SAS/STAT® 15.1 User's Guide

The ICPHREG 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 67
The ICPHREG Procedure

Contents

Overview: ICPHREG Procedure

The ICPHREG procedure is designed to fit proportional hazards regression models to interval-censored data. It can also fit proportional hazards regression models to failure time data that are uncensored, right censored, or left censored. The survival time of each member of a population is assumed to follow its own hazard function, $\lambda_i(t)$, which is expressed as

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

where $\lambda_0(t)$ is the baseline hazard function, $\mathbf{Z}_i$ is the vector of explanatory variables for the ith individual, and $\boldsymbol{\beta}$ is the vector of unknown regression coefficients that is associated with the explanatory variables. The vector $\boldsymbol{\beta}$ is assumed to be the same for all individuals.

The ICPHREG procedure enables you to use a variety of configurations with respect to the baseline function to fit a proportional hazards model; these configurations include a piecewise constant model (Friedman 1982), a cubic spline model (Royston and Parmar 2002), and a semiparametric model (Finkelstein 1986). To estimate the regression coefficients and the baseline parameters, the ICPHREG procedure maximizes the full likelihood instead of the Cox partial likelihood. Standard errors of the estimates are obtained by inverting the observed information matrix, which is derived from the full likelihood.

The ICPHREG procedure also enables you to do the following: include an offset variable in the model, weight the observations in the input data, test linear hypotheses about the regression coefficients, compute customized hazard ratios, and estimate and plot the survival function and the cumulative hazard function for a new set of covariates.

The ICPHREG 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 the PHREG Procedure

The ICPHREG procedure compares most closely to the PHREG procedure. Both procedures can fit proportional hazards models. They differ in the types of censored data that they are designed to handle and the forms of the baseline function. Table 67.1 matches the procedures with the types of censored data they can analyze and the form of the baseline function.

Table 67.1 Proportional Hazards Modeling

Data Type	Baseline	PROC ICPHREG	PROC PHREG
Interval-censored	Piecewise constant	Y	
	Cubic splines	Y	
	Unspecified	Y	
Right-censored	Piecewise constant	Y	Y
	Cubic splines	Y	
	Unspecified		Y

The PHREG procedure deals exclusively with right-censored data, and it mainly adopts a *semiparametric* approach by leaving the baseline hazard function unspecified. The ICPHREG procedure is specifically designed to handle interval-censored data and offers different options to parameterize the baseline hazard function. You can use the ICPHREG procedure to analyze data that are left-censored, interval-censored, or right-censored. However, if the data to be analyzed contain only exact or right-censored observations, it is recommended that you use the PHREG procedure because it provides specialized methods for dealing with right-censored data. For more information about PROC PHREG, see Chapter 89, "The PHREG Procedure."

Both the LIFEREG procedure and the ICPHREG procedure can handle interval-censored data. The LIFEREG procedure focuses on parametric analysis that uses accelerated failure time models, and it can fit only a proportional hazards model that assumes a Weibull baseline hazard function.

Getting Started: ICPHREG Procedure

This example demonstrates how you can fit a proportional hazards model on an interval-censored data set. By default, PROC ICPHREG uses a piecewise constant baseline hazard to fit the model.

The AIDS data (Larder, Darby, and Richman 1989) consist of observations from 31 patients who were followed up for the development of drug resistance to zidovudine. The following DATA step creates the SAS data set HIV:

```
data hiv;
   input Left Right Stage Dose CdLow CdHigh;
   if (Left=0) then Left=.;
   if (Right>=26) then Right=.;
   datalines;
0 16 0 0 0 1
15 26 0 0 0 1
12 26 0 0 0 1
17 26 0 0 0 1
13 26 0 0 0 1
0 24 0 0 1 0
6 26 0 1 1 0
0 15 0 1 1 0
14 26 0 1 1 0
12 26 0 1 1 0
13 26 0 1 0 1
12 26 0 1 1 0
12 26 0 1 1 0
0 18 0 1 0 1
0 14 0 1 0 1
0 17 0 1 1 0
0 15 0 1 1 0
3 26 1 0 0 1
4 26 1 0 0 1
1 11 1 0 0 1
13 19 1 0 0 1
0 6 1 0 0 1
```

```
0 11 1 1 0 0
6 26 1 1 0 0
0 6 1 1 0 0
2 12 1 1 0 0
1 17 1 1 1 0
0 14 1 1 0 0
0 25 1 1 0 1
2 11 1 1 0 0
0 14 1 1 0 0
;
```

The data set HIV contains the variables Left and Right, which are the starting time and ending time, both in months since the start of study; the variable Stage, which indicates the stage of disease (early (0) or late (1)); the variable Dose, a binary variable that indicates whether the dose is low (0) or high (1); the variable CdLow, which indicates whether the CD4 lymphocyte count is less than 100; and the variable CdHigh, which indicates that a count greater than or equal to 400 is recorded.

The following statements use PROC ICPHREG to fit a proportional hazards model to these data:

```
proc icphreg data=hiv;
   class Stage Dose / desc;
   model (Left, Right) = Stage Dose;
run;
```

The CLASS statement specifies that the variables Stage and Dose are classification variables. The DESC option sets the lower formatted value as the reference level for each CLASS variable. The MODEL statement specifies that the observed intervals are formed by Left and Right.

By default, the preceding statements produce information about the input data and the fitted model, as shown in Figure 67.1.

Figure 67.1 Model and Data Information from the ICPHREG Procedure

The ICPHREG Procedure

Model Information	
Data Set	WORK.HIV
Left Boundary	Left
Right Boundary	Right
Baseline Hazard	Piecewise Constant

Number of Observations Read	31
Number of Observations Used	31
Right Censored Observations	13
Interval Censored Observations	5
Left Censored Observations	13

Figure 67.1 shows 13 left-censored observations, 13 right-censored observations, and 5 interval-censored observations.

Figure 67.2 displays the "Class Level Information" table, which identifies the levels of the classification variables that are used in the model.

Figure 67.2 CLASS Variables Information from the ICPHREG Procedure

Class Level Information		
Name	**Levels**	**Values**
Stage	2	1 0
Dose	2	1 0

By default, PROC ICPHREG uses a baseline hazard that is partitioned into five disjoint intervals to fit a proportional hazards model. Figure 67.3 displays details about this partition.

Figure 67.3 Interval Partition

Constant Hazard Time Intervals		
Interval		
[Lower	Upper)	Hazard Parameter
0	5.5	Haz1
5.5	8	Haz2
8	12.5	Haz3
12.5	17	Haz4
17	Infty	Haz5

PROC ICPHREG determines the break points so that each time interval contains approximately an equal number of imputed middle points and boundary values in the input data set after excluding the right-censored observations. For more information about this method, see the section "Choosing Break Points" on page 5167. You can supply your own partition by using the INTERVALS= option in the MODEL statement.

The "Fit Statistics" table, shown in Figure 67.4, contains several statistics that summarize how well the model fits the data. These statistics are helpful in judging the adequacy of a model and in comparing it with other models under consideration.

Figure 67.4 Model Fit Statistics from the ICPHREG Procedure

Fit Statistics	
-2 Log Likelihood	21.813
AIC (Smaller is Better)	31.813
AICC (Smaller is Better)	34.213
BIC (Smaller is Better)	38.983

The table of parameter estimates is displayed in Figure 67.5. The columns display the parameter name, the degrees of freedom that are associated with the parameter, the estimated parameter value, the standard error of the parameter estimate, the confidence limits, the Wald chi-square statistic, and the associated p-value for testing the significance of the parameter. If a parameter has been fixed during the optimization process, or if a column of the Hessian matrix that corresponds to that parameter is found to linearly depend on columns that correspond to proceeding model parameters, PROC ICPHREG assigns zero degrees of freedom to that parameter and displays a value of zero for its standard error.

Figure 67.5 Model Parameter Estimates from the ICPHREG Procedure

Analysis of Maximum Likelihood Parameter Estimates

Effect	Stage	Dose	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Haz1			0	0.0000					
Haz2			1	0.0167	0.0205	0.0000	0.0568		
Haz3			0	0.0000					
Haz4			1	0.0842	0.0655	0.0000	0.2126		
Haz5			1	2.5641	366.4263	0.0000	720.7464		
Stage	1		1	2.9597	0.9358	1.1255	4.7939	10.00	0.0016
Stage	0		0	0.0000					
Dose		1	1	1.6229	0.8410	-0.0255	3.2713	3.72	0.0537
Dose		0	0	0.0000					

Two types of parameters are present in Figure 67.5: the hazard parameters (Haz1, Haz2, ..., Haz5) and the regression coefficients for the covariates. PROC ICPHREG does not display the chi-square statistic and associated *p*-value for the hazard parameters.

Two of the hazard parameters are constrained at 0, a sign of overparameterization that results from too many hazard parameters in the model. For more information about how the constraints are constructed, see the section "NOPOLISH" on page 5160. You can use fewer break points to fit the model by using the NINTERVAL= option or the INTERVALS= option. For example, the following statements request a model that has exactly two hazard parameters by specifying one break point at 10:

```
proc icphreg data=hiv ithistory;
   class Stage Dose / desc;
   model (Left, Right) = Stage Dose / basehaz=pch(intervals=(10));
run;
```

The table of parameter estimates is displayed in Figure 67.6. None of the hazard parameters are constrained.

Figure 67.6 Model Parameter Estimates from the ICPHREG Procedure

The ICPHREG Procedure

Analysis of Maximum Likelihood Parameter Estimates

Effect	Stage	Dose	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Haz1			1	0.0042	0.0051	0.0000	0.0142		
Haz2			1	0.0590	0.0360	0.0000	0.1296		
Stage	1		1	2.0810	0.7298	0.6506	3.5114	8.13	0.0044
Stage	0		0	0.0000					
Dose		1	1	1.0907	0.6766	-0.2354	2.4167	2.60	0.1069
Dose		0	0	0.0000					

The ITHISTORY option outputs the iteration history of the fitting algorithm, which is shown in Figure 67.7. This option also produces the gradient and Hessian of the likelihood function at the last evaluation. In Figure 67.7, all values of the gradient are close to zero.

Figure 67.7 Iteration History from the ICPHREG Procedure

Likelihood Optimization Iteration History

					Parameter Values				Gradient Values			
Iteration	Evaluations	-2 Log Likelihood	Change	Max Gradient	Stage1	Dose1	Haz1	Haz2	Stage1	Dose1	Haz1	Haz2
0	2	47.8895	.	74.9948	0	0	0.1245	0.0741	-1.8827	3.6394	74.9948	5.0972
1	10	39.1668	-8.7227	47.5072	0.1816	0.3012	0.0559	0.0848	-3.6580	0.5057	47.5072	-1.3723
2	6	37.6893	-1.4775	32.3796	0.2453	0.3347	0.0440	0.0944	-3.9790	-0.2883	32.3796	-1.7277
3	3	35.3576	-2.3317	50.9346	0.4701	0.4206	0.0185	0.1168	-4.6140	-2.0073	-50.9346	-3.9286
4	3	32.5990	-2.7586	6.4037	0.8039	0.5075	0.0200	0.1003	-2.8566	-0.7001	6.4037	0.3183
5	3	30.1026	-2.4964	94.9226	1.3800	0.7032	0.00785	0.0839	-2.1446	-1.3702	-94.9226	-6.0631
6	3	29.0224	-1.0802	23.8304	1.7623	0.8588	0.00724	0.0699	-0.3881	-0.2204	23.8304	-0.6369
7	3	28.8561	-0.1663	103.7	2.0115	1.0376	0.00384	0.0622	-0.5201	-0.4868	-103.7	-2.6791
8	3	28.7697	-0.0863	1.9794	2.0740	1.0858	0.00418	0.0593	-0.0194	-0.0145	-1.9794	-0.0585
9	3	28.7696	-0.00013	0.00309	2.0799	1.0898	0.00416	0.0590	-0.00170	-0.00121	0.00309	-0.00077
10	3	28.7696	-3.01E-6	0.000803	2.0809	1.0906	0.00416	0.0590	-0.00015	-0.00012	-0.00080	-0.00004
11	3	28.7696	-2.5E-8	6.991E-6	2.0810	1.0907	0.00416	0.0590	-6.99E-6	-5.67E-6	-6.93E-6	-1.38E-7
12	2	28.7696	0	6.991E-6	2.0810	1.0907	0.00416	0.0590	-6.99E-6	-5.67E-6	-6.93E-6	-1.38E-7

Last Evaluation of the Negative of the Gradient

Haz1	Haz2	Stage1	Dose1
-6.93E-6	-1.38E-7	-6.99E-6	-5.67E-6

Last Evaluation of the Negative of the Hessian

	Haz1	Haz2	Stage1	Dose1
Haz1	139133	4906.9	605.6	626.1
Haz2	4906.9	1967.0	46.4218	78.7404
Stage1	605.6	46.4218	5.2554	2.5003
Dose1	626.1	78.7404	2.5003	7.2471

One reason for fitting a proportional hazards model is to evaluate the hazard ratios between various disease groups. You can request customized hazard ratios by using the HAZARDRATIO statement, as follows:

```
proc icphreg data=hiv;
   class Stage / desc;
   model (Left, Right) = Stage / basehaz=pch(intervals=(10));
   hazardratio Stage;
run;
```

Figure 67.8 shows the estimated hazard ratio between the values 1 and 0 of the Stage variable and the corresponding 95% confidence limits.

Figure 67.8 Hazard Ratio Estimate between Stage Values 1 and 0

The ICPHREG Procedure

Hazard Ratios for Stage			
Description	Point Estimate	95% Wald Confidence Limits	
Stage 1 vs 0	5.624	1.734	18.241

The estimate of 5.624 indicates that patients who have Stage 1 disease tend to have a much higher risk of developing AIDS than those who have Stage 0. However, the confidence limits are wide because of the small sample size.

Syntax: ICPHREG Procedure

The following statements are available in the ICPHREG procedure:

PROC ICPHREG < *options* > **;**
BASELINE < **OUT=***SAS-data-set* > < **COVARIATES=***SAS-data-set* >< **TIMELIST=***list* >
< *keyword*=*name* . . . *keyword*=*name* > < / *options* > **;**
BY *variables* **;**
CLASS *variable* < **(***options***)** > . . . < *variable* < **(***options***)** > > < / *global-options* > **;**
FREQ *variable* < / *option* > **;**
HAZARDRATIO < '*label*' > *variable* < / *options* > **;**
ID *variable* **;**
MODEL (*t1***,** *t2***)=** *effects* < / *options* > **;**
MODEL *response* < ∗ *censor* **(***list***)** > **=** *effects* < / *options* > **;**
OUTPUT < **OUT=***SAS-data-set* > < *keyword*=*name* . . . *keyword*=*name* > **;**
TEST < *model-effects* > < / *options* > **;**
STRATA *variable* **;**

The MODEL statement is required, and only one MODEL statement is allowed. If multiple MODEL statements are present, only the last one is used. You can specify main effects and interaction terms in the MODEL statement, as in the GLM procedure. The CLASS statement, if present, must precede the MODEL statement. The BASELINE, HAZARDRATIO, and OUTPUT statements, if present, must follow the MODEL statement. If multiple BASELINE, OUTPUT, and STRATA statements are present, only the last one is used.

The following sections describe the PROC ICPHREG statement and then describe the other statements in alphabetical order.

PROC ICPHREG Statement

PROC ICPHREG < *options* > **;**

The PROC ICPHREG statement invokes the ICPHREG procedure. Table 67.2 summarizes the *options* available in the PROC ICPHREG statement.

Table 67.2 PROC ICPHREG Statement Options

Option	Description
ALPHA=	Specifies the level for confidence limits
DATA=	Names the SAS data set to be analyzed
ITHISTORY	Displays the iteration history, final gradient, and second derivative matrix
NAMELEN=	Specifies the length of effect names
NLOPTIONS	Specifies optimization parameters for fitting the specified model
NOPRINT	Suppresses all displayed output
NOTHREADS	Requests a single-threaded mode for the computation
PLOTS=	Controls the plots that are produced through ODS Graphics
SINGULAR=	Specifies the singularity tolerance
THREADS=	Specifies the number of threads for the computation

You can specify the following *options* in the PROC ICPHREG statement.

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 BASELINE, HAZARDRATIO, and MODEL statements. You can override this default by specifying the ALPHA= option in these statements.

DATA=*SAS-data-set*

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

ITHISTORY

displays the iteration history for computing maximum likelihood estimates, the final evaluation of the gradient, and the final evaluation of the negative of the second derivative matrix (that is, the negative of the Hessian).

NAMELEN=*n*

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

NLOPTIONS(*options***)**

specifies options for the nonlinear optimization methods that are used for fitting the specified model. You can specify the following *options*:

ABSCONV=*r*

ABSTOL=*r*

specifies an absolute function convergence criterion by which minimization stops when $f(\boldsymbol{\psi}^{(k)}) \leq r$, where $\boldsymbol{\psi}$ is the vector of parameters in the optimization and $f(\cdot)$ is the objective function. The default value of *r* is the negative square root of the largest double-precision value, which serves only as a protection against overflows.

ABSFCONV=*r*

ABSFTOL=*r*

specifies an absolute function difference convergence criterion. For all techniques except NMSIMP, termination requires a small change of the function value in successive iterations,

$$|f(\boldsymbol{\psi}^{(k-1)}) - f(\boldsymbol{\psi}^{(k)})| \leq r$$

where $\boldsymbol{\psi}$ denotes the vector of parameters that participate in the optimization and $f(\cdot)$ is the objective function. The same formula is used for the NMSIMP technique, but $\boldsymbol{\psi}^{(k)}$ is defined as the vertex that has the lowest function value, and $\boldsymbol{\psi}^{(k-1)}$ is defined as the vertex that has the highest function value in the simplex. By default, ABSFCONV=0.

ABSGCONV=*r*

ABSGTOL=*r*

specifies an absolute gradient convergence criterion. Termination requires the maximum absolute gradient element to be small,

$$\max_j |g_j(\boldsymbol{\psi}^{(k)})| \leq r$$

where $\boldsymbol{\psi}$ denotes the vector of parameters that participate in the optimization and $g_j(\cdot)$ is the gradient of the objective function with respect to the *j*th parameter. This criterion is not used by the NMSIMP technique. The default value is *r* = 1E–5.

FCONV=*r*

FTOL=*r*

specifies a relative function convergence criterion. For all techniques except NMSIMP, termination requires a small relative change of the function value in successive iterations,

$$\frac{|f(\boldsymbol{\psi}^{(k)}) - f(\boldsymbol{\psi}^{(k-1)})|}{|f(\boldsymbol{\psi}^{(k-1)})|} \leq r$$

where $\boldsymbol{\psi}$ denotes the vector of parameters that participate in the optimization and $f(\cdot)$ is the objective function. The same formula is used for the NMSIMP technique, but $\boldsymbol{\psi}^{(k)}$ is defined as the vertex that has the lowest function value, and $\boldsymbol{\psi}^{(k-1)}$ is defined as the vertex that has the highest function value in the simplex. The default is $r = 10^{-\mathrm{FDIGITS}}$, where FDIGITS is by default $-\log_{10}\{\epsilon\}$ and ϵ is the machine precision.

GCONV=*r*

GTOL=*r*

specifies a relative gradient convergence criterion. For all techniques except CONGRA and NMSIMP, termination requires the normalized predicted function reduction to be small,

$$\frac{\mathbf{g}(\boldsymbol{\psi}^{(k)})'[\mathbf{H}^{(k)}]^{-1}\mathbf{g}(\boldsymbol{\psi}^{(k)})}{|f(\boldsymbol{\psi}^{(k)})|} \leq r$$

where $\boldsymbol{\psi}$ denotes the vector of parameters that participate in the optimization, $f(\cdot)$ is the objective function, and $\mathbf{g}(\cdot)$ is the gradient. For the CONGRA technique (in which a reliable Hessian estimate $\mathbf{H}$ is not available), the following criterion is used:

$$\frac{\| \mathbf{g}(\boldsymbol{\psi}^{(k)}) \|_2^2 \quad \| \mathbf{g}(\boldsymbol{\psi}^{(k)}) \|_2}{\| \mathbf{g}(\boldsymbol{\psi}^{(k)}) - \mathbf{g}(\boldsymbol{\psi}^{(k-1)}) \|_2 \, |f(\boldsymbol{\psi}^{(k)})|} \leq r$$

This criterion is not used by the NMSIMP technique. The default value is *r* = 1E–8.

MAXFUNC=*n*

MAXFU=*n*

specifies the maximum number of function calls in the optimization process. The default values are as follows, depending on the optimization technique (which you can specify in the TECHNIQUE= option):

- TRUREG, NRRIDG, and NEWRAP: 125
- QUANEW and DBLDOG: 500
- CONGRA: 1000
- NMSIMP: 3000

The optimization can terminate only after completing a full iteration. Therefore, the number of function calls that are actually performed can exceed *n*.

MAXITER=*n*

MAXIT=*n*

specifies the maximum number of iterations in the optimization process. The default values are as follows, depending on the optimization technique (which you can specify in the TECHNIQUE= option):

- TRUREG, NRRIDG, and NEWRAP: 50
- QUANEW and DBLDOG: 200
- CONGRA: 400
- NMSIMP: 1000

These default values also apply when *n* is specified as a missing value.

MAXTIME=*r*

specifies an upper limit of *r* seconds of CPU time for the optimization process. The time is checked only at the end of each iteration. Therefore, the actual run time might be longer than *r*. By default, CPU time is not limited.

MINITER=*n*

MINIT=*n*

specifies the minimum number of iterations. If you request more iterations than are actually needed for convergence to a stationary point, the optimization algorithms can behave strangely. For example, the effect of rounding errors can prevent the algorithm from continuing for the required number of iterations. By default, MINITER=0.

TECHNIQUE=*keyword*

TECH=*keyword*

specifies the optimization technique to obtain maximum likelihood estimates. You can choose from the following techniques:

CONGRA	performs a conjugate-gradient optimization.
DBLDOG	performs a version of double-dogleg optimization.
EM	fits the model by using the expectation-maximization (EM) algorithm.

EMICM fits the model by using the EMICM algorithm, which is an extension of the EM algorithm that also uses the iterative convex minorant (ICM) algorithm.

NEWRAP performs a Newton-Raphson optimization that combines a line-search algorithm with ridging.

NMSIMP performs a Nelder-Mead simplex optimization.

NONE performs no optimization.

NRRIDG performs a Newton-Raphson optimization with ridging.

QUANEW performs a dual quasi-Newton optimization.

TRUREG performs a trust-region optimization.

For models other than the semiparametric model, the default is TECHNIQUE=NEWRAP. For the semiparametric model, only the EM and EMICM methods are supported and the default optimization method is EMICM.

For more information about these optimization methods, see the section "Choosing an Optimization Algorithm" on page 512 in Chapter 19, "Shared Concepts and Topics."

For more information about the EM and EMICM methods, see the section "EM Algorithm and Extensions" on page 5173.

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."

NOTHREADS

forces single-threaded execution of the analytic computations. This option overrides the SAS system option THREADS | NOTHREADS. Specifying this option is equivalent to specifying the THREADS=1 option.

PLOTS< **(***global-plot-options***)** **>** **=** *plot-request*

PLOTS< **(***global-plot-options***)** **>** **=** **(***plot-request* < . . . < *plot-request* > >**)**

specifies plots to be created using ODS Graphics. You can request plots of survival functions and cumulative hazard functions. Also, many of the observation statistics in the output data set can be plotted using this option. You are not required to create an output data set in order to produce a plot. When you specify only one *plot-request*, you can omit the parentheses around it.

You can specify the following *global-plot-options*:

CL

displays the pointwise confidence limits for the plot.

OVERLAY **<** **=***overlay-option* **>**

specifies how to overlay the functions that are plotted for the covariate sets. You can specify the following *overlay-options*:

BYGROUP
GROUP
overlays onto the same plot all functions that are plotted for the covariate sets and have the same GROUP= value in the COVARIATES= data set.

INDIVIDUAL
IND
displays a separate plot for each covariate set.

By default, OVERLAY=BYGROUP if the GROUP= option is specified in the BASELINE statement or if the COVARIATES= data set contains the _GROUP_ variable; otherwise, by default, OVERLAY=INDIVIDUAL.

UNPACK
displays multiple plots individually. The default is to display related multiple plots in a panel. The UNPACK option works for INTERVAL, RESDEV, and RESLAG plots only. See the section "OUTPUT Statement" on page 5161 for definitions of the statistics specified with these *plot-requests*.

TIMERANGE=(< *min* > < *,max* >**)**
TIMERANGE=< *min* > < *,max* >
RANGE=(< *min* > < *,max* >**)**
RANGE=< *min* > < *,max* >
specifies the range of values on the time axis to clip the display. The *min* and *max* values are the lower and upper bounds of the range. By default, *min* is 0 and *max* is the largest boundary value.

You can specify the following *plot-requests*:

CUMHAZ
plots the estimated cumulative hazard function for each set of covariates in the data set that you specify in the COVARIATES= option in the BASELINE statement. If the COVARIATES= data set is not specified, the estimated cumulative hazard function is plotted for the reference set of covariates, which consists of reference levels for the CLASS variables and average values for the continuous variables.

HAZARD
plots the estimated hazard function for each set of covariates in the data set that you specify in the COVARIATES= option in the BASELINE statement. If the COVARIATES= data set is not specified, the estimated hazard function is plotted for the reference set of covariates, which consists of reference levels for the CLASS variables and average values for the continuous variables.

INTERVAL
plots the observed interval length as a function of observation number.

NONE
suppresses all the plots in the procedure. Specifying this option is equivalent to disabling ODS Graphics for the entire procedure.

RESDEV< (*options*) **>**
plots deviance residuals. You can specify the following *options*:

INDEX
plots deviance residuals as a function of the observation number.

XBETA
plots deviance residuals as a function of the linear predictor.

If you do not specify an *option*, deviance residuals are plotted as a function of the observation number.

RESLAG< (*options*) >
plots Lagakos residuals. You can specify the following *options*:

INDEX
plots Lagakos residuals as a function of the observation number.

XBETA
plots Lagakos residuals as a function of the linear predictor.

If you do not specify an *option*, deviance residuals are plotted as a function of observation number.

SURVIVAL

S

SURV

SUR
plots the estimated survival function for each set of covariates in the data set that is specified in the COVARIATES= option in the BASELINE statement. If the COVARIATES= data set is not specified, the estimated survival function is plotted for the reference set of covariates, which consists of reference levels for the CLASS variables and average values for the continuous variables.

Each observation in the data set that is specified in the COVARIATES= option in the BASELINE statement provides a set of covariates for which a plot is produced for each *plot-request*. You can use the ROWID= option in the BASELINE statement to specify a variable in the COVARIATES= data set for identifying the functions that are plotted for the covariate sets. If the ROWID= option is not specified, the plots are identified by the covariate values if there is only a single covariate or by the observation numbers of the COVARIATES= data set if the model has two or more covariates. If the COVARIATES= data set is not specified, a reference set of covariates that consists of the reference levels for the CLASS variables and the average values for the continuous variables is used. When plotting more than one function, you can use the OVERLAY= option to group the functions. When you specify only one *plot-request*, you can omit the parentheses around the plot request. Here are some examples:

```
plots=survival
plots=(survival cumhaz)
```

ODS Graphics must be enabled before plots can be requested. For example:

```
ods graphics on;
proc icphreg plots(cl)=survival;
   model (Left, Right)=X1-X5;
   baseline covariates=One;
run;
```

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."

SINGULAR=*number*

EPSILON=*number*

specifies the tolerance for testing the singularity of the $\mathbf{Z}'\mathbf{Z}$ matrix that is formed from the design matrix $\mathbf{Z}$ 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 this 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}.

THREADS=*n*

NTHREADS=*n*

specifies the number of threads for analytic computations and overrides the SAS system option THREADS | NOTHREADS. If you do not specify the THREADS= option or if you specify THREADS=0, the number of threads is determined based on the data size and the number of CPUs on the host on which the analytic computations execute.

BASELINE Statement

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

The BASELINE statement creates a SAS data set (named by the OUT= option) that contains the predicted values at specified times that partition the time axis for every set of covariates in the COVARIATES= data set. If the COVARIATES= data set is not specified, PROC ICPHREG uses a reference set of covariates that consists of the reference levels for the CLASS variables and the average values for the continuous variables.

Table 67.3 summarizes the options that you can specify in the BASELINE statement.

Table 67.3 BASELINE Statement Options

Option	Description
Data Set and Time List Options	
OUT=	Specifies the output BASELINE data set
COVARIATES=	Specifies the SAS data set that contains the explanatory variables
TIMELIST=	Specifies a list of time points for computing the predicted values
Keyword Options for Variables	
CUMHAZ=	Specifies the cumulative hazard function estimate
HAZARD=	Specifies the hazard function estimate

Table 67.3 *continued*

Options	Description
LOGLOGS=	Specifies the log of the negative log of the survival function
LOGSURV=	Specifies the log of the survival function
LOWERCUMHAZ=	Specifies the lower pointwise confidence limit for the cumulative hazard function
LOWER=	Specifies the lower pointwise confidence limit for the survival function
STDCUMHAZ=	Specifies the estimated standard error of the cumulative hazard function
STDERR=	Specifies the standard error of the survival function
STDXBETA=	Specifies the estimated standard error of the linear predictor estimator
SURVIVAL=	Specifies the survival function estimate
UPPERCUMHAZ=	Specifies the upper pointwise confidence limit for the cumulative hazard function
UPPER=	Specifies the upper pointwise confidence limit for the survival function
XBETA=	Specifies the estimate of the linear predictor $\mathbf{z}'\boldsymbol{\beta}$
Other Options	
ALPHA=	Specifies the level of the confidence interval for the survival function
CLTYPE=	Specifies the transformation that is used to compute confidence limits for the survival function
GROUP=	Names a variable whose values identify or group predicted survival or cumulative hazard functions in plots
ROWID=	Names the variable in the COVARIATES= data set for identifying the predicted survival or cumulative hazard functions in plots

You can specify the following options in the BASELINE statement.

OUT=*SAS-data-set*
names the output data set. If you omit the OUT= option, the data set is created and given a default name by using the DATA*n* convention. For more information, see the section "OUT= Output Data Set in the BASELINE Statement" on page 5179.

COVARIATES=*SAS-data-set*
names the SAS data set that contains the sets of explanatory variable values for which the functions of interest are estimated. All variables in the COVARIATES= data set are copied to the OUT= data set. Thus, any variable in the COVARIATES= data set can be used to identify the covariate sets in the OUT= data set.

TIMELIST=*list*
specifies a list of time points at which the predicted values are computed. The following specifications are equivalent:

```
timelist=5,20 to 50 by 10
timelist=5 20 30 40 50
```

If you do not specify this option, predicted values are computed at all the times that partition the time axis.

keyword=*name*

specifies the statistics to be included in the OUT= data set and assigns names to the variables that contain these statistics. Specify a *keyword* for each desired statistic, an equal sign, and the name of the variable for the statistic. You can specify the following *keywords*:

CUMHAZ=*name*

specifies the cumulative hazard function estimate. Specifying CUMHAZ=_ALL_ is equivalent to specifying CUMHAZ=CumHaz, STDCUMHAZ=StdErrCumHaz, LOWERCUMHAZ=LowerCumHaz, and UPPERCUMHAZ=UpperCumHaz.

HAZARD=*name*

specifies the hazard function estimate.

LOGLOGS=*name*

CLOGLOGS=*name*

specifies the log of the negative log of the estimated survival function.

LOGSURV=*name*

specifies the log of the estimated survival function.

LOWER=*name*

L=*name*

LOWERSDF=*name*

specifies the lower pointwise confidence limit for the survival function. The confidence level is determined by the ALPHA= option.

LOWERCUMHAZ=*name*

specifies the lower pointwise confidence limit for the cumulative hazard function. The confidence level is determined by the ALPHA= option.

STDERR=*name*

STDSDF=*name*

specifies the standard error of the survival function estimator.

STDCUMHAZ=*name*

specifies the estimated standard error of the cumulative hazard function estimator.

STDXBETA=*name*

specifies the estimated standard error of the linear predictor estimator.

SURVIVAL=*name*

SDF=*name*

specifies the estimated survival function ($S(t) = [S_0(t)]^{\exp(\mathbf{z}'\boldsymbol{\beta})}$). Specifying SURVIVAL=_ALL_ is equivalent to specifying SURVIVAL=Survival, STDERR=StdErrSurvival, LOWER=LowerSurvival, and UPPER=UpperSurvival.

UPPER=*name*

U=*name*

UPPERSDF=*name*

specifies the upper pointwise confidence limit for the survival function. The confidence level is determined by the ALPHA= option.

UPPERCUMHAZ=*name*

specifies the upper pointwise confidence limit for the cumulative hazard function. The confidence level is determined by the ALPHA= option.

XBETA=*name*

specifies the estimate of the linear predictor $\mathbf{z}'\boldsymbol{\beta}$. If there is an offset, it is added to the predictor.

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

ALPHA=*value*

specifies the level of the confidence interval for the survival function. The *value* must be between 0 and 1. The default is the value of the ALPHA= option in the PROC ICPHREG statement, or 0.05 if that option is not specified.

CLTYPE=*method*

CITYPE=*method*

TYPE=*method*

CLTRANSFORM=*method*

TRANSFORM=*method*

specifies the transformation that is used to compute the confidence limits for $S(t, \mathbf{z})$, which is the survival function for a subject that has a fixed covariate vector $\mathbf{z}$ at event time t. You can specify the following *methods*:

LOG

uses normal theory approximation to compute the confidence limits for $\log(S(t, \mathbf{z}))$. The confidence limits for $S(t, \mathbf{z})$ are obtained by back-transforming the confidence limits for $\log(S(t, \mathbf{z}))$.

LOGLOG

uses normal theory approximation to compute the confidence limits for the $\log(-\log(S(t, \mathbf{z})))$. The confidence limits for $S(t, \mathbf{z})$ are obtained by back-transforming the confidence limits for $\log(-\log(S(t, \mathbf{z})))$.

NORMAL

IDENTITY

LINEAR

PLAIN

DIRECT

uses normal theory approximation to compute the confidence limits for $S(t, \mathbf{z})$.

By default, CLTYPE=LOG.

GROUP=*variable*
names a variable whose values identify or group the predicted curves. The *variable* must be a numeric variable in the COVARIATES= data set. Survival curves for observations that have the same value of the *variable* are overlaid in the same plot.

ROWID=*variable*

ID=*variable*

ROW=*variable*
names a variable in the COVARIATES= data set for identifying plotted survival functions and cumulative hazard functions. This option has no effect if the PLOTS= option in the PROC ICPHREG statement is not specified. Values of this variable are used to label the plotted functions for the corresponding rows in the COVARIATES= data set. You can specify ROWID=_OBS_ to use the observation numbers in the COVARIATES= data set for identification.

BY Statement

BY *variables* **;**

You can specify a BY statement in PROC ICPHREG 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 ICPHREG 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 ICPHREG 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 ICPHREG 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 ICPHREG 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 ICPHREG 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.

FREQ Statement

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

The FREQ statement identifies the *variable* (in the input data set) that contains the frequency of occurrence of each observation. PROC ICPHREG treats each observation as if it appears *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 the frequency value is missing, the observation is not used in the estimation of the regression coefficients.

You can specify the following *option* after a slash (/):

NOTRUNCATE

NOTRUNC

specifies that frequency values are not truncated to integers.

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 particular 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 you can provide an optional label (specified as a quoted string) to identify the output.

Table 67.4 summarizes the *options* that you can specify in the HAZARDRATIO statement.

Table 67.4 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 consider
E	Displays the log-hazard ratio
UNITS=	Specifies the units of change for a continuous variable of interest

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

ALPHA=*number*

specifies the level of confidence intervals for the hazard ratios. The *number* must be between 0 and 1. The default is the value of the ALPHA= option in the PROC ICPHREG 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 you specify a numeric *list* 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 you specify a *list*, at each level of the interacting variable if you specify ALL, or at the reference level of the interacting variable if you specify REF.

If you do not specify the AT option, PROC ICPHREG 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 average of all the sampled 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 average of the sampled 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

ALL

requests all comparisons of only the distinct combinations of pairs.

PAIRWISE

PERM

PERMUTATIONS

requests all possible pairwise comparisons of levels.

REF

REFERENCE

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), and suppose A3 is specified as the reference level. The following table shows 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 $\mathbf{h}$ of linear coefficients such that $\mathbf{h}'\boldsymbol{\beta}$ 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. By default, UNITS=1. This option is ignored in the computation of the hazard ratios for a CLASS variable.

ID Statement

ID *variable* **;**

The ID statement specifies additional variables for identifying observations in the input data. When you specify BASE=UNSPECIFIED, this statement can be used to identify the subjects.

MODEL Statement

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

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

The MODEL statement identifies the variables to be used as the failure time variables and the explanatory effects, including covariates, main effects, interactions, nested effects. For more information, see the section "Specification of Effects" on page 4020 in Chapter 50, "The GLM Procedure."

You can specify two forms of MODEL syntax: the first form allows two time variables, and the second form allows one time variable.

The first form of the MODEL statement enables you to analyze time-to-event data that have interval-censored outcomes. The MODEL syntax specifies two variables, *t1* and *t2*, that contain values of the endpoints of the censoring interval. Only nonnegative values are accepted. If the two values are the same (and not missing), it is assumed that there is no censoring and the actual response value is observed. If the lower value is missing, then the upper value is used as a left-censored value. If the upper value is missing, then the lower value is used as a right-censored value. If both values are present and the lower value is less than the upper value, it is assumed that the values specify a censoring interval. If the lower value is greater than the upper value or both values are missing, then the observation is not used in the analysis.

The following table summarizes the ways of specifying censoring.

Lower Value	Upper Value	Comparison	Interpretation
Not missing	Not missing	Equal	No censoring
Not missing	Not missing	Lower < upper	Censoring interval
Missing	Not missing		Upper used as left-censoring value
Not missing	Missing		Lower used as right-censoring value
Not missing	Not missing	Lower > upper	Observation not used
Missing	Missing		Observation not used

The second form of the MODEL statement enables you to analyze right-censored data or time-to-event data that contain repeated assessments and possibly time-dependent covariates (for more information, see the section "Semiparametric Model and Time-Dependent Covariates" on page 5175). 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) 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.

Table 67.5 summarizes the *options* that you can specify in the MODEL statement.

Table 67.5 MODEL Statement Options

Option	Description
Model Specification Options	
ALPHA=	Specifies the confidence level
BASE=	Specifies the functional form for the baseline function
ENTRY=	Specifies the left-truncation time of the model
HAZSCALE=	Requests parameterization of the hazard function in the original scale or in log scale
NOPOLISH	Suppresses polishing of parameter estimates of the hazard function
OFFSET=	Specifies an offset variable to be added to the linear predictor
PLVARIANCE	Computes the standard error estimates on the basis of the profile likelihood function
Output Options	
CORRB	Displays the estimated correlation matrix
COVB	Displays the estimated covariance matrix

ALPHA=*value*

specifies the level for the confidence intervals for parameters. The *value* must be between 0 and 1. By default, ALPHA=0.05.

CORRB

displays the estimated correlation matrix of the parameter estimates.

COVB
displays the estimated covariance matrix of the parameter estimates.

BASE=*baseline-type*
BASEHAZ=*baseline-type*
B=*baseline-type*
specifies a functional form for the baseline function. You can specify one of the following *baseline-types*:

PCH (< NINTERVAL=*number* **>, < INTERVALS=(***numeric-list***) >)**
PIECEWISE (< NINTERVAL=*number* **>, < INTERVALS=(***numeric-list***) >)**
PIECEWISEEXPONENTIAL (< NINTERVAL=*number* **>, < INTERVALS=(***numeric-list***) >)**
PCBH (< NINTERVAL=*number* **>, < INTERVALS=(***numeric-list***) >)**
partitions the time scale into disjoint intervals and assumes the baseline hazard function is piecewise constant within intervals. The parameters are the piecewise constant values of the baseline hazard functions and are named Haz1, Haz2, . . ., and so on. If HAZARDSCALE=LOGHAZ is specified, the names are LogHaz1, LogHaz2, . . ., and so on.

You can specify one of the following two options to control how to partition the time axis into intervals of constant baseline hazards:

NINTERVAL=*number*
N=*number*
specifies the number of intervals that have a constant hazard rate in each interval. PROC ICPHREG partitions the time axis into the *number* of intervals so that each interval contains an approximately equal number of unique boundary values and imputed middle points.

INTERVALS=(*numeric-list***)**
INTERVAL=(*numeric-list***)**
specifies a list of numbers that partition the time axis into disjoint intervals that have constant hazard rate in each interval. For example, INTERVALS=(100, 150, 200, 250, 300) specifies a model that has a constant hazard in the intervals [0,100), [100,150), [150,200), [200,250), [250,300), and [300,∞).

If you specify neither NINTERVAL= nor INTERVAL=, NINTERVAL=5 by default.

SPLINES (< DF=*number* **>)**
CUBICSPLINES (< DF=*number* **>)**
models the baseline cumulative hazard function by cubic splines (Royston and Parmar 2002). The parameters are the spline coefficients and are named Coef1, Coef2, . . ., and so on.

You can specify the degrees of freedom in the DF=*number* option, where *number* must be an integer. The number of knots equals *number* plus one. The actual positions of the knots are determined from an imputed data set as follows. First, PROC ICPHREG imputes a middle point for each observation in the input data set that is not right-censored. Then, it sorts these imputed times and the input boundary values in increasing order and selects only unique values. PROC ICPHREG places the *terminal knots* at the minimum and maximum of this sequence and chooses the *interval knots* by using the same method it uses to choose the break points for the piecewise constant model. For more information, see the section "Choosing Break Points" on page 5167.

By default, DF=2.

UNSPECIFIED

DISCRETE

models the cumulative hazard function as a discrete function in which jumps are identified according to Turnbull's formulation (1976). The parameters are named Eta1, Eta2, and so on.

The default fitting method for this type of model is EMICM. An alternative is the EM algorithm. For more information about these algorithms, see the section "EM Algorithm and Extensions" on page 5173.

If you do not specify the BASEHAZ= option, the ICPHREG procedure fits a piecewise constant model as if NINTERVAL=5.

ENTRYTIME=*variable*

ENTRY=*variable*

specifies the name of the variable that represents the left-truncation time. For more information, see the section "Left-Truncation of Failure Times" on page 5172.

NOPOLISH

suppresses polishing of parameter estimates of the baseline function. Occasionally, the parameter estimates of the baseline function can reach the default optimization lower bounds. This might indicate that the model is overparameterized. By default, the ICPHREG procedure "polishes" the hazard estimates by fixing these parameters at the lower bound value and refitting the model.

The lower bound values are set to 0 if the baseline parameters are on the original scale (HAZSCALE=HAZARD). The values are set to –10.0 if they are on the log scale (HAZSCALE=LOGHAZ).

This option does not apply to the cubic spline model because its baseline parameters are unbounded.

OFFSET=*variable*

specifies a variable in the input data set to be used as an offset variable. This variable cannot be a CLASS variable, the response variable, or any of the explanatory variables.

HAZSCALE=*hazard-type*

specifies a transformation to be applied to the baseline parameters for fitting the piecewise constant model. You can choose either of the following two options:

LOGHAZ

LOG

LOGHAZARD

uses the log transformed baseline parameters.

HAZARD

HAZ

does not transform the baseline parameters. A lower bound of 0 is used for fitting the models.

This option does not apply to the cubic spline model and the semiparametric model.

PLVARIANCE

computes the standard error estimates on the basis of the profile likelihood function, as opposed to the default Louis's method (Louis 1982). For more information, see the section "Variance Estimation" on page 5177. This option applies only to the semiparametric model.

OUTPUT Statement

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

The OUTPUT statement creates a new SAS data set that contains all the variables in the input data set and, optionally, contains the estimated linear predictors (XBETA) and their standard error estimates, residuals, and diagnostic statistics.

The estimated linear predictor and its standard error estimate are computed for all observations in which the explanatory variables are all nonmissing, even if the response is missing. By adding to the input data set observations for which both lower and upper values are missing, you can compute these statistics for new observations for settings of the explanatory variables not present in the data without affecting the model fit.

You can specify the following options:

OUT=*SAS-data-set*
specifies the output data set. If you omit the OUT=option, PROC ICPHREG creates the output data set and gives it a default name that uses the DATA*n* convention.

keyword=name
specifies the statistics to be included in the output data set and names the new variables that contain the statistics. Specify a keyword for each statistic that you want (see the following list of keywords), an equal sign, and the name of the new variable or variables to contain the statistic. You can list only one variable after the equal sign for all the statistics.

Although you can use the OUTPUT statement without specifying a *keyword=name*, if you do so, the output data set then contains only the original variables. Formulas for the statistics are given in the section "Residuals and Diagnostic Statistics" on page 5173.

You can specify the following *keywords*:

INTERVAL represents the interval length for the input observation. The value is missing if the observation is right-censored or missing.

RESDEV represents the deviance residual for identifying poorly fitted observations.

RESLAG represents the Lagakos residual for identifying poorly fitted observations.

STDXBETA represents the standard error estimate of XBETA (see the XBETA keyword).

XBETA represents the estimate of the linear predictor $\mathbf{x}_i'\boldsymbol{\beta}$ for observation i. If there is an offset, it is included in $\mathbf{x}_i'\boldsymbol{\beta}$.

The OUTPUT statement is not available when you specify the STRATA statement or the ENTRY= option in the MODEL statement.

STRATA Statement

STRATA *variable* < */ option* > **;**

The proportional hazards assumption might not be realistic for all data. However, it might still be reasonable to perform a stratified analysis. The STRATA statement names one variable that determines the stratification.

Strata are formed according to the nonmissing values of the STRATA variables unless you specify the MISSING option in the CLASS statement. In the STRATA statement, *variable* is a variable whose values are used to determine the strata levels. Multiple variables can appear in the STRATA statement.

You can specify the following *option*:

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

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 of Type I, II, and III estimable functions, see Chapter 15, "The Four Types of Estimable Functions."

Table 67.6 summarizes the *options* that you can specify in the TEST statement.

Table 67.6 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

For information about the syntax of the TEST statement, see the section "TEST Statement" on page 521 in Chapter 19, "Shared Concepts and Topics."

Details: ICPHREG Procedure

Model and Likelihood

Suppose that the observations to be analyzed consist of interval-censored outcomes $\{[L_i, R_i]; \mathbf{Z}_i\}$, $i = 1, ..., n$, where n is the number of subjects. $\mathbf{Z}_i$ denotes a p-dimensional vector of covariates for the ith subject. This notation allows for exact event times, right-censored data and left-censored data as special cases. When $L_i = R_i$, the observation is an exact time; when $R_i = \infty$, the observation is right-censored; when $L_i = 0$, the observation is left-censored.

Let $S(t;\mathbf{Z}_i)$ denote the survival function for a subject whose covariate is $\mathbf{Z}_i$. Assuming that t is continuous, denote $f(t;\mathbf{Z}_i)$ as the density function for the subject. The hazard function for the subject, $\lambda(t;\mathbf{Z}_i)$, is defined as the instantaneous failure rate at time t. Mathematically, the hazard function is determined as a ratio between the density function and the survival function:

$$\lambda(t;\mathbf{Z}_i) = f(t;\mathbf{Z}_i)/S(t;\mathbf{Z}_i)$$

A quantity that is closely related to the survival function is the cumulative hazard function, defined as

$$\Lambda(t;\mathbf{Z}_i) = \int_0^t \lambda(u;\mathbf{Z}_i)\,du$$

In turn, the cumulative hazard function determines the survival function:

$$S(t;\mathbf{Z}_i) = \exp(-\Lambda(t;\mathbf{Z}_i))$$

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

$$\begin{aligned}\log(\mathrm{L}) &= \sum \log\left[f(L_i;\mathbf{Z}_i)\right] + \sum \log\left[S(L_i;\mathbf{Z}_i)\right] \\ &+ \sum \log\left[1 - S(R_i;\mathbf{Z}_i)\right] + \sum \log\left[S(L_i;\mathbf{Z}_i) - S(R_i;\mathbf{Z}_i)\right]\end{aligned}$$

where the first sum is the total of the uncensored observations, the second sum is the total of the right-censored observations, the third sum is the total of the left-censored observations, and the last sum is the total of the interval-censored observations.

For the ith subject, the proportional hazards model (Cox 1972) assumes that

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

where $\boldsymbol{\beta}$ is a p-dimensional vector of coefficients for the covariate vector $\mathbf{Z}_i$ and $\lambda_0(t)$ is the baseline hazard function, which is the hazard rate when all the coefficients for the covariates are equal to 0.

Under the proportional hazards model, the cumulative hazard function for the ith subject is

$$\Lambda(t;\mathbf{Z}_i) = \int_0^t \lambda(u;\mathbf{Z}_i)\,du = \int_0^t \lambda_0(u)\,du\exp(\mathbf{Z}_i'\boldsymbol{\beta}) = \Lambda_0(t)\exp(\mathbf{Z}_i'\boldsymbol{\beta})$$

The survival function for the ith subject is

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

where $S_0(t)$ denotes the baseline survival function and $S_0(t) = \exp[-\Lambda_0(t)]$.

The density function for the subject is obtained by differentiating the survival function:

$$f(t;\mathbf{Z}_i) = -\frac{dS(t;\mathbf{Z}_i)}{dt} = \lambda(t;\mathbf{Z}_i)S(t;\mathbf{Z}_i) = \lambda_0(t)\exp(\mathbf{Z}_i'\boldsymbol{\beta})S_0(t)^{\exp(\mathbf{Z}_i'\boldsymbol{\beta})}$$

Given these quantities, the likelihood function under the proportional hazards model can be expressed as

$$\begin{aligned}\log(\mathrm{L}) &= \sum \log\left[\lambda_0(L_i)\exp(\mathbf{Z}_i'\boldsymbol{\beta})S_0(L_i)^{\exp(\mathbf{Z}_i'\boldsymbol{\beta})}\right] + \sum \log\left[S_0(L_i)^{\exp(\mathbf{Z}_i'\boldsymbol{\beta})}\right] \\ &+ \sum \log\left[1 - S_0(R_i)^{\exp(\mathbf{Z}_i'\boldsymbol{\beta})}\right] + \sum \log\left[S_0(L_i)^{\exp(\mathbf{Z}_i'\boldsymbol{\beta})} - S_0(R_i)^{\exp(\mathbf{Z}_i'\boldsymbol{\beta})}\right]\end{aligned}$$

where the first sum is the total of the uncensored observations, the second sum is the total of the right-censored observations, the third sum is the total of the left-censored observations, and the last sum is the total of the interval-censored observations.

This likelihood function is often referred as the *full likelihood* as compared to the partial likelihood (Cox 1972) because it involves parameters for the baseline hazard function in addition to the regression coefficients $\boldsymbol{\beta}$. The full likelihood is often used for analyzing interval-censored data because constructing a likelihood function that contains only the regression coefficients as conveniently as the Cox partial likelihood does for right-censored data is not straightforward (Finkelstein 1986).

Baseline Parameterization

Because any one of the baseline hazard, cumulative hazard, and survival functions determines the others, it is sufficient to parameterize one of them. For the baseline function, PROC ICPHREG supports the parameterizations that are described in the following subsections.

Piecewise Constant Model

As its name suggests, the piecewise constant hazard rate model parameterizes the baseline hazard function as a union of several disjoint intervals, within each of which the hazard rate is constant:

$$\lambda_0(t) = r_j \text{ if } a_{j-1} \le t < a_j, j = 1, \ldots, J$$

It follows that the baseline cumulative hazard function is

$$\Lambda_0(t) = \sum_{j=1}^{J} r_j \Delta_j(t)$$

where

$$\Delta_j(t) = \begin{cases} 0 & t < a_{j-1} \\ t - a_{j-1} & a_{j-1} \le t < a_j \\ a_j - a_{j-1} & t \ge a_j \end{cases}$$

To produce a meaningful hazard function, the r_j need to be bounded below by 0. Such a constraint can be removed by transforming the parameters to a natural log scale:

$$\alpha_j = \log(r_j), \quad j = 1, \ldots, J$$

PROC ICPHREG uses either the original or the transformed scale to fit piecewise constant models. You can change the scale by using the HAZSCALE= option. By default, the original scale is used.

Cubic Splines Model

For the proportional hazards model, Royston and Parmar (2002) propose modeling the log of the baseline cumulative hazard function in terms of natural cubic splines,

$$\log[\Lambda_0(t)] = \gamma_0 + \gamma_1 x + \gamma_2 v_1(x) + \cdots + \gamma_{m+1} v_J(x)$$

where $x = \log(t)$ represents the time on a log scale. The v_j are the basis functions, which are computed as

$$v_j(x) = (x - k_j)_+^3 - e_j(x - k_{\min})_+^3 - (1 - e_j)(x - k_{\max})_+^3$$

where

$$\begin{aligned} e_j &= \frac{k_{\max} - k_j}{k_{\max} - k_{\min}} \\ (x-a)_+ &= \max(0, x - a) \end{aligned}$$

Here, $k_{\min}$ and $k_{\max}$ are two terminal knots, and $k_1 < \cdots < k_J$ are m interval knots that are placed between $k_{\min}$ and $k_{\max}$. The degrees of freedom equals $m + 1$. When $m = 0$, the log of the baseline hazard becomes $\gamma_0 + \gamma_1 x$, which corresponds to a common form of the Weibull model. When $\gamma_1 = 1$, the Weibull model further reduces to the exponential model.

Semiparametric Model

Assuming that the underlying baseline hazard is a discrete function, Finkelstein (1986) proposed parameterizing the baseline survival function as a discrete function in which jumps occur only at the discrete times. You typically apply the algorithm proposed by Turnbull (1976) to reduce the number of parameters to estimate. First, you rank observed interval boundaries $\{L_i, R_i, i = 1, \ldots, n\}$ with labels of L and R attached. Then you identify the set of intervals for which L is the left boundary and R is the right boundary, and the reduced set of discrete times $s_0 = 0 < s_1 < \cdots < s_{m+1} = \infty$ is simply the ranked right boundaries of these intervals. The cumulative hazard function is assumed to be

$$\Lambda_0(t) = \sum_{j:s_j<t} \gamma_j, j = 1, \ldots, J$$

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 identified 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 joining two or more continuous variables with asterisks: 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 the preceding 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, 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 two or fewer variables: A B C A*B A*C B*C.

Computational Details

Design Matrix

The linear predictor part of a proportional hazards model is

$$\boldsymbol{\mu} = \mathbf{Z}'\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ is a vector of unknown regression coefficients and $\mathbf{Z}$ is a known design matrix. 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 use the PARAM=GLM option in the CLASS statement to specify an overparameterized model, some columns of $\mathbf{Z}$ can be linearly dependent on other columns. For example, when you specify a model that consists of 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 $\mathbf{Z}$. The columns of $\mathbf{Z}'\mathbf{Z}$ 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 and its standard error are set to 0 to indicate that it is not estimated. The test for linear dependence is controlled by the SINGULAR= option in the MODEL statement. You can use the ORDER= option in the CLASS statement to specify the order in which the levels of a classification variable are checked for dependencies. For full-rank parameterizations, the columns of the $\mathbf{Z}$ matrix are designed to be linearly independent.

Initial Values

The initial values of the regression coefficients $\boldsymbol{\beta}$ are all set to 0.

For the piecewise constant model, the initial values of the hazard parameters are set equal to the exponential rate that is estimated from an imputed data set. The data set is obtained by imputing a middle point for the

interval-censored and left-censored observations while retaining the right-censored and exact observations. For the cubic spline model, the first spline coefficient, γ_0, is set to be the log of the exponential rate estimated with the previous imputed data, and the second spline coefficient, γ_1, is set to 1. The remaining spline coefficients, if there are any, are set to 0.

Maximum Likelihood Estimation

By default, the ICPHREG procedure uses a Newton-Raphson algorithm to maximize the log-likelihood function with respect to the parameters.

Denote the set of parameters that need to be estimated as $\boldsymbol{\omega} = \{\omega_j\}$, which consists of the parameters that determine baseline hazard function $\Lambda_0(t)$ and the regression coefficients $\boldsymbol{\beta}$. On the rth iteration, the algorithm updates the parameter vector $\boldsymbol{\omega}_r$ with

$$\boldsymbol{\omega}_{r+1} = \boldsymbol{\omega}_r - \mathbf{H}^{-1}\mathbf{g}$$

where $\mathbf{H}$ is the Hessian (second derivative) matrix, and $\mathbf{g}$ is the gradient (first derivative) vector of the log-likelihood function, both evaluated at the current value of the parameter vector. That is,

$$\mathbf{g} = [g_j] = \left[\frac{\partial l}{\partial \omega_j}\right]$$

and

$$\mathbf{H} = [h_{ij}] = \left[\frac{\partial^2 l}{\partial \omega_i \partial \omega_j}\right]$$

The ICPHREG procedure also supports other optimization methods, such as quasi-Newton and Newton-Raphson with ridging. These methods are described in the section "Choosing an Optimization Algorithm" on page 512 in Chapter 19, "Shared Concepts and Topics."

Covariance and Correlation Matrix

The estimated covariance matrix of the parameter estimator is

$$\boldsymbol{\Sigma} = -\mathbf{H}^{-1}$$

where $\mathbf{H}$ is the Hessian matrix that is evaluated using the parameter estimates on the last iteration. If some parameters in the baseline function are held fixed, they are not incorporated in $\mathbf{H}$. Rows and columns that correspond to aliased parameters are not included in $\boldsymbol{\Sigma}$.

The correlation matrix is the normalized covariance matrix. That is, if σ_{ij} is an element of $\boldsymbol{\Sigma}$, then the corresponding element of the correlation matrix is $\sigma_{ij}/\sigma_i\sigma_j$, where $\sigma_i = \sqrt{\sigma_{ii}}$.

Choosing Break Points

There are no obvious ways to choose break points for parameterizing the baseline function in terms of a piecewise constant function or a cubic spline curve. For right-censored data, PROC PHREG chooses a set of points such that the resulting time intervals contain approximately equal numbers of event times. This is difficult for interval-censored data because event times are not fully observed. Friedman (1982) recommends choosing the points so that the expected number of events is comparable among the time intervals. For an

interval-censored spline model, Cai and Betensky (2003) propose an ad hoc approach that uses the quantile values of the unique time points among $\{L_i, R_i, (L_i + R_i)/2, i = 1, ..., n\}$ for choosing the knot values.

Ibrahim, Chen, and Sinha (2001) propose the equally spaced quantile partition (ESQP) method for selecting break points in the right-censored data to fit the piecewise constant model. Suppose there are Q break points to be determined. The ICPHREG procedure modifies this method to handle interval-censored data. First, it imputes a middle point for each observation that is not right-censored. Then, it merges these values with the observed boundary values in the input data set, except for the right-censored observations. Next, it sorts these values in increasing order.

Suppose the unique values of the sorted sequence are $u_1 < u_2 < \cdots < u_M$. First, PROC ICPHREG computes the targeted quantile for each break point as $q_j = j/(Q + 1)(j = 1, \cdots, Q)$. Then, it chooses the point u_{m+1}, where m equals the integer part of the product $q_j M$. If $q_j M$ is already an integer, then the chosen break point is set to be $(u_m + u_{m+1})/2$. When there are no ties in the sorted sequence for right-censored data, this method is identical to the original ESQP method.

Fit Statistics

Suppose that the model contains q estimated parameters and that n observations are used in model fitting. The fit criteria displayed by the ICPHREG procedure are calculated as follows:

- –2 log likelihood:

  $$-2\log(\mathrm{L})$$

  where L is the maximized likelihood for the model.

- Akaike's information criterion:

  $$\mathrm{AIC} = -2\log(\mathrm{L}) + 2q$$

- corrected Akaike's information criterion:

  $$\mathrm{AICC} = \mathrm{AIC} + \frac{2q(q+1)}{n-q-1}$$

- Bayesian information criterion:

  $$\mathrm{BIC} = -2\log(\mathrm{L}) + q\log(n)$$

For more information about AIC and BIC, see Akaike (1981, 1979). For a discussion of using AIC, AICC, and BIC in statistical modeling, see Simonoff (2003).

Predicted Values

Given a new vector of covariates $\mathbf{Z}_{\text{new}}$, the linear predictor is computed as $\hat{\mu}_{\mathbf{Z}_{\text{new}}} = \mathbf{Z}'_{\text{new}}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$. The variance of $\hat{\mu}_{\mathbf{Z}_{\text{new}}}$ is estimated by

$$\hat{\sigma}^2_{\mathbf{Z}_{\text{new}}} = \mathbf{Z}'_{\text{new}}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}\mathbf{Z}_{\text{new}}$$

where $\mathbf{\Sigma}_{\hat{\boldsymbol{\beta}}}$ denotes the estimated covariance matrix for $\hat{\boldsymbol{\beta}}$.

Suppose the estimated baseline hazard is $\hat{\Lambda}_0(t)$. Given $\mathbf{Z}_{\text{new}}$, the cumulative hazard function can be predicted by

$$\hat{\Lambda}(t;\mathbf{Z}_{\text{new}}) = \hat{\Lambda}_0(t)e^{\mathbf{Z}'_{\text{new}}\hat{\boldsymbol{\beta}}}$$

Denote the vector of parameters that is used for obtaining $\hat{\Lambda}_0(t)$ as $\boldsymbol{\tau}$. It is apparent that $\boldsymbol{\tau} \cap \boldsymbol{\beta} = \emptyset$. The vector of parameters that need to be estimated can be represented as $\boldsymbol{\omega} = (\boldsymbol{\beta}, \boldsymbol{\tau})$.

The variance of $\hat{\Lambda}(t;\mathbf{Z}_{\text{new}})$ can be estimated by applying the delta method:

$$\hat{\sigma}^2(\hat{\Lambda}(t;\mathbf{Z}_{\text{new}})) = P(t,\hat{\boldsymbol{\omega}})'\mathbf{\Sigma}\,P(t,\hat{\boldsymbol{\omega}})$$

where

$$P(t,\boldsymbol{\omega}) = \frac{\partial\Lambda(t;\mathbf{Z}_{\text{new}})}{\partial\boldsymbol{\omega}}$$

and $\mathbf{\Sigma}$ denotes the estimated covariance matrix for $\hat{\boldsymbol{\omega}}$.

Given $\mathbf{Z}_{\text{new}}$, the predicted survival function is estimated by

$$\hat{S}(t;\mathbf{Z}_{\text{new}}) = \exp(\hat{\Lambda}(t;\mathbf{Z}_{\text{new}}))$$

The standard error of $\hat{S}(t;\mathbf{Z}_{\text{new}})$ can be conveniently estimated by an application of the delta method:

$$\hat{\sigma}(\hat{S}(t;\mathbf{Z}_{\text{new}})) = \hat{S}(t;\mathbf{Z}_{\text{new}})\hat{\sigma}(\hat{\Lambda}(t;\mathbf{Z}_{\text{new}}))$$

By default, a natural log transformation is applied to obtain the pointwise confidence limits for $S(t;\mathbf{Z}_{\text{new}})$ and $\Lambda(t;\mathbf{Z}_{\text{new}})$. You can use the CLTYPE= option to specify a different transformation for $S(t;\mathbf{Z}_{\text{new}})$.

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

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

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

The hazard ratio ψ is defined as the ratio of the hazard for those who have the risk factor ($X = 1$) to the hazard for those who do not have the risk factor ($X = 0$). The log of the hazard ratio is

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

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 that is associated with X represents the change in the log-hazard 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:

	Design Variables		
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

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

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

	Design Variables		
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

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

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

	Design Variables			
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

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

Consider Cell as the only risk factor. The computation of the hazard ratio of Adeno versus Squamous for various parameterization schemes is shown in Table 67.7.

Table 67.7 Hazard Ratio of Adeno to Squamous

	Parameter Estimates				
PARAM=	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Hazard Ratio Estimates**
EFFECT	0.5772	–0.2115	0.2454		$\exp(2 \times 0.5772 - 0.2115 + 0.2454) = 3.281$
REF	1.8830	0.3996	0.8565		$\exp(1.8830) = 3.281$
GLM	1.8830	0.3996	0.8565	0.0000	$\exp(1.8830) = 3.281$

The fact that the log-hazard ratio ($\log(\psi)$) is a linear function of the parameters enables the HAZARDRATIO statement to compute the hazard ratio of the main effect even in the presence of interactions and nest effects.

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 shown in a separate table. Let (V_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 $\exp(cV_j)$ and $\exp(cU_j)$, respectively for $c > 0$, or $\exp(cU_j)$ and $\exp(cV_j)$, respectively for $c < 0$.

Let $\mathbf{e}_j$ be the jth unit vector—that is, the jth entry of the vector is 1 and all other entries are 0. The hazard ratio for the explanatory variable with regression coefficient $\beta_j = \mathbf{e}_j'\boldsymbol{\beta}$ is defined as $\exp(\beta_j)$. In general, a log-hazard ratio can be written as $\mathbf{h}'\boldsymbol{\beta}$ (a linear combination of the regression coefficients), and the hazard ratio $\exp(\mathbf{h}'\boldsymbol{\beta})$ is obtained by replacing $\mathbf{e}_j$ with $\mathbf{h}$.

Point Estimate

The hazard ratio $\exp(\mathbf{e}_j'\boldsymbol{\beta})$ is estimated by $\exp(\mathbf{e}_j'\hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate of the regression coefficients $\boldsymbol{\beta}$.

Wald's Confidence Limits

The $100(1-\alpha)\%$ confidence limits for the hazard ratio are calculated as

$$\exp\left(\mathbf{e}_j'\hat{\boldsymbol{\beta}} \pm z_{\alpha/2}\sqrt{\mathbf{e}_j'\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}\mathbf{e}_j}\right)$$

where $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$ is estimated covariance matrix and $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile point of the standard normal distribution.

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). Any individuals whose left-truncation time exceeds the event time cannot be observed. 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, T3) represent the left-truncation time and the observed interval, respectively. To account for left-truncation, you specify the following statements:

```
proc icphreg;
   model (T2,T3)=X1-X10/entry=T1;
   title 'The ENTRY= option is Specified';
run;
```

For the specified truncation time to be meaningful, its value must be smaller than the values of the interval boundaries. Otherwise, the observation is removed from the analysis.

Residuals and Diagnostic Statistics

The ICPHREG procedure computes two types of residuals. Residuals are available for the piecewise constant hazard model and the cubic splines model.

For the *i*th subject whose observed interval is (L_i, R_i), the Lagakos residual is defined in Farrington (2000) as

$$r_i^L = \frac{\hat{S}_i(L_i)\log[\hat{S}_i(L_i)] - \hat{S}_i(R_i)\log[\hat{S}_i(R_i)]}{\hat{S}_i(L_i) - \hat{S}_i(R_i)}$$

You can request Lagakos residuals in an output data set by using the keyword RESLAG in the OUTPUT statement.

The deviance residual is defined as

$$r_i^D = \text{sign}(r_i^L)\left\{2\log\left[\frac{\hat{S}_0(L_i)^{\eta_i} - \hat{S}_0(R_i)^{\eta_i}}{\hat{S}_0(L_i)^{\exp(\mathbf{Z}_i'\hat{\boldsymbol{\beta}})} - \hat{S}_0(R_i)^{\exp(\mathbf{Z}_i'\hat{\boldsymbol{\beta}})}}\right]^{1/2}\right\}$$

where

$$\eta_i = \frac{\log[\hat{\Lambda}_0(R_i)] - \log[\hat{\Lambda}_0(L_i)]}{\hat{\Lambda}_0(R_i) - \hat{\Lambda}_0(L_i)}$$

and where $\eta_i = 0$ if $R_i = \infty$ and $\eta_i = \infty$ if $L_i = 0$.

You can request deviance residuals in an output data set by using the keyword RESDEV in the OUTPUT statement.

The ICPHREG procedure computes the length of input intervals. For the *i*th subject whose observed interval is (L_i, R_i), the length is defined as

$$\text{Len}_i = \begin{cases} R_i - L_i & \text{if } R_i \geq L_i > 0 \\ R_i & \text{if } R_i > 0 \text{ and } L_i = 0 \text{ or } L_i \text{ is missing} \\ -1 & \text{otherwise} \end{cases}$$

You can request interval lengths in an output data set by using the keyword INTERVAL in the OUTPUT statement.

EM Algorithm and Extensions

The expectation-maximization (EM) algorithm, as described in Wang et al. (2016) and Zeng, Mao, and Lin (2016), can be used to fit certain types of proportional hazards models to interval-censored data.

Suppose that the observations to be analyzed consist of interval-censored outcomes $\mathcal{D} = \{[L_i, R_i]; \mathbf{Z}_i\}$, $i = 1, \ldots, n$, where n is the number of subjects. $\mathbf{Z}_i$ denotes a p-dimensional vector of covariates for the *i*th subject.

Assuming that there is no exact observation ($L_i = R_i$), the full likelihood function is

$$\begin{aligned} \mathrm{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} [S(L_i;\mathbf{Z}_i) - S(R_i;\mathbf{Z}_i)] \\ &= \prod_{i=1}^{n} [1 - S(R_i;\mathbf{Z}_i)]^{\Delta_{i1}} [S(L_i;\mathbf{Z}_i) - S(R_i;\mathbf{Z}_i)]^{\Delta_{i2}} [S(L_i;\mathbf{Z}_i)]^{\Delta_{i3}} \end{aligned}$$

where Δ_{i1} indicates whether the ith subject is left-censored ($L_i = 0$), Δ_{i2} indicates whether the ith subject is interval-censored ($0 < L_i < R_i < \infty$), and Δ_{i3} indicates whether the ith subject is right-censored ($R_i = \infty$).

Assume that the baseline hazard function is of the following form,

$$\Lambda_0(t) = \sum_{k=1}^{K} \gamma_k b_k(t)$$

where $b_k(t)$ are known functions that are nondecreasing and nonnegative, and γ_k are nonnegative baseline parameters.

Let $\{W_{ik} : i = 1,\ldots,n; k = 1,\ldots,K\}$ be a set of latent variables that follow Poisson distributions with means $\gamma_k b_k(L_i)\exp(\mathbf{Z}_i'\boldsymbol{\beta})$. Let $\{U_{ik} : i = 1,\ldots,n; k = 1,\ldots,K\}$ be a set of latent variables that follow Poisson distributions with means $\gamma_k(b_k(R_i) - b_k(L_i))\exp(\mathbf{Z}_i'\boldsymbol{\beta})$. Define $W_i = \sum_{k=1}^{K} W_{ik}$ and $U_i = \sum_{k=1}^{K} U_{ik}$.

The full likelihood can be rewritten as

$$\mathrm{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} P(W_i > 0)^{\Delta_{i1}} P(W_i = 0, U_i > 0)^{\Delta_{i2}} P(W_i = 0, U_i = 0)^{\Delta_{i3}}$$

The complete-data likelihood is

$$\mathrm{L}_c(\boldsymbol{\theta}) = \prod_{i=1}^{n}\prod_{k=1}^{K} f_{W_{ik}}(W_{ik}) f_{U_{ik}}(U_{ik})^{\Delta_{i2}+\Delta_{i3}}$$

where $f_V(\cdot)$ denotes the Poisson probability mass function for the variable V. It is straightforward to verify that the integration of $\mathrm{L}_c(\boldsymbol{\theta})$ with respect to latent variables leads to the full likelihood $\mathrm{L}(\boldsymbol{\theta})$.

The EM algorithm proceeds as follows. Let the current parameter estimates be $\boldsymbol{\theta}^{(d)} = (\boldsymbol{\beta}^{(d)}, \boldsymbol{\gamma}^{(d)})'$. Define

$$\begin{aligned} w_{ik} &= E(W_{ij}|\mathcal{D}, \boldsymbol{\theta}^{(d)}) \\ u_{ik} &= E(U_{ij}|\mathcal{D}, \boldsymbol{\theta}^{(d)}) \\ w_i &= E(W_i|\mathcal{D}, \boldsymbol{\theta}^{(d)}) \\ u_i &= E(U_i|\mathcal{D}, \boldsymbol{\theta}^{(d)}) \end{aligned}$$

The expected complete-data log likelihood $\mathrm{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)}) = E[\log(\mathrm{L}_c(\boldsymbol{\theta}))|\mathcal{D}, \boldsymbol{\theta}^{(d)}]$ is computed as

$$\begin{aligned} \mathrm{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} \Big\{ [w_{ik} + (\Delta_{i2} + \Delta_{i3})u_{ik}] \left[\log(\gamma_k) + \mathbf{Z}_i'\boldsymbol{\beta}\right] \\ &\quad -\gamma_k \exp(\mathbf{Z}_i'\boldsymbol{\beta}) [(\Delta_{i2} + \Delta_{i1})b_k(R_i) + \Delta_{i3} b_k(L_i)] \Big\} + B(\boldsymbol{\theta}^{(d)}) \end{aligned}$$

where $B(\boldsymbol{\theta}^{(d)})$ is a constant.

The quantities w_{ik} and u_{ik} are computed as follows,

$$w_{ik} = \frac{\gamma_k^{(d)} b_k(R_i) w_i}{\Lambda_0^{(d)}(R_i)}$$

$$u_{ik} = \frac{\gamma_k^{(d)} [b_k(R_i) - b_k(L_i)] u_i}{\Lambda_0^{(d)}(R_i) - \Lambda_0^{(d)}(L_i)}$$

where

$$w_i = \frac{\Lambda_0^{(d)}(R_i) \exp(\mathbf{Z}_i' \boldsymbol{\beta}^{(d)}) \Delta_{i1}}{1 - \exp\left[\Lambda_0^{(d)}(R_i) \exp(\mathbf{Z}_i' \boldsymbol{\beta}^{(d)})\right]}$$

$$u_i = \frac{[\Lambda_0^{(d)}(R_i) - \Lambda_0^{(d)}(L_i)] \exp(\mathbf{Z}_i' \boldsymbol{\beta}^{(d)}) \Delta_{i2}}{1 - \exp\left\{\left[\Lambda_0^{(d)}(R_i) - \Lambda_0^{(d)}(L_i)\right] \exp(\mathbf{Z}_i' \boldsymbol{\beta}^{(d)})\right\}}$$

Solve $\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})/\partial \gamma_k = 0$ for $k = 1, \ldots, K$. It follows that the $\gamma_k^{(d)}$ can be updated as

$$\gamma_k^{(d+1)} = \frac{\sum_{i=1}^n (z_{ik} + \Delta_{i2} w_{ik})}{\sum_{i=1}^n [(\Delta_{i1} + \Delta_{i2}) b_k(R_i) + \Delta_{i3} b_k(L_i)] \exp(\mathbf{Z}_i' \boldsymbol{\beta})}$$

The partial derivative of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})$ with respect to $\boldsymbol{\beta}$ is

$$\frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \left\{(w_i + (\Delta_{i2} + \Delta_{i3}) u_i) - \left[(\Delta_{i1} + \Delta_{i2}) \Lambda_0^{(d)}(R_i) + \Delta_{i3} \Lambda_0^{(d)}(L_i)\right] \exp(\mathbf{Z}_i' \boldsymbol{\beta})\right\} \mathbf{Z}_i$$

After plugging in $\{\gamma_k^{(d+1)}, k = 1, \ldots, K\}$, you can update the parameters $\boldsymbol{\beta}$ by using the one-step Newton-Raphson method (Zeng, Mao, and Lin 2016).

The EM algorithm alternates between updating $\boldsymbol{\gamma}$ and updating $\boldsymbol{\beta}$ until convergence.

You can use the EM algorithm to fit the semiparametric model and the piecewise constant hazard model in PROC ICPHREG. The option is NLOPTIONS(TECH=EM) in the PROC ICPHREG statement.

Semiparametric Model and Time-Dependent Covariates

A typical way that interval-censored data are generated is through a process of repeated assessments. Suppose that $V_1 < V_2 < \cdots < V_M$ are a random sequence of assessment times. Denote $\tilde{V} = (V_0 = 0, V_1, V_2, \ldots, V_M, V_{M+1} = \infty)$ and $\tilde{D} = (D_0 = 0, D_1, D_2, \ldots, D_M)$, where $D_m = I(V_{m-1} < T < V_m), m = 1, \ldots, M$.

For the ith subject, $i = 1, \ldots, n$, let K_i, T_i, $\tilde{V}_i = (V_{i0}, V_{i1}, \ldots, V_{i(K_i+1)})$, $\tilde{D}_i = (D_{i0}, D_{i1}, \ldots, D_{iK_i})$, and $\mathbf{Z}_i(\cdot)$ be the number of assessments, event time, assessment time vector, the indicator vector, and time-dependent covariate process, respectively. Suppose that T_i is interval-censored between two assessment times, L_i and R_i, where $L_i = \max_{V_j}\{V_j < T_i, j = 0, \ldots, K_i\}$ and $R_i = \min_{V_j}\{V_j < T_i, j = 1, \ldots, K_i + 1\}$. Let $s_1 < \cdots < s_J$ be the sorted right boundaries of the Turnbull intervals for $\{(L_i, R_i] : i = 1, \ldots, n\}$.

Suppose that the time-dependent covariates process $\mathbf{Z}_i(\cdot)$ change value only at assessment times. Let $(\mathbf{Z}_{i1}, \mathbf{Z}_{i2}, \ldots, \mathbf{Z}_{i(J+1)})$ be the observed covariate vectors at times $(0, s_1, \ldots, s_J)$.

Under the semiparametric model, the baseline cumulative hazard function is

$$\Lambda_0(t) = \sum_{j:s_j<t} \gamma_j, \; j = 1, \ldots, J$$

For the *i*th subject, the cumulative hazard function is computed as

$$\Lambda_i(t) = \sum_{j:s_j<t} \gamma_j \exp(\mathbf{Z}'_{ij}\boldsymbol{\beta})$$

The full likelihood function is

$$\mathrm{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \{1 - \exp[-\Lambda_i(R_i)]\}^{\Delta_{i1}} \{\exp[-\Lambda_i(L_i)] - \exp[-\Lambda_i(R_i)]\}^{\Delta_{i2}} \{\exp[-\Lambda_i(L_i)]\}^{\Delta_{i3}}$$

where Δ_{i1} indicates whether the *i*th subject is left-censored ($L_i = 0$), Δ_{i2} indicates whether the *i*th subject is interval-censored ($0 < L_i < R_i < \infty$), and Δ_{i3} indicates whether the *i*th subject is right-censored ($R_i = \infty$).

As the following derivation shows, the EM algorithm can be adapted straightforwardly to fit the semiparametric model that contains time-dependent covariates.

Let $b_j(t) = I(s_j \le t)$, and redefine the latent Poisson variables as

$$E(W_{ij}) = \gamma_j \exp(\mathbf{Z}'_{ij}\boldsymbol{\beta}) b_j(R_i) = \gamma_j \exp(\mathbf{Z}'_{ij}\boldsymbol{\beta}) I(s_j < R_i)$$
$$E(U_{ij}) = \gamma_j \exp(\mathbf{Z}'_{ij}\boldsymbol{\beta})[b_j(R_i) - b_j(L_i)] = \gamma_j \exp(\mathbf{Z}'_{ij}\boldsymbol{\beta}) I(L_i < s_j < R_i)$$

The expected complete-data log likelihood $\mathrm{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)}) = E[\log(\mathrm{L}_c(\boldsymbol{\theta}))|\mathcal{D}, \boldsymbol{\theta}^{(d)}]$ becomes

$$\begin{aligned}\mathrm{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)}) \;=\; & \sum_{i=1}^{n}\sum_{k=1}^{J} \Big\{ [w_{ik} + (\Delta_{i2} + \Delta_{i3})u_{ik}] \Big[\log(\gamma_k) + \mathbf{Z}'_{ij}\boldsymbol{\beta}\Big] \\ & -\gamma_k \exp(\mathbf{Z}'_{ij}\boldsymbol{\beta}) \left[(\Delta_{i2} + \Delta_{i1}) b_j(R_i) + \Delta_{i3} b_j(L_i)\right] \Big\} + B(\boldsymbol{\theta}^{(d)})\end{aligned}$$

where $B(\boldsymbol{\theta}^{(d)})$ is a constant and w_{ik} and u_{ik} are computed as follows:

$$w_{ik} = \frac{\gamma_k^{(d)} b_j(R_i) \exp(\mathbf{Z}'_{ij}\boldsymbol{\beta}) w_i}{\Lambda_i^{(d)}(R_i)}$$

$$u_{ik} = \frac{\gamma_k^{(d)} [b_j(R_i) - b_j(L_i)] \exp(\mathbf{Z}'_{ij}\boldsymbol{\beta}) u_i}{\Lambda_i^{(d)}(R_i) - \Lambda_i^{(d)}(L_i)}$$

$$w_i = \frac{\Lambda_i^{(d)}(R_i)\Delta_{i1}}{1 - \exp\left[\Lambda_i^{(d)}(R_i)\right]}$$

$$u_i = \frac{[\Lambda_i^{(d)}(R_i) - \Lambda_i^{(d)}(L_i)]\Delta_{i2}}{1 - \exp\left\{\left[\Lambda_i^{(d)}(R_i) - \Lambda_i^{(d)}(L_i)\right]\right\}}$$

You use the ID statement to fit the semiparametric model that contains time-dependent covariates. The levels of the ID variable identify the subjects to be analyzed.

Variance Estimation

Louis's Method

Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ be the maximum likelihood estimates as found by the EM algorithm. Under suitable conditions, you can apply Louis's method (Louis 1982) to obtain the covariance matrix of $\hat{\boldsymbol{\theta}}$.

The observed information matrix is computed as

$$I(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2 \mathrm{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} - \mathrm{cov}\left\{ \frac{\partial \log \mathrm{L}_c(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right\}$$

and its inverse $I^{-1}(\hat{\boldsymbol{\theta}})$ is the estimated covariance of $\hat{\boldsymbol{\theta}}$.

Louis's method is the default method of calculating standard errors for the semiparametric model.

Profile Likelihood Method

You can use the profile likelihood method of Murphy and Van Der Vaart (2000) to estimate the covariance matrix of $\hat{\boldsymbol{\beta}}$. The profile log-likelihood function is defined as

$$\mathrm{PL}(\boldsymbol{\beta}) = \max_{\boldsymbol{\gamma} \in (D)} \log \mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\gamma})$$

where $\mathcal{D}$ is the parameter space of $\boldsymbol{\gamma}$.

The Hessian matrix of $\mathrm{PL}(\boldsymbol{\beta})$ can be computed using numerical differentiation. Let $\mathbf{e}_k$ be the identity vector for $k = 1, \ldots, p$, and let l be a small perturbation. The (j,k)th element of the Hessian matrix can be approximated by

$$H_{ij} = \frac{\mathrm{PL}(\boldsymbol{\beta}) - \mathrm{PL}(\boldsymbol{\beta} + l \cdot \mathbf{e}_j) - \mathrm{PL}(\boldsymbol{\beta} + l \cdot \mathbf{e}_k) - \mathrm{PL}(\boldsymbol{\beta} + l \cdot \mathbf{e}_j + l \cdot \mathbf{e}_k)}{l^2}$$

The covariance matrix of $\hat{\boldsymbol{\beta}}$ is estimated by inverting the negative of the Hessian matrix.

You can use the profile likelihood method for the semiparametric model by specifying the PLVARIANCE option in the MODEL statement. But be aware that this computation is iterative and can consume a relatively large amount of CPU time.

EMICM Algorithm

Pan (1999) proposes using the iterative convex minorant (ICM) algorithm to fit semiparametric proportional hazards models to interval-censored data.

Define $\alpha_j = \sum_{k=1}^{j} \gamma_k$. Denote $\alpha_0 = 0$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{J-1})'$. The full likelihood function can be rewritten in terms of $\boldsymbol{\alpha}$ and the regression coefficients $\boldsymbol{\beta}$.

Maximizing the likelihood with respect to $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta})$ is equivalent to maximizing it with respect to $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Because the α_j are naturally ordered, the optimization is subject to the following constraint:

$$C = \{\mathrm{x} = (\alpha_1, \ldots, \alpha_{J-1}) : 0 \le \alpha_1 \le \cdots \le \alpha_{J-1} \le 1\}$$

Denote the log-likelihood function as $l(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Because the regression coefficients $\boldsymbol{\beta}$ are not constrained, you can update them by using the one-step Newton-Raphson method as in the EM algorithm. Pan (1999) suggests using the ICM algorithm to update the baseline parameters $\boldsymbol{\alpha}$; doing so essentially treats $\boldsymbol{\beta}$ as fixed

and maximizes the function $l(\boldsymbol{\alpha}) = l(\boldsymbol{\alpha}|\boldsymbol{\beta})$. Suppose that the maximum of $l(\boldsymbol{\alpha})$ occurs at $\hat{\boldsymbol{\alpha}}$. Mathematically, it can be proved that $\hat{\boldsymbol{\alpha}}$ equals the maximizer of the following quadratic function,

$$g^*(\mathbf{x}|\mathbf{y}, \mathbf{W}) = -\frac{1}{2}(\mathbf{x}-\mathbf{y})'\mathbf{W}(\mathbf{x}-\mathbf{y})$$

where $\mathbf{y} = \hat{\boldsymbol{\alpha}} + \mathbf{W}^{-1}\nabla l(\hat{\boldsymbol{\alpha}})$, $\nabla l(\cdot)$ denotes the derivatives of $l(\cdot)$ with respect to $\boldsymbol{\alpha}$, and $\mathbf{W}$ is a positive definite matrix of size $(J-1)\times(J-1)$ (Groeneboom and Wellner 1992).

The ICM algorithm updates $\boldsymbol{\alpha}$ as follows. For the dth iteration, the algorithm updates the quantity

$$\mathbf{y}^{(d)} = \hat{\boldsymbol{\alpha}}^{(d-1)} - \mathbf{W}^{-1}(\hat{\boldsymbol{\alpha}}^{(d-1)})\nabla l(\hat{\boldsymbol{\alpha}}^{(d-1)})$$

where $\hat{\boldsymbol{\alpha}}^{(d-1)}$ is the parameter estimate from the previous iteration and $\mathbf{W}(\hat{\boldsymbol{\alpha}}^{(d-1)}) = \mathrm{diag}(w_j, j = 1,\ldots,J-1)$ is a positive definite diagonal matrix that depends on $\hat{\boldsymbol{\alpha}}^{(l-1)}$. A convenient choice for $\mathbf{W}(\boldsymbol{\alpha})$ is the negative of the second-order derivative of the log-likelihood function $l(\boldsymbol{\alpha})$:

$$w_j = w_j(\boldsymbol{\alpha}) = -\frac{\partial^2}{\partial\alpha_j^2}l(\boldsymbol{\alpha})$$

Given $\mathbf{y} = \mathbf{y}^{(d)} = (y_1^{(d)},\ldots,y_{J-1}^{(d)})'$ and $\mathbf{W} = \mathbf{W}(\hat{\boldsymbol{\alpha}}^{(d-1)})$, the parameter estimate $\hat{\boldsymbol{\alpha}}^{(d)}$ maximizes the quadratic function $g^*(\mathbf{x}|\mathbf{y}, \mathbf{W})$.

Define the cumulative sum diagram $\{P_k, k = 0,\ldots,J-1\}$ as a set of m points in the plane, where $P_0 = (0,0)$ and

$$P_k = \Big(\sum_{i=1}^{k} w_i, \sum_{i=1}^{k} w_i\, y_i^{(l)}\Big)$$

Technically, $\hat{\boldsymbol{\alpha}}^{(d)}$ equals the left derivative of the convex minorant, or in other words, the largest convex function below the diagram $\{P_k, k = 0,\ldots,J-1\}$. You can solve this optimization problem by using the pool-adjacent-violators algorithm (Groeneboom and Wellner 1992).

The EMICM algorithm combines the EM algorithm and the ICM algorithm by alternating the two different steps in its iterations. Whereas the EM step updates both the baseline parameters and the regression coefficients, the ICM step updates only the baseline parameters. If the ICM step does not increases the likelihood value, the parameter changes are halved for the next iteration. The process repeats a maximum of five times, until an increase in the likelihood value is found.

The EMICM algorithm is the default method of fitting the semiparametric model. You can use it to fit the piecewise constant hazard model by specifying the NLOPTIONS(TECH=EMICM) option in the PROC ICPHREG statement.

Input and Output Data Sets

OUT= Output Data Set in the BASELINE Statement

The OUT= data set in the BASELINE statement contains all the variables in the COVARIATES= data set, along with statistics you request by specifying *keyword=name* options. There are *n* observations in the OUT= data set for each observation in the COVARIATES= data set, where *n* is the number of break points for the fitted model.

Missing Values

Observations that contain a missing value for both the left boundary value and the right boundary 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 unless you specify the MISSING option in the STRATA statement. If any explanatory variable that you specify in the MODEL statement has a missing value, that observation is not used in the model fitting. Predicted values, residuals, and regression diagnostics are computed for all observations that have no missing explanatory variable values.

Displayed Output

PROC ICPHREG 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 type of model fitted, the parameterization type that is used for the baseline function, and the name and label of the dependent variables. If you specify the FREQ statement, this table also displays the name and label of the frequency variable.

Class Level Information

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

Class Variable Levels

If you specify a CLASS statement, PROC ICPHREG outputs the "Class Variable Levels" table. This table displays the levels of classification variables when the parameterization is not full rank.

Fit Statistics

The "Fit Statistics" table displays the negative of twice the log likelihood, Akaike's information criterion (AIC), the corrected Akaike's information criterion (AICC), and the Bayesian information criterion (BIC). This table is displayed by default.

Analysis of Maximum Likelihood Parameter Estimates

The "Analysis of Maximum Likelihood Parameter Estimates" table displays the parameter name, the degrees of freedom for each parameter, the maximum likelihood 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. This table is displayed by default.

Iteration History for Parameter Estimates

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

Last Evaluation of the Gradient

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

Last Evaluation of the Hessian

If you specify the ITHISTORY option in the PROC ICPHREG statement, the procedure displays the last evaluation of the Hessian matrix.

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.

Hazard Ratios for *label*

If you specify the HAZARDRATIO statement, PROC ICPHREG outputs the "Hazard Ratios for *label*" table. The table displays the estimate and confidence limits for each hazard ratio. The ODS name of the "Hazard Ratios for *label*" table is HazardRatios.

Parameter Information

If you specify the ITHISTORY, COVB, or CORRB option in the MODEL statement, PROC ICPHREG outputs the "Parameter Information" table. This table displays the names of the parameters and the corresponding level information of effects that contain the CLASS variables. The ODS name of the "Parameter Information" table is ParmInfo.

Constant Hazard Time Intervals

If a piecewise constant model is used (the default model), PROC ICPHREG outputs the "Constant Hazard Time Intervals" table. This table displays information about the parameterization of the baseline function under the piecewise constant model. The ODS name of the "Constant Hazard Time Intervals" table is HazardParms.

Cubic Spline Parameters

If you fit a cubic spline model, PROC ICPHREG outputs the "Cubic Spline Parameters" table. This table displays information about the parameterization of the baseline function under the cubic spline model. The ODS name of the "Cubic Spline Parameters" table is HazardParms.

ODS Table Names

PROC ICPHREG assigns a name to each table it creates. You can use these names to refer to the table when you use the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in Table 67.8. For more information about ODS, see Chapter 20, "Using the Output Delivery System."

Table 67.8 ODS Tables Produced by PROC ICPHREG

ODS Table Name	Description	Statement	Option
ClassLevelInfo	Design information for CLASS variables	CLASS	
ClassLevels	Classification variables levels	CLASS	PARAM=GLM
ConvergenceStatus	Convergence status	MODEL	
CorrB	Parameter estimate correlation matrix	MODEL	CORRB
CovB	Parameter estimate covariance matrix	MODEL	COVB
FitStatistics	Fit statistics	MODEL	
HazardRatios	Customized hazard ratio estimates	HAZARDRATIO	
LastGrad	Last evaluation of the gradient	PROC	ITHISTORY
LastHess	Last evaluation of the Hessian	PROC	ITHISTORY
IterHist	Iteration history	PROC	ITHISTORY
ModelInfo	Model and data information	MODEL	
NObs	Number of observations	MODEL	
OptInfo	Optimization Information	MODEL	
ParameterEstimates	Parameter estimates	MODEL	
ParmInfo	Regression effect names	CLASS	
HazardParms	Baseline parameters	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."

PROC ICPHREG assigns a name to each table that it creates. You can use these names to refer to the table when you use the Output Delivery System (ODS) to select tables and create output data sets. These names are listed separately in Table 67.9. For more information about ODS, see Chapter 20, "Using the Output Delivery System."

Table 67.9 Graphs Produced by PROC ICPHREG

ODS Graph Name	Plot Description	Statement	Option
CumhazPlot	Cumulative hazard function plot	PROC	PLOTS=CUMHAZ
DevianceResidByXBeta	Deviance residuals by linear predictor	PROC	PLOTS=RESDEV(XBETA)
DevianceResPlot	Deviance values	PROC	PLOTS=RESDEV
DiagnosticPlot	Panel of residuals and diagnostic plots	PROC	PLOTS=RESDEV RESLAG
HazardPlot	Hazard function plot	PROC	PLOTS=HAZARD
IntervalPlot	Interval lengths	PROC	PLOTS=INTERVAL
LagakosResidByXBeta	Lagakos residuals by linear predictor	PROC	PLOTS=RESLAG(XBETA)
LagakosResPlot	Lagakos residuals	PROC	PLOTS=RESLAG
SurvivalPlot	Survival function plot	PROC	PLOTS=SURVIVAL

Examples: ICPHREG Procedure

Example 67.1: Fitting Cubic Spline Models

This example illustrates how to use a cubic spline baseline hazard to fit a proportional hazards model.

Consider the HIV data set in the section "Getting Started: ICPHREG Procedure" on page 5135. The following statements request a cubic spline proportional hazards model and the hazard ratio between the two levels of the Stage variable.

```
proc icphreg data=hiv;
   class Stage / desc;
   model (Left, Right) = Stage / basehaz=splines;
   hazardratio Stage;
run;
```

Output 67.1.1 displays information about the fitted spline model.

Output 67.1.1 Model Information

The ICPHREG Procedure

Model Information	
Data Set	WORK.HIV
Left Boundary	Left
Right Boundary	Right
Baseline Hazard	Cubic Splines

If no suboption is specified for the spline model, PROC ICPHREG uses three knots, generating three spline coefficients. Output 67.1.2 shows the selected knots.

Output 67.1.2 Cubic Spline Coefficients

Cubic Spline Parameters	
Coefficient	**Knot**
Coef1	1
Coef2	11
Coef3	25

The table of parameter estimates for the spline model is displayed in Output 67.1.3.

Output 67.1.3 Parameter Estimates for the Spline Model

Analysis of Maximum Likelihood Parameter Estimates

Effect	Stage	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Coef1		1	-6.0630	3.2263	-12.3865	0.2605		
Coef2		1	1.4921	2.2568	-2.9311	5.9152		
Coef3		1	-0.3086	0.6708	-1.6233	1.0060		
Stage	1	1	1.9016	0.6662	0.5959	3.2072	8.15	0.0043
Stage	0	0	0.0000					

Output 67.1.4 shows the estimated hazard ratio between the two stages and the 95% confidence limits.

Output 67.1.4 Hazard Ratio Estimate for Stage Values 1 and 0

Hazard Ratios for Stage			
Description	Point Estimate	95% Wald Confidence Limits	
Stage 1 vs 0	6.697	1.815	24.711

The cubic spline model can be considered a generalization of the Weibull proportional hazards model. It reduces to the Weibull model when there are only two knots, in which case the degrees of freedom is one (DF=1). The Weibull model assumes that the cumulative hazard function is a straight line in the log time scale whereas cubic splines offer a richer set of shapes that have more knots. The following statements fit the spline model with DF=1:

```
proc icphreg data=hiv;
   class Stage / desc;
   model (Left, Right) = Stage / basehaz=splines(df=1);
   hazardratio Stage;
run;
```

The "Fit Statistics" table is displayed in Output 67.1.5.

Output 67.1.5 Fit Statistics for the Spline Model When DF=1

The ICPHREG Procedure

Fit Statistics	
-2 Log Likelihood	30.025
AIC (Smaller is Better)	36.025
AICC (Smaller is Better)	36.914
BIC (Smaller is Better)	40.327

The table of parameter estimates for the fitted spline model is displayed in Output 67.1.6.

Output 67.1.6 Parameter Estimates for the Spline Model When DF=1

Analysis of Maximum Likelihood Parameter Estimates								
Effect	Stage	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Coef1		1	-7.3481	2.4438	-12.1378	-2.5584		
Coef2		1	2.5420	0.8974	0.7831	4.3008		
Stage	1	1	1.8265	0.6132	0.6247	3.0283	8.87	0.0029
Stage	0	0	0.0000					

You can request that PROC LIFEREG fit an accelerated failure lifetime model by using the default distribution (Weibull). This would be equivalent to fitting the proportional hazards model by using a Weibull baseline hazard (Klein and Moeschberger 1997). The following statements fit the Weibull model:

```
proc lifereg data=hiv;
   class Stage;
   model (Left, Right) = Stage;
run;
```

The table of fit statistics is displayed in Output 67.1.7.

Output 67.1.7 Fit Statistics That Are Produced by PROC LIFEREG

The LIFEREG Procedure

Fit Statistics (Unlogged Response)	
-2 Log Likelihood	30.025
Weibull AIC (smaller is better)	36.025
Weibull AICC (smaller is better)	36.914
Weibull BIC (smaller is better)	40.327

The table of parameter estimates for the Weibull model is displayed in Output 67.1.8.

Output 67.1.8 Parameter Estimates That Are Produced PROC LIFEREG

Analysis of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	2.1722	0.1791	1.8211	2.5233	147.06	<.0001
Stage	0	1	0.7185	0.2711	0.1871	1.2499	7.02	0.0080
Stage	1	0	0.0000	.	.	.	.	.
Scale		1	0.3934	0.1389	0.1969	0.7858		
Weibull Shape		1	2.5420	0.8974	1.2726	5.0776		

Comparing Output 67.1.7 with Output 67.1.5, you can see that the two model fits produce identical likelihood values.

The Weibull shape estimate is equal to the second spline coefficient, but the rest of the parameter estimates are different. This is because PROC LIFEREG fits the Weibull model under the configuration of accelerated failure time models. The estimates of regression coefficients from PROC LIFEREG and PROC ICPHREG are proportional; their ratio equals the negative of the Weibull shape parameter. For example, the estimate –0.7185 from PROC LIFEREG can also be obtained by dividing the estimate 1.8265 from PROC ICPHREG by –2.5420.

Example 67.2: Plotting Predicted Survival and Cumulative Hazard Functions

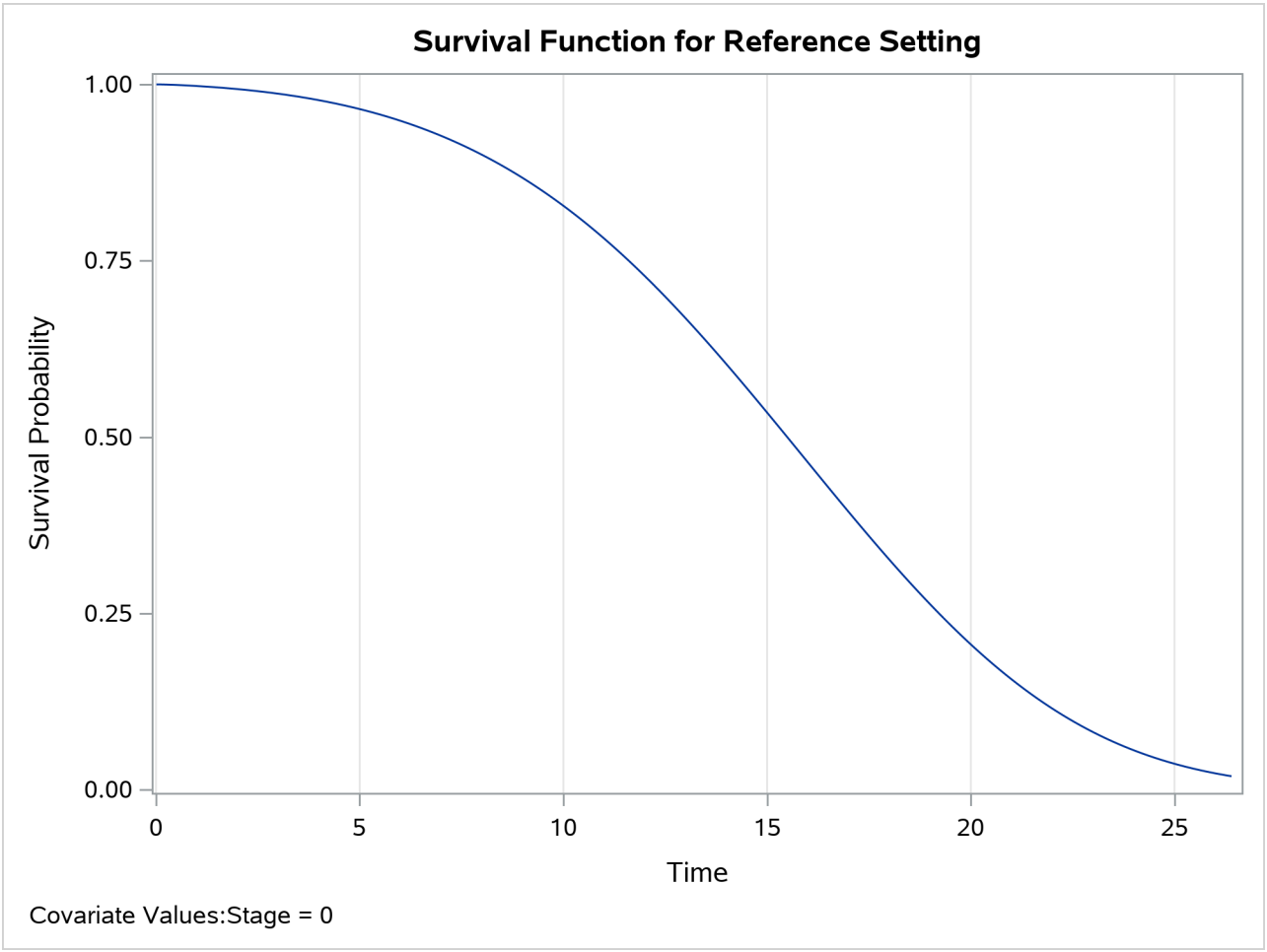
This example illustrates how to plot the predicted survival and cumulative hazard functions for specified covariate patterns.

The following statements request a plot of the estimated baseline survival function:

```
ods graphics on;
proc icphreg data=hiv plot=surv;
   class Stage / desc;
   model (Left, Right) = Stage / basehaz=splines;
run;
```

Output 67.2.1 shows the predicted survival curve at the reference level.

Output 67.2.1 Estimated Survival Curve for the Reference Set



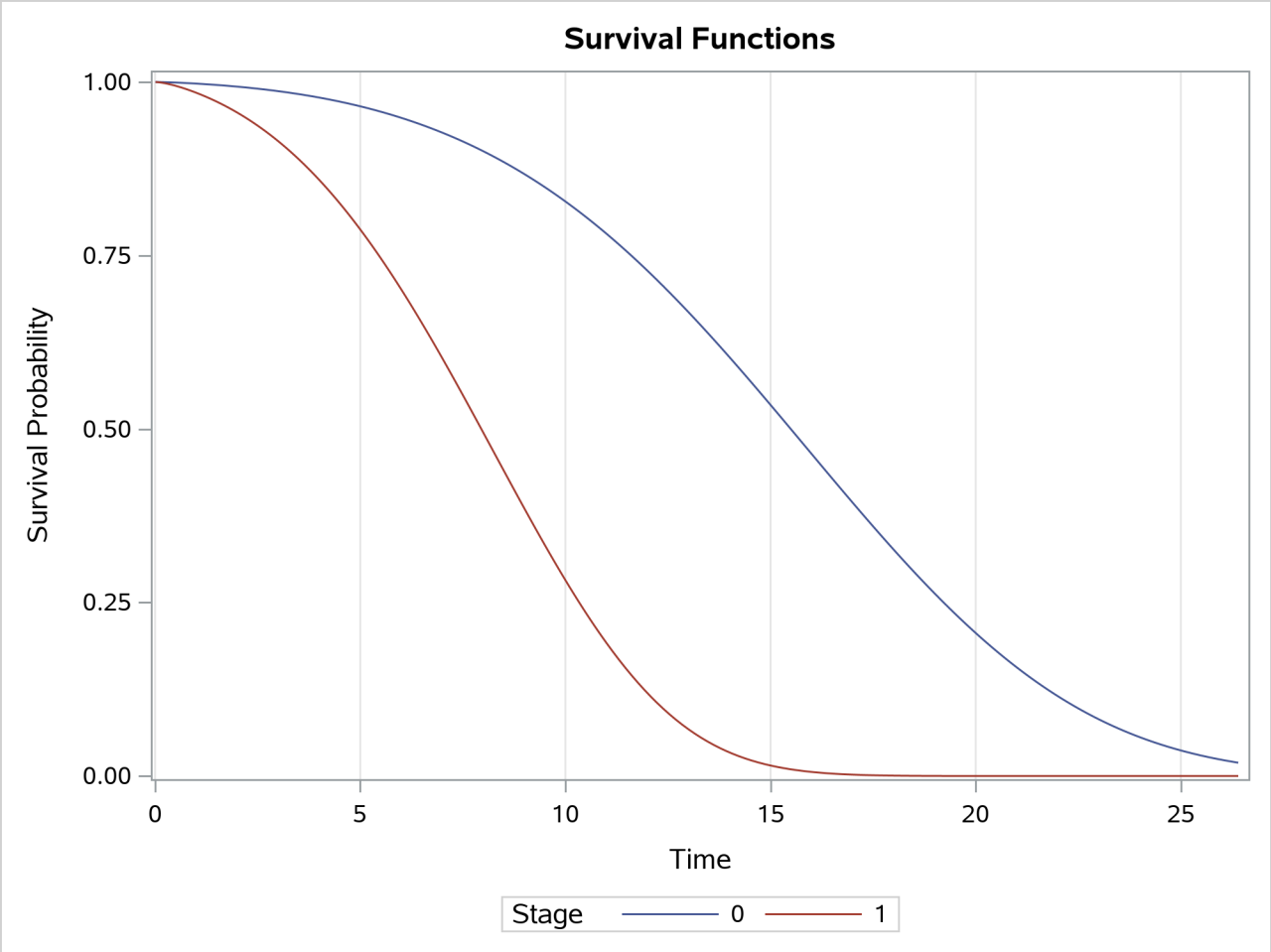
To produce curves for general covariate patterns, you can specify the COVARIATES= option in the BASELINE statement. The following statements create observations for two levels of Stage and plot the corresponding predicted curves:

```
data cov;
   Stage=0; output;
   Stage=1; output;
run;

proc icphreg data=hiv plot=surv;
   class Stage / desc;
   model (Left, Right) = Stage / basehaz=splines;
   baseline covariates=cov;
run;
```

Under the proportional hazards assumption, the two curves do not cross each other. As shown in Output 67.2.2, patients at Stage 1 have much lower survival rates than patients at Stage 0.

Output 67.2.2 Predicted Survival Curves for Specified Covariate Patterns

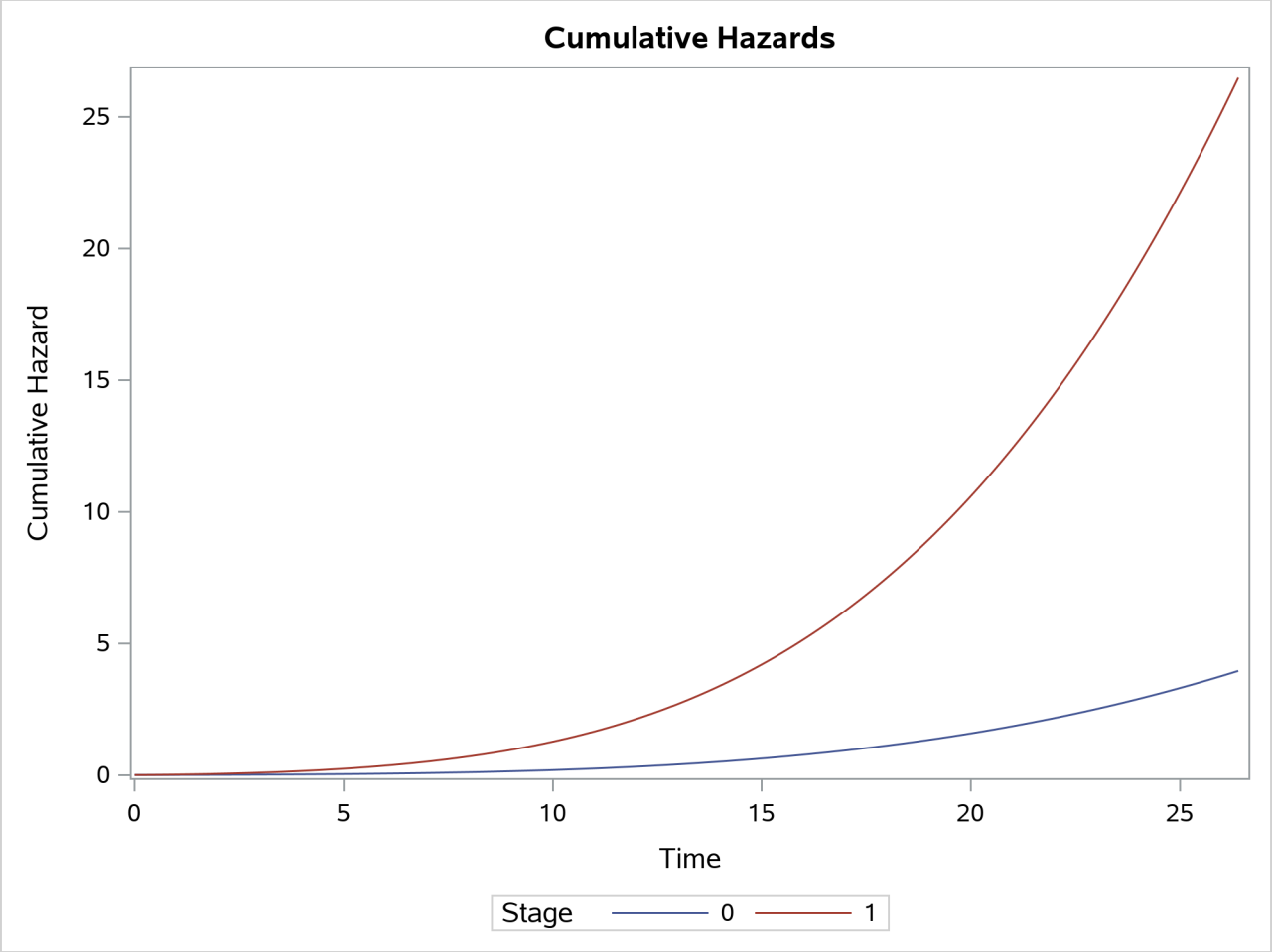


The following statements request a plot of the predicted cumulative hazard functions for the two levels of Stage:

```
proc icphreg data=hiv plot=cumhaz;
   class Stage / desc;
   model (Left, Right) = Stage / basehaz=splines;
   baseline covariates=cov;
run;
```

Output 67.2.3 shows the plot.

Output 67.2.3 Predicted Cumulative Hazards for Specified Covariate Patterns



Example 67.3: Fitting Stratified Weibull Models

This example illustrates how to fit stratified Weibull models by using the STRATA statement.

The following data set contains survival times for 36 patients who were diagnosed with a malignant kidney tumor. Most of the patients received the therapy of nephrectomy (removal of all or part of the kidney).

```
data hyper;
   input nephrectomy age time status @@;
   datalines;
0   1    9      1
0   1    6      1
0   1   21      1
0   2   15      1
0   2    8      1
0   2   17      1
0   3   12      1
1   1  104      0
1   1    9      1
1   1   56      1
```

```
1   1   35      1
1   1   52      1
1   1   68      1
1   1   77      0
1   1   84      1
1   1    8      1
1   1   38      1
1   1   72      1
1   1   36      1
1   1   48      1
1   1   26      1
1   1  108      1
1   1    5      1
1   2  108      0
1   2   26      1
1   2   14      1
1   2  115      1
1   2   52      1
1   2    5      0
1   2   18      1
1   2   36      1
1   2    9      1
1   3   10      1
1   3    9      1
1   3   18      1
1   3    6      1
;
```

The following statements convert the censored times into intervals:

```
data hyper;
   set hyper;
   left = time;
   if status = 0 then right = .;
   else right = time;
run;
```

The following statements fit a stratified Weibull proportional hazards model:

```
ods graphics on;
proc icphreg data=hyper plot(timerange=(0,125))=surv;
   class Age(desc);
   strata Nephrectomy;
   model (Left, Right) = Age / basehaz=splines(df=1);
run;
```

The "Cubic Splines Parameters" table, shown in Output 67.3.1, contains the parameters for the cubic splines. Because DF=1, it is equivalent to a Weibull distribution.

Output 67.3.1 Baseline Hazard Parameters

The ICPHREG Procedure

Cubic Spline Parameters		
Strata	**Coefficient**	**Knot**
0	**Coef1**	6
0	**Coef2**	21
1	**Coef3**	5
1	**Coef4**	115

The "Fit Statistics" table, shown in Output 67.3.2, contains the fit statistics.

Output 67.3.2 Model Fit Statistics

Fit Statistics	
-2 Log Likelihood	279.422
AIC (Smaller is Better)	291.422
AICC (Smaller is Better)	294.319
BIC (Smaller is Better)	300.923

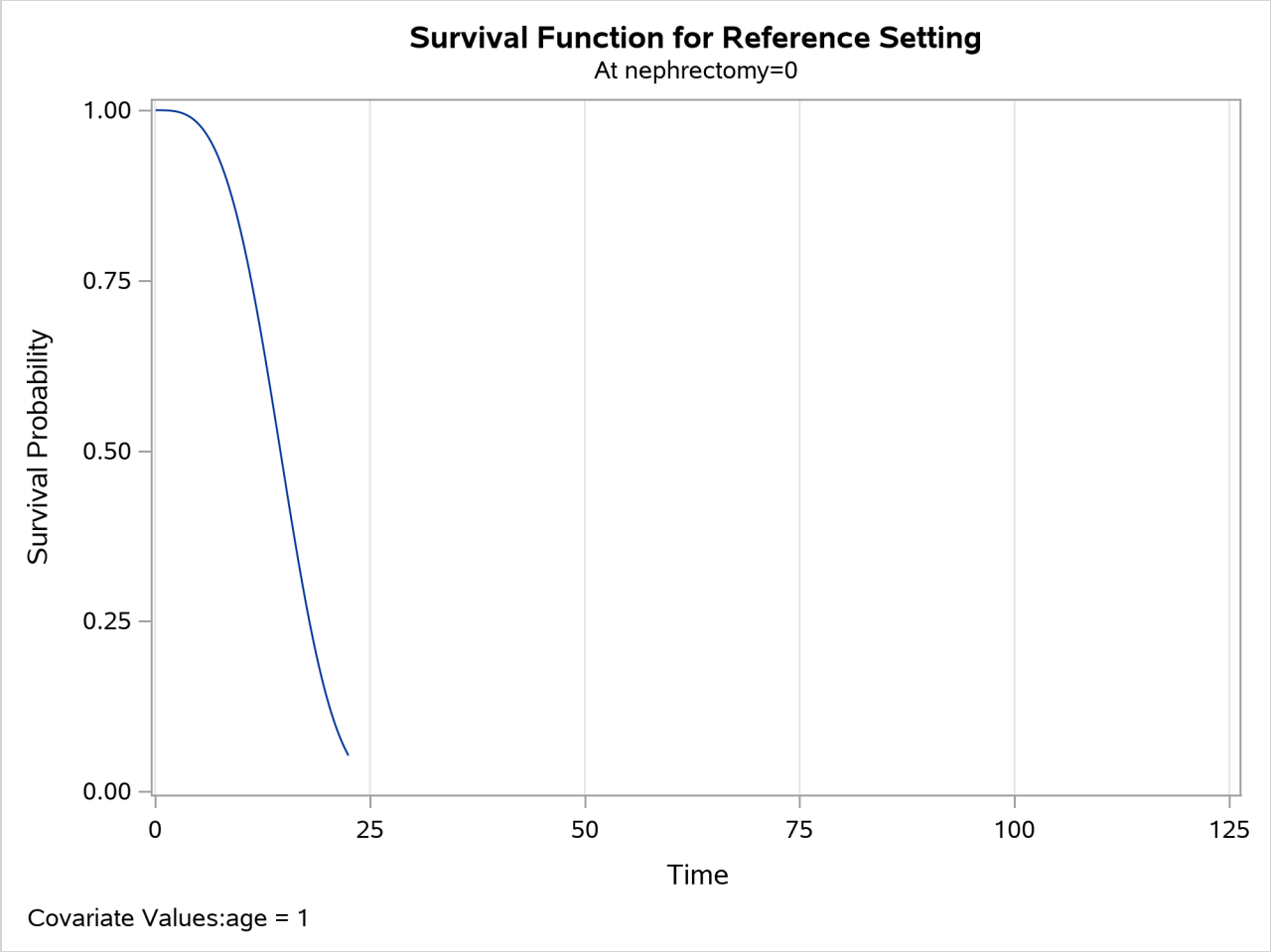
The table of parameter estimates is displayed in Output 67.3.3.

Output 67.3.3 Model Parameter Estimates

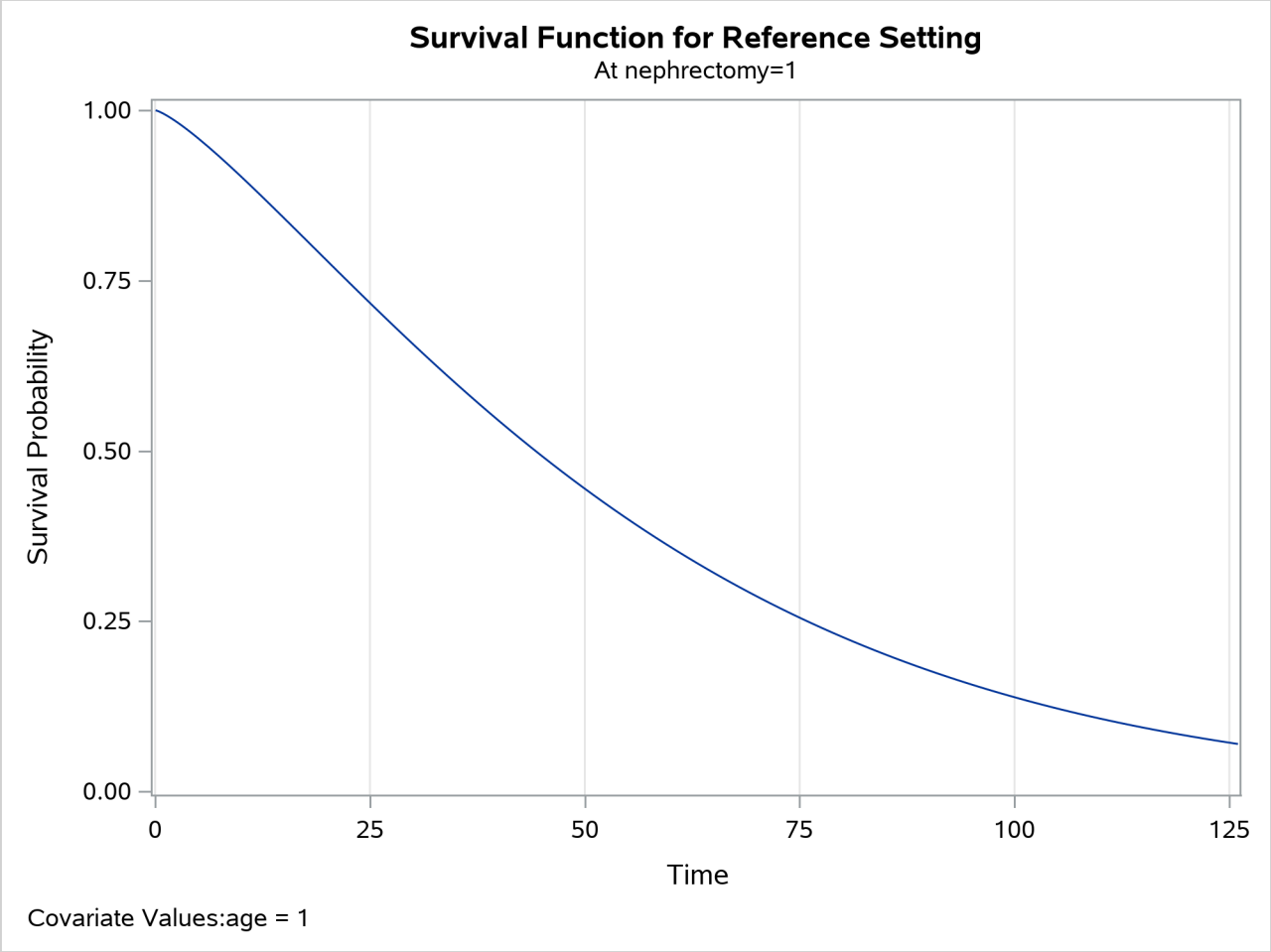
Analysis of Maximum Likelihood Parameter Estimates								
Effect	**age**	**DF**	**Estimate**	**Standard Error**	**95% Confidence Limits**		**Chi-Square**	**Pr > ChiSq**
Coef1		1	-9.2750	2.8327	-14.8269	-3.7231		
Coef2		1	3.3252	0.9918	1.3814	5.2690		
Coef3		1	-5.2408	0.8785	-6.9627	-3.5188		
Coef4		1	1.2861	0.2013	0.8915	1.6806		
age	3	1	1.8005	0.5966	0.6312	2.9697	9.11	0.0025
age	2	1	0.1272	0.4017	-0.6602	0.9146	0.10	0.7515
age	1	0	0.0000					

Output 67.3.4 and Output 67.3.5 show the predicted survival curves for the two strata at the reference level.

Output 67.3.4 Estimated Survival Curve at the Reference Level for Stratum 1



Output 67.3.5 Estimated Survival Curve at the Reference Level for Stratum 2



Example 67.4: Fitting Semiparametric Proportional Hazards Models

This example illustrates how to use PROC ICPHREG to fit a semiparametric proportional hazards model.

Gómez et al. (2009) analyzed data from a sample of 4,386 Belgian schoolchildren who were followed up for the emergence times of permanent upper-left first premolars (teeth 24 in European dental notation). The following DATA step creates the SAS data set Tooth:

```
proc format;
   value Sex 0='Male'
             1='Female';
run;

data Tooth;
   input Id Left Right Sex Dmf @@;
   format Sex Sex.;
   datalines;
   1 2.7 3.5 1 1      2 2.4 3.4 0 1      3 4.5 5.5 1 0      4 5.9   . 1 0
   5 4.1 5.0 1 1      6 3.7 4.5 0 1      7 4.9 5.8 0 1      8 5.4 6.5 1 1
   9 4.0 4.9 1 1     10 5.9 6.7 1 0     11 5.8 6.6 1 1     12 6.0   . 1 0
  13 3.3   . 0 0     14 3.6 4.6 0 1     15 6.1 7.0 0 0     16 2.9 3.7 0 1

   ... more lines ...

4421 2.5   . 0 0   4422 3.9 4.8 1 0   4423 3.0   . 1 0   4424 2.8   . 1 0
4425 3.3 4.3 1 1   4426 4.4 5.2 1 0   4427 2.8   . 1 0   4428 3.5   . 1 0
4429 5.0 6.0 0 0   4430 4.2   . 1 0
;
```

The data set Tooth contains the variables Left and Right, which are the two assessment times, in years since the age of five; Sex, which indicates the child's gender (0 for male or 1 for female); and Dmf, a binary variable that indicates whether the primary predecessor of this tooth was sound (0) or was decayed, missing, or filled (1).

The following statements use PROC ICPHREG to fit a semiparametric proportional hazards model to these data:

```
proc icphreg data=Tooth;
   class Sex;
   model (Left, Right) = Sex Dmf / base=unspecified;
   hazardratio Sex;
run;
```

Output 67.4.1 displays information about the semiparametric proportional hazards model.

Output 67.4.1 Model Information

The ICPHREG Procedure

Model Information	
Data Set	WORK.TOOTH
Left Boundary	Left
Right Boundary	Right
Baseline Hazard	Discrete

Output 67.4.2 displays information about the number of observations in the analysis.

Output 67.4.2 Number of Observations

Number of Observations Read	4386
Number of Observations Used	4386
Right Censored Observations	1611
Interval Censored Observations	2775

As you can see, there are 2,775 interval-censored observations and 1,611 right-censored observations.

Because the baseline function is not specified a priori in the semiparametric model, PROC ICPHREG finds potential jumps in the baseline cumulative hazard function by using the algorithm of Turnbull (1976). Output 67.4.3 shows the identified locations where jumps can occur.

Output 67.4.3 Parameters in the Baseline Cumulative Hazard Function

Hazard Increments		
Interval		
[Lower	Upper)	Hazard Parameter
2.5	2.6	Eta1
2.6	2.7	Eta2
2.7	2.8	Eta3
2.8	2.9	Eta4
2.9	3	Eta5
3	3.1	Eta6
3.1	3.2	Eta7
3.2	3.3	Eta8
3.3	3.4	Eta9
3.4	3.5	Eta10
3.5	3.6	Eta11
3.6	3.7	Eta12
3.7	3.8	Eta13
3.8	3.9	Eta14
3.9	4	Eta15
4	4.1	Eta16
4.1	4.2	Eta17
4.2	4.3	Eta18
4.3	4.4	Eta19
4.4	4.5	Eta20
4.5	4.6	Eta21
4.6	4.7	Eta22
4.7	4.8	Eta23
4.8	4.9	Eta24
4.9	5	Eta25
5	5.1	Eta26
5.1	5.2	Eta27
5.2	5.3	Eta28
5.3	5.4	Eta29
5.4	5.5	Eta30
5.5	5.6	Eta31
5.6	5.7	Eta32
5.7	5.8	Eta33
5.8	5.9	Eta34
5.9	6	Eta35
6	6.1	Eta36
6.1	6.2	Eta37
6.2	6.3	Eta38
6.3	6.4	Eta39
6.4	6.5	Eta40
6.5	6.6	Eta41
6.6	6.7	Eta42
6.7	6.8	Eta43
6.8	6.9	Eta44
6.9	7	Eta45
7	7.1	Eta46

Output 67.4.3 *continued*

Hazard Increments		
Interval		
[Lower	Upper)	Hazard Parameter
7.1	7.2	Eta47
7.2	7.3	Eta48
7.3	7.4	Eta49
7.4	Infty	Eta50

The last parameter is redundant.

Output 67.4.4 displays information about the optimization.

Output 67.4.4 Optimization Information

Optimization Information	
Optimization Technique	EMICM
Parameters in Optimization	51
Fixed Parameters	7

The EMICM algorithm is used to fit the semiparametric model. There are 51 parameters to be estimated, and 7 of them are set to zero during the iterative fitting process.

The table of parameter estimates for the semiparametric model is displayed in Output 67.4.5. It appears that both the child's gender and the previous status of the primary predecessor of the tooth are strongly associated with the outcome of a tooth emergency.

Output 67.4.5 Parameter Estimates for the Semiparametric Model

Analysis of Maximum Likelihood Parameter Estimates

Effect	Sex	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Eta1		1	0.0036	0.0025	0.0000	0.0085		
Eta2		1	0.0042	0.0041	0.0000	0.0122		
Eta3		1	0.0007	0.0046	0.0000	0.0097		
Eta4		1	0.0009	0.0040	0.0000	0.0087		
Eta7		1	0.0028	0.0049	0.0000	0.0124		
Eta8		1	0.0200	0.0078	0.0046	0.0353		
Eta9		1	0.0055	0.0082	0.0000	0.0215		
Eta11		1	0.0158	0.0088	0.0000	0.0330		
Eta12		1	0.0073	0.0109	0.0000	0.0287		
Eta13		1	0.0027	0.0103	0.0000	0.0230		
Eta15		1	0.0030	0.0109	0.0000	0.0244		
Eta16		1	0.0438	0.0120	0.0204	0.0673		
Eta18		1	0.0183	0.0142	0.0000	0.0462		
Eta19		1	0.0178	0.0167	0.0000	0.0506		
Eta20		1	0.0120	0.0180	0.0000	0.0472		
Eta21		1	0.0002	0.0181	0.0000	0.0357		
Eta22		1	0.0397	0.0197	0.0011	0.0783		
Eta23		1	0.0284	0.0233	0.0000	0.0740		
Eta24		1	0.0428	0.0249	0.0000	0.0915		
Eta25		1	0.0520	0.0286	0.0000	0.1080		
Eta26		1	0.0229	0.0312	0.0000	0.0842		
Eta27		1	0.0481	0.0329	0.0000	0.1126		
Eta28		1	0.0357	0.0353	0.0000	0.1048		
Eta29		1	0.0631	0.0366	0.0000	0.1348		
Eta30		1	0.0346	0.0336	0.0000	0.1006		
Eta32		1	0.0607	0.0385	0.0000	0.1362		
Eta33		1	0.0673	0.0482	0.0000	0.1617		
Eta34		1	0.1016	0.0522	0.0000	0.2039		
Eta35		1	0.0609	0.0571	0.0000	0.1729		
Eta36		1	0.0259	0.0601	0.0000	0.1436		
Eta37		1	0.0908	0.0619	0.0000	0.2121		
Eta38		1	0.0283	0.0683	0.0000	0.1622		
Eta39		1	0.0731	0.0732	0.0000	0.2165		
Eta40		1	0.1113	0.0800	0.0000	0.2681		
Eta41		1	0.0553	0.0892	0.0000	0.2301		
Eta42		1	0.1737	0.0873	0.0025	0.3448		
Eta44		1	0.0724	0.1145	0.0000	0.2969		
Eta45		1	0.0882	0.1490	0.0000	0.3802		
Eta46		1	0.1572	0.1918	0.0000	0.5332		
Eta47		1	0.1524	0.2622	0.0000	0.6664		
Eta48		1	0.3081	0.3630	0.0000	1.0196		
Eta49		1	0.3726	0.6425	0.0000	1.6319		
Sex	Female	1	0.3216	0.0387	0.2457	0.3975	69.04	<.0001
Dmf		1	0.3352	0.0387	0.2594	0.4111	75.04	<.0001

Two types of parameters are present in Output 67.4.5: the hazard parameters (Eta1, Eta2, . . . , Eta5) and the regression coefficients for the covariates. If a hazard parameter has been set to zero during the optimization process, or if a column of the Hessian matrix that corresponds to that parameter is found to linearly depend on columns that correspond to preceding model parameters, then PROC ICPHREG does not display that parameter. The procedure does not display the chi-square statistic and associated *p*-value for the hazard parameters.

Output 67.4.6 shows the estimated hazard ratio between boys and girls and the 95% confidence limits. The hazard of tooth emergency for girls is 1.377 times that for boys.

Output 67.4.6 Hazard Ratio Estimate for Girls and Boys

Hazard Ratios for Sex

Description	Point Estimate	95% Wald Confidence Limits	
Sex Female vs Male	1.379	1.279	1.488

The default method of computing the standard errors is Louis's method (Louis 1982). Alternatively, you can calculate the standard error estimates by using the profile likelihood method. The following statements use PROC ICPHREG to fit the same model and compute the profile-likelihood-based standard errors:

```
proc icphreg data=Tooth;
   class Sex;
   model (Left, Right) = Sex Dmf / base=unspecified plvariance;
   hazardratio Sex;
run;
```

Output 67.4.7 shows the estimated hazard ratio between boys and girls and the 95% confidence limits.

Output 67.4.7 Hazard Ratio Estimate for Girls and Boys

The ICPHREG Procedure

Hazard Ratios for Sex

Description	Point Estimate	95% Wald Confidence Limits	
Sex Female vs Male	1.379	1.279	1.488

The hazard ratio and standard errors based on the profile likelihood are almost identical to the previous ones, which were computed by Louis's method.

Example 67.5: Fitting Semiparametric Models with Time-Dependent Covariates

This example illustrates how to use PROC ICPHREG to fit a semiparametric proportional hazards model that contains time-dependent covariates.

Sparling et al. (2006) analyzed data from a long-term study about the effect of glucose exposure on retinopathy in young diabetes patients. The blood protein hemoglobin A1c (HbA1c) measures the level of glucose exposure (glycemia) over the preceding 68 weeks. More than a thousand patients participated in the study. Researchers measured their level of HbA1c at initial screening and their mean level of HbA1c during the study. Patients were assessed multiple times for retinopathy over a period of four years.

The following DATA step creates the data set Retinopathy:

```
data Retinopathy;
   input Duration Id Status CurrentHbA Time InitHbA PreHbA PreTime PreRetin;
   datalines;
   178       1     0     7.7500   988   9.63   7.9954   114    0
   178       1     0     7.7250  1618   9.63   7.9954   114    0
   178       1     0     7.9400  1956   9.63   7.9954   114    0
   175       3     0     8.8000  1111   7.93   8.1532   108    0
   175       3     0     8.6750  1859   7.93   8.1532   108    0
    72       5     1     8.2500  1027   8.53   7.8703   108    0
   106       6     0     8.2000  1110   7.26   8.5547   108    0
   106       6     0     8.5000  1838   7.26   8.5547   108    0
   147       8     0     8.3000  1202   9.27   6.8011   108    0
   147       8     0     8.3500  1834   9.27   6.8011   108    0

   ... more lines ...

    39    1438     0     8.6500  1610   7.30   8.0938    42    1
    30    1439     0     8.6750  1647   9.30   9.2750    42    1
    22    1440     0     8.0000  1702   6.60   6.9733    42    1
;
```

The variable Time represents the assessment time in days. The variable Status indicates whether retinopathy occurred before the associated assessment time. The variable Id identifies the subjects. The variable Duration represents how long the subject has had diabetes, in months. The variable PreHbA represents the mean HbA1c level in a previous study. The variable CurrentHbA represents the current mean HbA1c level at each assessment. The variable InitHbA represents the HbA1c level at initial screening. The variable PreTime represents the time in months that the subject spent in the previous study. The binary variable PreRetin indicates whether retinopathy was present prior to the current study.

The following statements use PROC ICPHREG to fit a semiparametric proportional hazards model to these data:

```
proc icphreg data=Retinopathy;
   model Time*Status(0) = Duration PreRetin InitHbA PreHbA PreTime CurrentHbA
                        / base=unspecified plvariance;
   id id;
   hazardratio CurrentHbA;
run;
```

The PLVARIANCE option computes the standard error estimates of the regression coefficients on the basis of the profile likelihood function (Zeng, Mao, and Lin 2016). The ID statement identifies the variable that represents the study subjects. The HAZARDRATIO statement displays hazard ratios for the current HbA1c level.

Output 67.5.1 displays information about the semiparametric proportional hazards model.

Output 67.5.1 Model Information

The ICPHREG Procedure

Model Information	
Data Set	WORK.RETINOPATHY
Dependent Variable	Time
Censoring Variable	Status
Censoring Value(s)	0
Baseline Hazard	Discrete

Output 67.5.2 displays the levels of the ID variable.

Output 67.5.2 Subject Identification Values

Class Level Information		
Name	**Levels**	**Values**
Id	1315	1 3 5 6 8 9 11 14 15 16 17 18 19 20 21 22 24 25 27 28 30 32 35 36 37 38 39 40 42 43 44 45 46 47 48 49 50 51 52 54 55 56 57 58 59 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 77 78 79 80 81 82 83 84 86 87 88 89 90 91 92 93 94 95 96 97 98 99 101 102 103 ...

Output 67.5.3 displays information about the number of observations in the analysis.

Output 67.5.3 Number of Observations

Number of Observations Read	2615
Number of Observations Used	2615
Right Censored Observations	1085
Interval Censored Observations	105
Left Censored Observations	125

As you can see, 1,085 patients are right-censored, 105 patients are interval-censored, and 125 patients are left-censored.

Output 67.5.4 displays information about the optimization.

Output 67.5.4 Optimization Information

Optimization Information	
Optimization Technique	EMICM
Parameters in Optimization	181
Fixed Parameters	155

The EMICM algorithm is used to fit the semiparametric model. There are 181 parameters that need to be estimated, and 155 of them are set to zero during the iterative fitting process.

The table of parameter estimates for the semiparametric model is displayed in Output 67.5.5.

Output 67.5.5 Parameter Estimates for the Semiparametric Model

Analysis of Maximum Likelihood Parameter Estimates

Effect	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Duration	1	0.0039	0.0020	0.0000	0.0077	3.90	0.0484
PreRetin	1	-0.2739	0.1948	-0.6556	0.1079	1.98	0.1597
InitHbA	1	-0.0845	0.0451	-0.1729	0.0039	3.51	0.0611
PreHbA	1	0.6918	0.0617	0.5709	0.8126	125.86	<.0001
PreTime	1	0.0062	0.0036	-0.0009	0.0133	2.94	0.0864
CurrentHbA	1	0.1262	0.0530	0.0223	0.2301	5.67	0.0173

Output 67.5.6 shows the estimated hazard ratio for the current HbA1c level and the 95% confidence limits.

Output 67.5.6 Hazard Ratio Estimate for HbA1c

Hazard Ratios for CurrentHbA

Description	Point Estimate	95% Wald Confidence Limits	
CurrentHbA Unit=1	1.135	1.023	1.259

For the time-dependent variable HbA1c, the 95% confidence interval for the hazard ratio is above 1. This indicates that a higher level of current HbA1c is associated with an increased risk of retinopathy.

References

Akaike, H. (1979). "A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting." *Biometrika* 66:237–242.

Akaike, H. (1981). "Likelihood of a Model and Information Criteria." *Journal of Econometrics* 16:3–14.

Cai, T., and Betensky, R. A. (2003). "Hazard Regression for Interval-Censored Data with Penalized Spline." *Biometrics* 59:570–579. http://dx.doi.org/10.1111/1541-0420.00067.

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

Farrington, C. P. (2000). "Residuals for Proportional Hazards Models with Interval-Censored Survival Data." *Biometrics* 56:473–482.

Finkelstein, D. M. (1986). "A Proportional Hazards Model for Interval-Censored Failure Time Data." *Biometrics* 42:845–854.

Friedman, M. (1982). "Piecewise Exponential Models for Survival Data with Covariates." *Annals of Statistics* 10:101–113.

Gómez, G., Calle, M. L., Oller, R., and Langohr, K. (2009). "Tutorial on Methods for Interval-Censored Data and Their Implementation in R." *Statistical Modelling* 9:259–297.

Groeneboom, P., and Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Basel: Birkhäuser.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer-Verlag.

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

Larder, B. A., Darby, G., and Richman, D. D. (1989). "HIV with Reduced Sensitivity to Zidovudine (AZT) Isolated during Prolonged Therapy." *Science* 243:1731–1734. http://www.sciencemag.org/content/243/4899/1731.full.pdf.

Louis, T. A. (1982). "Finding the Observed Information Matrix When Using the EM Algorithm." *Journal of the Royal Statistical Society, Series B* 44:226–233.

Murphy, S. A., and Van Der Vaart, A. W. (2000). "On Profile Likelihood." *Journal of the American Statistical Association* 95:449–465.

Pan, W. (1999). "Extending the Iterative Convex Minorant Algorithm to the Cox Model for Interval-Censored Data." *Journal of Computational and Graphical Statistics* 8:109–120.

Royston, P., and Parmar, M. K. B. (2002). "Flexible Parametric Proportional-Hazards and Proportional-Odds Models for Censored Survival Data, with Application to Prognostic Modelling and Estimation of Treatment Effects." *Statistics in Medicine* 21:2175–2197.

Simonoff, J. S. (2003). *Analyzing Categorical Data*. New York: Springer-Verlag.

Sparling, Y. H., Younes, N., Lachin, J. M., and Bautista, O. M. (2006). "Parametric Survival Models for Interval-Censored Data with Time-Dependent Covariates." *Biostatistics* 7:599–614.

Turnbull, B. W. (1976). "The Empirical Distribution Function with Arbitrarily Grouped, Censored, and Truncated Data." *Journal of the Royal Statistical Society, Series B* 38:290–295.

Wang, L., McMahan, C. S., Hudgens, M. G., and Qureshi, Z. P. (2016). "A Flexible, Computationally Efficient Method for Fitting the Proportional Hazards Model to Interval-Censored Data." *Biometrics* 72:222–231.

Zeng, D., Mao, L., and Lin, D. Y. (2016). "Maximum Likelihood Estimation for Semiparametric Transformation Models with Interval-Censored Data." *Biometrika* 103:253–271.

Subject Index

Syntax Index