2, \ldots, n\}$.

For the ith subject, the weight is computed as

$$w_i = \frac{\tilde{\Delta}_i}{\hat{G}_{k=B_i}(R_i)}$$

The following quantities are also adjusted accordingly in the estimation:

$$\hat{\mathbf{q}}_k(u) = \sum_{i=1}^{n} w_i \left(R_i - g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}) \right) \mathbf{x}_i \, I(U_i \geq u, B_j = k)$$

$$\pi_k(u) = \sum_j I(U_j \geq u, B_j = k)$$

$$\hat{M}_i^c(t) = I(U_i \leq t, \Delta_i = 0) - \int_0^t I(U_i \geq u) d\hat{\Lambda}_{k=B_i}^c(u)$$

$$\hat{\Lambda}_k^c(t) = \int_0^t \frac{dN_k^c(u)}{\pi_k(u)}$$

$$N_k^c(u) = \sum_j I(U_j \leq u, \Delta_j = 0, B_j = k)$$

Fitting Algorithm

The following is an algorithm for solving generalized estimating equations (GEEs) (Liang and Zeger 1986). The algorithm is generic and can be conveniently adapted to solve the estimating equations from a pseudovalue regression or the inverse probability censoring weighting (IPCW) method.

Typically, the estimating equations of a GEE have the form

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{U}_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} f_i \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)' \mathbf{V}_i^{-1} \left(\mathbf{Y}_i - \boldsymbol{\mu}_i \right) = \mathbf{0}$$

where $\mathbf{Y}_i = (y_{i1}, \ldots, y_{in_i})'$ is the outcome vector for the ith subject, $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})'$ is the model-based means of $\mathbf{Y}_i$, $\mathbf{V}_i$ is the covariance matrix of $\mathbf{Y}_i$, and f_i is the subject-specific weight.

Let $\mathbf{R}_i(\boldsymbol{\alpha})$ be an $n_i \times n_i$ "working" correlation matrix that is fully specified by the vector of parameters $\boldsymbol{\alpha}$. The covariance matrix of $\mathbf{Y}_i$ is modeled as

$$\mathbf{V}_i = \phi \mathbf{A}_i^{\frac{1}{2}} \mathbf{W}_i^{-\frac{1}{2}} \mathbf{R}(\boldsymbol{\alpha}) \mathbf{W}_i^{-\frac{1}{2}} \mathbf{A}_i^{\frac{1}{2}}$$

where $\mathbf{A}_i$ is an $n_i \times n_i$ diagonal matrix with $v(\mu_{ij})$ as the jth diagonal element and $\mathbf{W}_i$ is an $n_i \times n_i$ diagonal matrix with w_{ij} as the jth diagonal, where w_{ij} is a weight that is specified a priori. If $\mathbf{R}_i(\boldsymbol{\alpha})$ is the true correlation matrix of $\mathbf{Y}_i$, then $\mathbf{V}_i$ is the true covariance matrix of $\mathbf{Y}_i$.

The working correlation matrix is usually unknown and must be estimated. It is estimated in the iterative fitting process by using the current value of the parameter vector $\boldsymbol{\beta}$ to compute appropriate functions of the Pearson residual

$$e_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})/w_{ij}}}$$

If you specify the working correlation as $\mathbf{R_0} = \mathbf{I}$, which is the identity matrix, the GEE reduces to the independence estimating equation.

The dispersion parameter ϕ is estimated by

$$\hat{\phi} = \frac{1}{N-p} \sum_{i=1}^{K} f_i \sum_{j=1}^{n_i} e_{ij}^2$$

where $N = \sum_{i=1}^{K} f_i n_i$ is the total number of measurements and p is the number of regression parameters.

To solve the estimating equations, the algorithm proceeds as follows:

1. Computes an initial estimate of $\boldsymbol{\beta}$ by using an ordinary generalized linear model that assumes independence.
2. Computes the working correlations $\mathbf{R}$ on the basis of the standardized residuals, the current $\boldsymbol{\beta}$, and the assumed structure of $\mathbf{R}$.
3. Computes an estimate of the covariance:

$$\mathbf{V}_i = \phi \mathbf{A}_i^{\frac{1}{2}} \mathbf{W}_i^{-\frac{1}{2}} \hat{\mathbf{R}}(\boldsymbol{\alpha}) \mathbf{W}_i^{-\frac{1}{2}} \mathbf{A}_i^{\frac{1}{2}}$$

4. Updates $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}_{r+1} = \boldsymbol{\beta}_r + \left[\sum_{i=1}^{K} f_i \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right]^{-1} \left[\sum_{i=1}^{K} f_i \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) \right]$$

5. Repeats steps 2–4 until convergence.

Missing Values

Observations that contain a missing or negative value in the response are not used in the analysis. If a FREQ variable value is missing or 0, the observation is not used. If a STRATA variable is missing, the observation is not used. If any explanatory variable that you specify in the MODEL statement has a missing value, that observation is not used in the model fitting.

Displayed Output

PROC RMSTREG displays the following information as results of the model fitting.

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, the values that indicate censored times, the link function, the estimation method and inference method used, the name and label of the FREQ variable, and the name and label of the STRATA variable.

Number of Observations

The "Number of Observations" table displays the number of observations that were read and used in the analysis.

CLASS Level Information

If you specify a CLASS statement, PROC RMSTREG outputs the "Class Level Information" table. This table displays the design information for the classification variables when the parameterization is full rank.

Summary of the Number of Event and Censored Values

The "Summary of the Number of Event and Censored Values" table displays a breakdown of the number of events and censored values.

Analysis of Parameter Estimates

The "Analysis of Parameter Estimates" table displays the parameter name, the degrees of freedom for each parameter, the estimate of each parameter, the estimated standard error of the parameter estimator, confidence limits for each parameter, a chi-square statistic for testing whether the parameter is 0, and the associated p-value for the statistic.

Iteration History for Parameter Estimates

If you specify the ITHISTORY option in the PROC RMSTREG statement, the procedure outputs a table that contains the following for each iteration in the iterative process of model fitting: the iteration number, the maximum gradient, and the values of all parameters in the model.

Last Evaluation of the Gradient

If you specify the ITHISTORY option in the PROC RMSTREG statement, the procedure displays the last evaluation of the gradient vector.

Convergence Status

The "Convergence Status" table displays the convergence status of the iterative estimation routine.

Estimated Covariance Matrix

If you specify the COVB option in the MODEL statement, the procedure displays the estimated covariance matrix. This matrix is defined as the inverse of the information matrix at the final iteration and is based on the Hessian matrix that is used at the final iteration.

Estimated Correlation Matrix

If you specify the CORRB option in the MODEL statement, the procedure displays the estimated correlation matrix, which is based on the Hessian matrix that is used at the final iteration.

Type 3 Tests

The "Type 3 Tests" table is displayed if the model contains a CLASS variable or if you specify the TYPE3 option in the MODEL statement. The table displays, for each specified statistic, the Type 3 chi-square, the degrees of freedom, and the *p*-value for each effect in the model.

ODS Table Names

PROC RMSTREG assigns a name to each table that it creates. You can use these names to refer to the tables when you use the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in Table 103.10. For more information about ODS, see Chapter 20, "Using the Output Delivery System."

Table 103.10 ODS Tables Produced by PROC RMSTREG

ODS Table Name	Description	Statement	Option
ClassLevelInfo	Design information for CLASS variables	CLASS	
Coef	Coefficients for LS-means	LSMEANS	E
ConvergenceStatus	Convergence status	MODEL	
CorrB	Parameter estimate correlation matrix	MODEL	CORRB
CovB	Parameter estimate covariance matrix	MODEL	COVB
Diffs	Differences of LS-means	LSMEANS	DIFF
IterHistory	Iteration history	PROC RMSTREG	ITHISTORY
LastGradient	Last evaluation of the gradient	PROC RMSTREG	ITHISTORY
LSMeans	LS-means	LSMEANS	Default
LSMLines	Lines display for LS-means	LSMEANS	LINES
ModelInfo	Model and data information	MODEL	
NObs	Number of observations	MODEL	
ParameterEstimates	Parameter estimates	MODEL	
Type3Test	Type 3 tests	MODEL	

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 RMSTREG procedure produces plots through the LSMEANS statement.

Examples: RMSTREG Procedure

The following examples illustrate some of the capabilities of the RMSTREG procedure. These examples are not intended to represent definitive analyses of the data sets that are presented here.

Example 103.1: Comparison of PROC RMSTREG with Other Procedures

This example compares the RMSTREG procedure with the PHREG and LIFEREG procedures.

Krall, Uthoff, and Harley (1975) analyzed data from a study on multiple myeloma in which researchers treated 65 patients by using alkylating agents. Of those patients, 48 died during the study and 17 survived. The following DATA step creates the data set Myeloma:

```
data Myeloma;
   input Time VStatus LogBUN HGB Platelet Age LogWBC Frac
         LogPBM Protein SCalc;
   label Time='Survival Time'
         VStatus='0=Alive 1=Dead';
   datalines;
 1.25  1  2.2175   9.4  1  67  3.6628  1  1.9542  12  10
 1.25  1  1.9395  12.0  1  38  3.9868  1  1.9542  20  18
 2.00  1  1.5185   9.8  1  81  3.8751  1  2.0000   2  15
 2.00  1  1.7482  11.3  0  75  3.8062  1  1.2553   0  12
 2.00  1  1.3010   5.1  0  57  3.7243  1  2.0000   3   9
 3.00  1  1.5441   6.7  1  46  4.4757  0  1.9345  12  10
 5.00  1  2.2355  10.1  1  50  4.9542  1  1.6628   4   9

   ... more lines ...

53.00  0  1.1139  12.0  1  66  3.6128  1  2.0000   1  11
57.00  0  1.2553  12.5  1  66  3.9685  0  1.9542   0  11
77.00  0  1.0792  14.0  1  60  3.6812  0  0.9542   0  12
;
```

The variable Time represents the survival time in months from diagnosis. The variable VStatus indicates whether the patient survived the study. If the value of VStatus is 0, then the patient survived and thus the corresponding value of Time is censored. The variables that are thought to be related to survival are LogBUN (log(BUN) at diagnosis), HGB (hemoglobin at diagnosis, in g/dl), Platelet (platelets at diagnosis: 0=abnormal, 1=normal), Age (age at diagnosis, in years), LogWBC (log(WBC) at diagnosis), Frac (fractures at diagnosis: 0=none, 1=present), LogPBM (log percentage of plasma cells in bone marrow), Protein (proteinuria at diagnosis, in mg/dl), and SCalc (serum calcium at diagnosis, in mg/dl). The goal of the study is to identify important prognostic factors from these nine explanatory variables.

Suppose that T_i represents the survival time of patient i and μ_i represents the RMST at the specified time τ for the patient. A log-linear model for the RMST assumes that

$$\log(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$$

The following statements fit the log-linear model for the RMST at $\tau = 50$ with covariates LogBUN and HGB:

```
proc rmstreg data=Myeloma tau=50;
   model Time*VStatus(0)=LogBUN HGB / method=ipcw link=log;
run;
```

Output 103.1.1 displays the parameter estimates for the fitted log-linear regression model.

Output 103.1.1 Parameter Estimates

The RMSTREG Procedure

Analysis of Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	4.2187	0.6091	3.0249	5.4126	47.97	<.0001
LogBUN	1	-1.0925	0.3053	-1.6908	-0.4942	12.81	0.0003
HGB	1	0.0438	0.0416	-0.0378	0.1254	1.11	0.2923

The following statements fit the linear model with covariates LogBUN and HGB:

```
proc rmstreg data=Myeloma tau=50;
   model Time*VStatus(0)=LogBUN HGB / method=ipcw link=linear;
   store rr;
run;
```

Output 103.1.2 displays the resulting parameter estimates for the fitted linear regression model.

Output 103.1.2 Parameter Estimates

The RMSTREG Procedure

Analysis of Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	45.2318	14.0717	17.6518	72.8118	10.33	0.0013
LogBUN	1	-23.1680	5.5614	-34.0681	-12.2679	17.35	<.0001
HGB	1	1.1493	1.0259	-0.8615	3.1601	1.25	0.2626

For comparison, a Cox proportional hazards (PH) model is fitted to the data, using the same set of covariates. The PH model assumes that

$$\lambda(t|\mathbf{x}_i) = \lambda_0(t)\exp(\mathbf{x}_i{}'\boldsymbol{\beta}_{\mathrm{PH}})$$

where $\lambda_0(t)$ is the baseline hazard function and $\boldsymbol{\beta}_{\mathrm{PH}}$ is a vector of unknown regression parameters.

The following PROC PHREG statements fit a Cox PH model with covariates LogBUN and HGB:

```
proc phreg data=Myeloma;
   model Time*VStatus(0)=LogBUN HGB;
   store ph;
run;
```

Output 103.1.3 displays the parameter estimates for the fitted Cox regression model.

Output 103.1.3 Parameter Estimates from Cox Regression

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
LogBUN	1	1.67440	0.61209	7.4833	0.0062	5.336
HGB	1	-0.11899	0.05751	4.2811	0.0385	0.888

If you compare the results from PROC RMSTREG with results from PROC PHREG, you see that the estimated effects have opposite signs. This is because Cox regression models the hazard function, and RMST regression models the restricted mean at a certain time via a log link. Thus, you interpret the estimated effects from PROC PHREG as ratios of the hazard functions; you interpret the estimated effects from PROC RMSTREG as log ratios of the RMST. An increase in the hazard rate would correspond roughly to a decrease in the restricted mean, although the exact relationship is difficult to quantify (Karrison 1987).

Also, note that the effect of HGB is significant at the 5% level in the PROC PHREG results but not significant in the PROC RMSTREG results. This might be due to the particular time τ that is specified for the RMST analysis, and it might also be due to the different characteristics of PROC PHREG's likelihood-based approach versus PROC RMSTREG's approach, which uses estimating equations.

The RMST regression model can also be compared to the accelerated failure time (AFT) model fitted by the LIFEREG procedure. The AFT model assumes that

$$\log(T_i) = \mathbf{x}_i\boldsymbol{\beta}_{\mathrm{AFT}} + \sigma\epsilon_i$$

where $\boldsymbol{\beta}_{\mathrm{AFT}}$ is a vector of unknown regression parameters, ϵ_i is the error term that is sampled from a known distribution, and σ is an unknown scale parameter.

The following PROC LIFEREG statements fit the AFT model with a Weibull distribution, using the same set of covariates as in the previous examples:

```
proc lifereg data=Myeloma;
   model Time*VStatus(0)=LogBUN HGB;
   store aft;
run;
```

Output 103.1.4 displays the parameter estimates for the fitted AFT regression model.

Output 103.1.4 Parameter Estimates from AFT Model

The LIFEREG Procedure

Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	4.5458	0.8939	2.7937	6.2979	25.86	<.0001
LogBUN	1	-1.5304	0.5031	-2.5165	-0.5444	9.25	0.0024
HGB	1	0.0975	0.0492	0.0011	0.1939	3.93	0.0474
Scale	1	0.8770	0.0945	0.7100	1.0833		
Weibull Shape	1	1.1403	0.1229	0.9231	1.4085		

The estimated effects from the AFT model exhibit the same pattern as the estimated effects from the log-linear model of PROC RMSTREG. However, the interpretations of the two models are quite different. Instead of modeling a specific measure, as PROC RMSTREG does, the AFT model assumes that the effect of the independent variables on the event time is additive on the log scale. As with the Cox regression model fit by PROC PHREG, it is difficult to translate the estimated effects from PROC LIFEREG to the scale of the RMST model.

Example 103.2: Comparing the RMSTs of Two Groups

This example illustrates how to use PROC RMSTREG to compare the RMSTs of two groups while adjusting for other factors.

In a study of the human immunodeficiency virus (HIV), subjects were followed after a confirmed HIV-positive diagnosis (Hosmer and Lemeshow 1999). The primary goal is to evaluate the effect of two different covariates on mortality: the subject's age and history of prior intravenous drug use.

The following DATA step creates the data set HIV, which contains the variables Time (the follow-up time in days), Status (with a value of 0 if Time was censored and 1 otherwise), Drug (with a value of 1 for prior intravenous drug use and 0 otherwise), and Age (the patient's age in years at the beginning of the follow-up):

```
data HIV;
   input Time Age Drug Status;
   datalines;
    5      46      0        1
    6      35      1        0
    8      30      1        1
    3      30      1        1
   22      36      0        1
    1      32      1        0

   ... more lines ...

    1      34      1        1
;
```

The following statements fit a linear model for the RMST at $\tau = 48$:

```
proc rmstreg data=hiv tau=48;
   class Drug;
   model Time*Status(0) = Drug Age / link=linear;
   lsmeans Drug;
run;
```

Output 103.2.1 shows the "Model Information" table, which provides information about the specified linear model of the RMST and the input data set.

Output 103.2.1 Model Information

The RMSTREG Procedure

Model Information	
Data Set	WORK.HIV
Time Variable	Time
Censoring Variable	Status
Censoring Value(s)	0
Link Function	Linear
Estimation Method	Pseudo Value
Tau Value	48

Output 103.2.2 displays the parameter estimates for the fitted linear model:

Output 103.2.2 Parameter Estimates

Analysis of Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	47.3681	9.0332	29.6633	65.0728	27.50	<.0001
Drug	0	1	10.1327	2.8592	4.5288	15.7366	12.56	0.0004
Drug	1	0	0.0000					
Age		1	-1.0778	0.2342	-1.5367	-0.6188	21.18	<.0001

The results show both age and drug use history are strongly associated with mortality.

The following statements fit the linear model for the RMST at $\tau = 48$ by using the method of inverse probability censoring weighting (IPCW) to illustrate that estimation method.

```
proc rmstreg data=hiv tau=48;
   class Drug;
   model Time*Status(0) = Drug Age / link=linear method=ipcw;
   lsmeans Drug;
run;
```

Output 103.2.3 displays the parameter estimates for the fitted linear model:

Output 103.2.3 Parameter Estimates

The RMSTREG Procedure

Analysis of Parameter Estimates

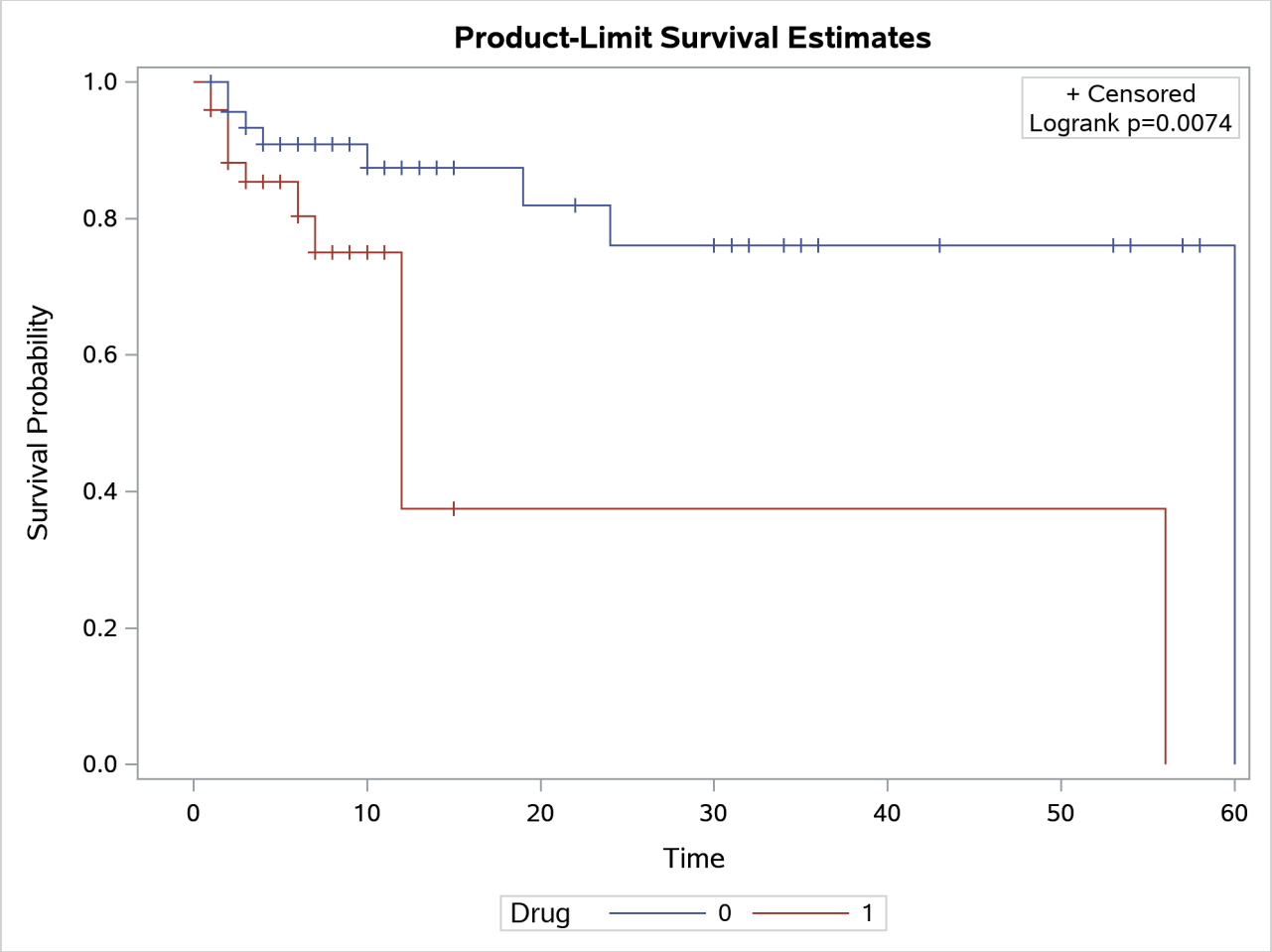
Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	44.6305	8.8479	27.2889	61.9720	25.44	<.0001
Drug	0	1	12.5483	2.6029	7.4468	17.6499	23.24	<.0001
Drug	1	0	0.0000					
Age		1	-1.0748	0.2289	-1.5234	-0.6262	22.05	<.0001

This method assumes homogeneity of the censoring mechanism. Such an assumption can be evaluated. The following statements use PROC LIFETEST to estimate the survival functions of censoring for the two levels of Drug. Note that the censoring value in the TIME statement uses 1 instead of 0.

```
ods graphics on;
proc lifetest data=hiv plot=s(test);
   strata Drug;
   time Time*Status(1);
run;
```

Output 103.2.4 displays the Kaplan-Meier curves for the two Drug groups.

Output 103.2.4 Plot of Estimated Survival Functions



The curves show substantial separation over time. The small p-value from the log-rank test also suggests that the censoring patterns of the two groups are quite different.

By default, the IPCW method uses the Kaplan-Meier technique to obtain the weights, and this approach implicitly assumes that the censoring mechanism is homogeneous among all subjects. Sometimes this is not a reasonable assumption—for example, if there are distinct groups, such as treatment arms in randomized clinical trials. Under such circumstances, it is more appropriate to use group-specific weights by applying the Kaplan-Meier technique to different groups.

The following statements fit the linear model for the RMST at $\tau = 48$ by using the IPCW technique, with weights estimated separately for the two Drug groups:

```
proc rmstreg data=hiv tau=48;
   class Drug;
   model Time*Status(0) = Drug Age / link=linear method=ipcw(strata=Drug);
   lsmeans Drug;
run;
```

Output 103.2.5 displays the parameter estimates for the fitted linear model of the RMST.

Output 103.2.5 Parameter Estimates

The RMSTREG Procedure

Analysis of Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	47.5159	10.5939	26.7522	68.2795	20.12	<.0001
Drug	0	1	11.4049	2.5667	6.3743	16.4354	19.74	<.0001
Drug	1	0	0.0000					
Age		1	-1.1380	0.2654	-1.6582	-0.6179	18.39	<.0001

The parameter estimates are similar to those displayed in Output 103.2.3.

Example 103.3: Making Model-Based Inferences

This example reconsiders the Liver data set introduced in the section "Getting Started: RMSTREG Procedure" on page 8615. In this case, you are interested in studying RMST differences over different levels of edema, at $\tau = 10$.

The following statements fit a log-linear model for the RMST at $\tau = 10$ with covariates Bilirubin, Age, and Edema:

```
proc rmstreg data=liver tau=10;
   class Edema;
   model Time*Status(0) = Age Bilirubin Edema / link=log method=pv;
   estimate '0 vs 0.5' Edema 1 -1  0 / exp cl;
   estimate '0.5 vs 1' Edema 0  1 -1 / exp cl;
   estimate '0 vs 1'   Edema 1  0 -1 / exp cl;
run;
```

The METHOD=PV option in the MODEL statement specifies pseudovalue regression to fit the model. The ESTIMATE statements compute differences in predicted RMST on the log scale for different pairs of levels of the variable Edema. The EXP option in the ESTIMATE statements exponentiates these differences to form ratio estimates of the RMST. Standard errors and confidence intervals are also computed.

Figure 103.3.1 displays the parameter estimates from the fitted model.

Output 103.3.1 Parameter Estimates

The RMSTREG Procedure

Analysis of Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	1.7815	0.3241	1.1462	2.4168	30.21	<.0001
Age		1	-0.0098	0.0020	-0.0138	-0.0058	23.27	<.0001
Bilirubin		1	-0.0848	0.0115	-0.1074	-0.0621	53.99	<.0001
Edema	0	1	0.9362	0.2952	0.3576	1.5148	10.06	0.0015
Edema	0.5	1	0.7356	0.3083	0.1314	1.3398	5.69	0.0170
Edema	1	0	0.0000					

Output 103.3.2 displays the results of the ESTIMATE statement.

Output 103.3.2 Ratio Estimates

Estimate										
Label	Estimate	Standard Error	z Value	Pr > |z|	Alpha	Lower	Upper	Exponentiated	Exponentiated Lower	Exponentiated Upper
0 vs 0.5	0.2006	0.09844	2.04	0.0416	0.05	0.007668	0.3936	1.2222	1.0077	1.4822

Estimate										
Label	Estimate	Standard Error	z Value	Pr > |z|	Alpha	Lower	Upper	Exponentiated	Exponentiated Lower	Exponentiated Upper
0.5 vs 1	0.7356	0.3083	2.39	0.0170	0.05	0.1314	1.3398	2.0867	1.1404	3.8182

Estimate										
Label	Estimate	Standard Error	z Value	Pr > |z|	Alpha	Lower	Upper	Exponentiated	Exponentiated Lower	Exponentiated Upper
0 vs 1	0.9362	0.2952	3.17	0.0015	0.05	0.3576	1.5148	2.5503	1.4299	4.5487

The results indicate that the different levels of Edema have significantly different predicted RMST values.

An analysis that uses LS-means can provide more detail for these Edema differences in predicted RMST. In the following statements, the LSMEANS statement is specified with several options: the E option displays the coefficients that are used to compute the LS-means for each Treatment level; the DIFF option takes all pairwise differences of the LS-means for the levels of the Treatment variable; the ILINK option exponentiates these differences to form ratio estimates; the CL option produces confidence intervals for the differences and ratios; and the ADJUST=BON option performs a very conservative adjustment of the *p*-values and confidence intervals.

```
proc rmstreg data=liver tau=10;
   class Edema;
   model Time*Status(0) = Age Bilirubin Edema / link=log method=pv;
   lsmeans Edema / e diff ilink cl adjust=bon;
run;
```

The results from the LSMEANS statement are displayed in Output 103.3.3 through Output 103.3.5.

The LS-means are computed by constructing each of the $\boldsymbol{l}$ coefficient vectors shown in Output 103.3.3 and then computing $\boldsymbol{l}'\boldsymbol{\beta}$. The LS-means are not estimates of the RMST; they are estimates of the linear predictors on the logit scale and therefore are estimated log differences. To obtain the RMST, you need to apply the inverse-link transformation by specifying the ILINK option in the LSMEANS statement. For more information about the construction of LS-means, see the section "Construction of Least Squares Means" on page 4056 in Chapter 50, "The GLM Procedure."

Output 103.3.3 Edema LS-Means Coefficients

The RMSTREG Procedure

Coefficients for Edema Least Squares Means

Parameter	Edema	Row1	Row2	Row3
Intercept		1	1	1
Age		50.741	50.741	50.741
Bilirubin		3.2208	3.2208	3.2208
Edema 0	0	1		
Edema 0.5	0.5		1	
Edema 1	1			1

The Edema LS-means shown in Output 103.3.4 are all significantly nonzero at the 0.05 level. These LS-means are *predicted population margins* of the RMST; that is, they estimate the marginal means over a balanced population, and they are effectively the within-Edema means appropriately adjusted for the other effects in the model.

Output 103.3.4 Edema LS-Means

Edema Least Squares Means

Edema	Estimate	Standard Error	z Value	Pr > |z|	Alpha	Lower	Upper	Mean	Standard Error of Mean	Lower Mean	Upper Mean
0	1.9480	0.02857	68.19	<.0001	0.05	1.8920	2.0040	7.0147	0.2004	6.6327	7.4187
0.5	1.7474	0.09663	18.08	<.0001	0.05	1.5580	1.9368	5.7396	0.5546	4.7494	6.9364
1	1.0118	0.2914	3.47	0.0005	0.05	0.4406	1.5830	2.7505	0.8016	1.5536	4.8696

Pairwise differences between the Edema LS-means, requested using the DIFF option, are displayed in Output 103.3.5. The LS-mean for the level that is displayed in the _Edema column is subtracted from the LS-mean for the level in the Edema column, so the first row displays the LS-mean for Edema level 0 minus the LS-mean for Edema level 0.5. The difference (0.2006) is the estimated difference in log scale, or equivalently the log ratio of the two Edema levels. The Pr > |z| column indicates that the 0 and 0.5 levels are not significantly different; however, both levels are significantly different from level 1. If the inverse-link transformation is specified by the ILINK option, then these differences do not transform back to differences in the RMST.

Because multiple tests are performed, you can protect yourself from falsely significant results by adjusting your *p*-values for multiplicity. The ADJUST=BON option performs the conservative Bonferroni adjustment and adds the columns labeled with 'Adj' to Output 103.3.5. By comparing the Pr > |z| column to the Adj P column, you can see that the *p*-values are adjusted upward and two of the three values are now greater than the alpha level. The confidence intervals are also adjusted for multiplicity—all adjusted intervals are wider than the unadjusted intervals, but the results are unchanged. You can specify other adjustment methods by using the ADJUST= option.

Output 103.3.5 Differences and RMST Ratios for the Edema LS-Means

Differences of Edema Least Squares Means
Adjustment for Multiple Comparisons: Bonferroni

Edema	_Edema	Estimate	Standard Error	z Value	Pr > |z|	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
0	0.5	0.2006	0.09844	2.04	0.0416	0.1247	0.05	0.007668	0.3936	-0.03506	0.4363
0	1	0.9362	0.2952	3.17	0.0015	0.0046	0.05	0.3576	1.5148	0.2295	1.6430
0.5	1	0.7356	0.3083	2.39	0.0170	0.0511	0.05	0.1314	1.3398	-0.00237	1.4736

If you want to jointly test whether the three Edema levels are different, you can specify a custom hypothesis test among LS-means by using the LSMESTIMATE statement. In the following statements, the LS-means for the two Edema levels are contrasted with the LS-mean for the third level, and the JOINT option performs a joint test that the three Edema levels are the same:

```
proc rmstreg data=liver tau=10;
   class Edema;
   model Time*Status(0) = Age Bilirubin Edema / link=log method=pv;
   lsmestimate Edema 1 0 -1, 0 1 -1 / joint;
run;
```

Output 103.3.6 displays the results from the LSMESTIMATE statement. The "Least Squares Means Estimates" table displays the requested differences, and the results are identical to the second and third rows of Output 103.3.5. The "Chi-Square Test for Least Squares Means Estimates" table displays the joint test. In all these tests, you reject the null hypothesis that the three Edema levels are the same.

Output 103.3.6 Custom LS-Mean Tests

The RMSTREG Procedure

Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	z Value	Pr > |z|
Edema	Row 1	0.9362	0.2952	3.17	0.0015
Edema	Row 2	0.7356	0.3083	2.39	0.0170

Chi-Square Test for Least Squares Means Estimates

Effect	Num DF	Chi-Square	Pr > ChiSq
Edema	2	13.82	0.0010

References

Andersen, P. K., Hansen, M. G., and Klein, J. P. (2004). "Regression Analysis of Restricted Mean Survival Time Based on Pseudo-observations." *Lifetime Data Analysis* 10:335–350.

Andersen, P. K., Klein, J. P., and Rosthøj, S. (2003). "Generalised Linear Models for Correlated Pseudo-observations, with Applications to Multi-state Models." *Biometrika* 90:15–27.

Hosmer, D. W., Jr., and Lemeshow, S. (1999). *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. New York: John Wiley & Sons.

Karrison, T. (1987). "Restricted Mean Life with Adjustment for Covariates." *Journal of the American Statistical Association* 82:1169–1176.

Klein, J. P., and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd ed. New York: Springer-Verlag.

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

Liang, K.-Y., and Zeger, S. L. (1986). "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73:13–22.

Lin, D. Y., Wei, L. J., and Ying, Z. (1993). "Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals." *Biometrika* 80:557–572.

Littell, R. C., Freund, R. J., and Spector, P. C. (1991). *SAS System for Linear Models*. 3rd ed. Cary, NC: SAS Institute Inc.

Muller, K. E., and Fetterman, B. A. (2002). *Regression and ANOVA: An Integrated Approach Using SAS Software*. Cary, NC: SAS Institute Inc.

Nelder, J. A., and Wedderburn, R. W. M. (1972). "Generalized Linear Models." *Journal of the Royal Statistical Society, Series A* 135:370–384.

Royston, P., and Parmar, M. K. B. (2013). "Restricted Mean Survival Time: An Alternative to the Hazard Ratio for the Design and Analysis of Randomized Trials with a Time-to-Event Outcome." *BMC Medical Research Methodology* 13:152–166.

Tian, L., Zhao, L., and Wei, L. J. (2014). "Predicting the Restricted Mean Event Time with the Subject's Baseline Covariates in Survival Analysis." *Biostatistics* 15:222–233.

Trinquart, L., Jacot, J., Conner, S. C., and Porcher, R. (2016). "Comparison of Treatment Effects Measured by the Hazard Ratio and by the Ratio of Restricted Mean Survival Times in Oncology Randomized Controlled Trials." *Journal of Clinical Oncology* 34:1813–1819.

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., et al. (2014). "Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis." *Journal of Clinical Oncology* 32:2380–2385.

Zucker, D. M. (1998). "Restricted Mean Life with Covariates: Modification and Extension of a Useful Survival Analysis Method." *Journal of the American Statistical Association* 93:702–709.

Subject Index

Syntax Index