Paper 254-2010

Not Hazardous to Your Health: Proportional Hazards Modeling for Survey Data with the SURVEYPHREG Procedure

Pushpal K Mukhopadhyay
SAS Institute Inc., Cary, NC

ABSTRACT

The Cox proportional hazards model is widely used in practice to estimate the effects of covariates on survival function. The SAS/STAT® PHREG procedure is a long-standing workhorse for performing regression analysis for event-time data based on the proportional hazards model. However, special techniques are required when the subjects are selected through complex surveys with stratification, clustering, unequal selection probabilities, or any combination of these.

This paper introduces the SURVEYPHREG procedure, a new SAS/STAT procedure for finite population inference for the Cox proportional hazards model for complex surveys. PROC SURVEYPHREG maximizes a weighted partial likelihood to estimate the model parameters, where the weights are the inverse of the selection probabilities. Features include: Taylor series linearization and replication methods for variance estimation, domain estimation for subpopulations, extensive residual analysis for assessing lack of fit, different likelihoods for handling ties, and hazard ratios. Examples of using PROC SURVEYPHREG are drawn from national longitudinal health surveys.

INTRODUCTION

Sample surveys commonly use complex designs that include stratification, clustering, and unequal selection probabilities. Thus, analyses of complex surveys might require modeling techniques that incorporate this survey design information (Särndal, Swensson, and Wretman 1992; Lohr 2009). SAS/STAT software provides specialized procedures to analyze complex survey data: SURVEYMEANS for means, SURVEYFREQ for frequencies, SURVEYREG for regression, and SURVEYLOGISTIC for logistic regression analysis. The new tool in this survey analysis toolbox is the SURVEYPHREG procedure, experimental in SAS/STAT 9.22, for Cox proportional hazards regressions for survey data. PROC SURVEYPHREG incorporates the design weights to estimate the proportional hazards regression parameters and computes design-based variance estimates, confidence intervals, and hypothesis tests for the parameters and model effects. This paper introduces the SURVEYPHREG procedure and discusses some of its important capabilities by using a data set from the National Health and Epidemiology Followup Study (NHEFS).

Survival analysis is used to model *time-to-event data*. For example, you might be interested in the time until the first heart attack for a given population. All persons with heart attacks have experienced an event, and the age of the first attack can be used as the time (response) variable. A key element of survival analysis is that it also includes the information from persons who have not yet had a heart attack—that is, censored observations. Persons without heart attacks are considered to be censored, and their current ages can be used as the responses. The Cox proportional hazards regression is a powerful tool for modeling the hazard function, the instantaneous rate of failure. This semiparametric model can adjust for covariates and can account for censored observations. See Kalbfleisch and Prentice (2002) and Lawless (2003) for details about survival analysis. The SAS/STAT PHREG procedure can be used to fit a proportional hazards model, but it does not account for stratification and clustering, which often occur in complex surveys. PROC SURVEYPHREG fills that gap. See Chapter 87, "The SURVEYPHREG Procedure" (*SAS/STAT User's Guide*), for detailed documentation of this procedure.

Although descriptive analyses such as means and frequency tables are the most common inferential objectives for sample surveys, survey data are also often used for modeling purposes (Korn and Graubard, 1999; Chamber and Skinner, 2003; and Kasprzyk et al., 1989). The application of proportional hazards models for surveys has become popular since the seminal works of Binder (1983, 1990, 1992) and Chambless and Boyle (1985). Binder (1992) estimated the parameters of a Cox proportional hazards model by incorporating the design weights in the partial likelihood. Binder also suggested a Taylor series linearized variance estimator for the proposed estimators of the regression parameters, and the theoretical properties of this estimator were explored by Lin and Wei (1989) and Lin (2000). You can use the SURVEYPHREG procedure to perform Binder's variance estimation technique and replication-based variance estimations.

The syntax for PROC SURVEYPHREG closely follows the syntax of PROC PHREG and the syntax of the other survey analysis procedures, SURVEYREG and SURVEYLOGISTIC. Specifically, consider the following SAS statements that fit a proportional hazards model of Age as a function of Gender, BloodChol, and Income, with the variable HeartAttack as the censor indicator.

```
proc surveyphreg data = nhefs;
   class gender;
   weight analysisweight;
   strata stratum;
   cluster psu;
   model age*heartattack(2) = gender bloodchol income;
   lsmeans gender / diff cl;
run;
```

The details of the syntax are discussed later, but for now note that the model is specified by using the CLASS and MODEL statements with the familiar PROC PHREG syntax; the survey design information is specified by using the WEIGHT, STRATA, and CLUSTER statements as in the other SAS/STAT survey procedures; and the postfitting least squares means analysis is specified by using the LSMEANS statement, which is common to many SAS/STAT linear modeling procedures (Tobias and Cai 2010).

The outline of this paper is as follows: the section "The NHEFS Data Set" discusses the example data set; the section "Estimates of Regression Parameters" discusses the model parameters and the output from the SURVEYPHREG procedure; the section "Replication-Based Variance Estimation Methods" discusses some other methods to estimate the variances of the proportional hazards regression estimators; the section "Postprocessing and Residual Analysis" discusses some postprocessing features; the section "Domain Analysis" discusses the domain (subpopulation) analysis (Lohr 2009); and the section "Conclusion" discusses some concluding remarks. In an appendix the mathematical, statistical, and computational details for key methodology are discussed.

THE NHEFS DATA SET

This paper uses a public use data set from the NHEFS survey. The NHEFS is a national longitudinal survey that is conducted by the National Center for Health Statistics and some other agencies of the Public Health Service in the United States. A cohort of size 14,407, comprised of all persons 25 to 74 years old who completed a medical examination during the National Health and Nutrition Examination Survey (NHANES) I in 1971–1975, was selected for the NHEFS. Personal interviews were conducted for every selected unit during the first wave of data collection from the years 1982 to 1984. Follow-up studies were conducted in 1986, 1987, and 1992. Vital and tracing status data, interview data, health care facility stay data, and mortality data for all four waves are available for public use. See National Center for Health Statistics (2010) for more information about the survey and the data sets.

The examples in this paper use 4,676 observations, associated with the detailed person records for NHANES Locations 1 to 100, from the 1992 NHEFS interview data set. All eligible units (interviewees) were asked if they ever experienced a heart attack. The age of all reported heart attacks were recorded for persons that have observed at least one heart attack. The age of the first reported heart attack is the event-time variable. If no heart attack is reported for a person, then his or her current age is the response and the person's heart attack time is considered to be censored. The examples use the following variables:

- Age, age at first reported heart attack or current age in 1992
- HeartAttack, event indicator (1 = reported heart attack, 2 = never experienced a heart attack)
- Gender, gender (1 = male, 2 = female)
- Income, individual income standardized to mean zero (if the reported income is not available for a person, then a random mean imputed income is used)
- BloodChol, blood cholesterol (1 = high blood cholesterol, 0 = high blood cholesterol is not reported)
- Race, race (1 = black, 2 = white, 3 = other)
- ObservationWeight, design weight for the observation
- AnalysisWeight, a standardized version of ObservationWeight used to facilitate convergence
- Stratum, stratum identification for variance estimation
- PSU, primary sampling unit identification for variance estimation

The data set and the SAS statements are used only for illustration purposes. The following SAS statements create the SAS data set NHEFS:

```
data nhefs;
   input stratum psu observationweight analysisweight age heartattack gender bloodchol income race;
datalines;
03 002 4821 0.0398429752 66 2 1 0 -11958.98 2
03 027 9875 0.0816115702 63 -1 1 1 -15102.98 1
03 049 39996 0.3305454545 70 2 2 0 65.97393285 2
03 062 51901 0.4289338843 57 2 1 1 -4558.98 2
03 068 17716 0.1464132231 68 2 2 0 -11958.98 2
03 006 3543 0.0292809917 78 2 2 0 -16566.98 1
03 014 5273 0.0435785124 50 2 2 0 -9558.98 1
03 018 3543 0.0292809917 77 -1 2 0 -10558.98 1
03 030 3462 0.0286115702 74 2 1 0 -1479.753849 1

   ... more lines ...

29 003 10795 0.089214876 73 2 2 1 -18930.98 2
29 003 10795 0.089214876 78 2 2 0 -13158.98 2
29 003 11044 0.0912727273 46 2 1 0 -11862.98 2
;
```

ESTIMATES OF REGRESSION PARAMETERS

Most surveys are designed to provide a cross-sectional look at a finite population. In contrast, longitudinal analysis of changes might require sampling a cohort of subjects from a finite population and following them through time. The inferential objectives for these surveys are usually either event-time analysis or change analysis across time. Every unit in the initial cohort can either experience an event (success or failure) or is assumed to be censored. The proportional hazards model is a semiparametric model that regresses the hazard function on a set of covariates and accounts for the censoring time. Consider the problem where T is a failure time random variable, $\mathbf{z}$ is a vector of covariates, and $\boldsymbol{\beta}$ is a set of regression parameters. The proportional hazards model is specified as

$$\lambda(t;\mathbf{z}) = \lambda_0(t)\exp(\mathbf{z}\boldsymbol{\beta})$$

where $\lambda(t;\mathbf{z})$ is the hazard function at time t for the observation unit with covariate $\mathbf{z}$, and $\lambda_0(t)$ is an arbitrary and unspecified baseline function. See Kalbfleisch and Prentice (2002) and Lawless (2003) for details about proportional hazards models. If all the units of a finite population U of size N are observed, then the regression parameters $\boldsymbol{\beta}$ can be estimated by maximizing an appropriate likelihood

$$l(\boldsymbol{\beta}) = \sum_{i \in U} \log\{L(\boldsymbol{\beta};\mathbf{z}_i,t_i)\}$$

where the responses are assumed to be uncorrelated for the working model and $L(\boldsymbol{\beta}) = \prod_{i\in U} L(\boldsymbol{\beta};\mathbf{z}_i,t_i)$ is the partial likelihood function defined in Cox (1972, 1975). Let $\boldsymbol{\beta}_N$ be the desired estimator. This is the likelihood that PROC PHREG works with.

In survey sampling, you do not observe the entire population. Instead a subset A of U is observed. Suppose the units in A are selected through a probability design $\pi(A)$ that assigns a selection probability π_i to every unit in the finite population. Then a sample-based estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_N$ is obtained by maximizing the *weighted partial likelihood*

$$l_\pi(\hat{\boldsymbol{\beta}}) = \sum_{i \in A} \pi_i^{-1} \log\{L(\boldsymbol{\beta};\mathbf{z}_i,t_i)\}$$

The SURVEYPHREG procedure maximizes this weighted partial likelihood (partial pseudo-likelihood) to obtain a sample-based estimator $\hat{\boldsymbol{\beta}}$ of the corresponding finite population quantity $\boldsymbol{\beta}_N$. The variance estimators of $\hat{\boldsymbol{\beta}}$ are obtained by assuming that the values of the finite population are fixed but unknown. This variance is commonly known as the design-based variance in the sample survey literature (Lohr 2009). Different forms of the likelihood function and of the variance estimators available in SURVEYPHREG are defined in the APPENDIX.

Returning to the NHEFS data introduced in the previous section, suppose you want to estimate the parameters of a proportional hazards regression model for the 1971–1975 base year survey population when the age at first reported heart attack is regressed on gender, blood cholesterol, and individual income. The following SAS statements request a proportional hazards model for Age on Gender, BloodChol, and Income, where HeartAttack is the censor indicator and –1, –7, –8, –9, and 2 are the censoring values. A negative value for HeartAttack denotes responses such as "inapplicable," "refused," "don't know," and "not ascertained"—different ways that a non-event (no reported heart attack) can be observed. For the examples used in this paper, an event is defined only if the person reports a heart attack. The PROC SURVEYPHREG statement invokes the procedure, and the MODEL statement specifies the analysis model. The CLASS statement specifies that Gender and BloodChol are categorical variables. The WEIGHT, STRATA, and CLUSTER statements identify the design weights, variance strata, and variance PSUs, respectively.

```
proc surveyphreg data = nhefs;
   class gender bloodchol;
   weight observationweight;
   strata stratum;
   cluster psu;
   model age*heartattack(-7 -8 -9 -1 2) = gender bloodchol income;
run;
```

The procedure first displays some summary information about the data and the model as shown in Figure 1.The "Model Information" table displays information about the analysis model, such as the name of the dependent variable, the censoring variable, the censoring values, the weight variable, the strata, and the cluster variable. The Ties Handling field shows that the Breslow likelihood is used, which is the default. Alternatively, you can use the TIES=EFRON option in the MODEL statement to specify the Efron likelihood. The Number of Observations Read and the Number of Observations Used are useful to determine the number of observation units that are not used by the procedure. The Sum of Weights Read and the Sum of Weights Used could represent the estimated population size and the estimated size of the set of respondents, respectively. A total of 4,676 observations are read from the input data set NHEFS, and they represent over 75 million units in the 1971–1975 NHEFS population. All observations are used to fit the model. The design summary table shown in Figure 2 displays the number of strata and PSUs. There are a total of 644 PSUs, and they are divided into 35 strata. The censored summary table shows that 95.42% units in the sample are censored, and the weighted censored count shows that an estimated 95.65% units are censored in the population—that is, 95.65% units in the population have never observed a heart attack. Finally, the "Variance Estimation" table shows that the procedure uses the Taylor series (linearization) method for variance estimation.

Figure 1 Summary Information

The SURVEYPHREG Procedure

Model Information

Data Set	WORK.NHEFS
Dependent Variable	age
Censoring Variable	heartattack
Censoring Value(s)	-7 -8 -9 -1 2
Weight Variable	observationweight
Stratum Variable	stratum
Cluster Variable	psu
Ties Handling	BRESLOW

Number of Observations Read	4676
Number of Observations Used	4676
Sum of Weights Read	75048207
Sum of Weights Used	75048207

Figure 2 Design Summary and Censored Summary

Design Summary

Number of Strata	35
Number of Clusters	644

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
4676	214	4462	95.42

Summary of the Weighted Number of Event and Censored Values

Total	Event	Censored	Percent Censored
75048207	3262099	71786108	95.65

Variance Estimation

Method	Taylor Series

Figure 3 displays the results of fitting the proportional hazards model. Although the likelihood ratio test (LRT) is a standard component of such analyses, the procedure uses only a weighted likelihood. Therefore, the LRT is sensitive to the scaling of the weights, making its interpretation problematic for survey data (Rao, Scott, and Skinner 1998). You should focus on the Wald test instead, which accounts for stratification and clustering. The Wald F test has a p-value of 0.004, which is based on a F distribution with two numerator degrees of freedom and 609 denominator degrees of freedom, which equal the number of PSUs minus the number of strata. The regression coefficient for Gender is estimated as 0.308 with a standard error of 0.136, the regression coefficient for BloodChol is estimated as –0.308 with a standard error of 0.138, and the regression coefficient for Income is estimated as –1.31E–6 with a standard error of 1.88E–6. Finally, the estimated hazard for males is 1.36 times the estimated hazard for females.

Figure 3 Global Tests and Parameter Estimates

Testing Global Null Hypothesis: BETA=0

Test	Test Statistic	Num DF	Den DF	p-Value
Likelihood Ratio	136447.250	3	Infty	<.0001
Wald	4.5252	3	609	0.0038

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
gender 1	609	0.307914	0.135731	2.27	0.0236	1.361
gender 2	609	0	.	.	.	1.000
bloodchol 0	609	−0.308293	0.137616	−2.24	0.0254	0.735
bloodchol 1	609	0	.	.	.	1.000
income	609	−0.000001306	0.000001882	−0.69	0.4880	1.000

The preceding example directly uses the observation weights. Observation weights range from 1,164 to 121,040 with a mean of 16,049 and a median of 12,323. Several observation weights have large values; therefore it is reasonable to rescale the observation weights to facilitate the optimization. The resulting analysis weights are created by dividing each observation weight by an arbitrary large number (121,000 in this case). These are the weights used in the rest of this paper. Because of this rescaling, you must take care in interpreting some output for this example, such as weighted counts.

REPLICATION-BASED VARIANCE ESTIMATION METHODS

Several variance estimation methods are available in PROC SURVEYPHREG. The Taylor series (linearization) method is the most commonly used variance estimation method for survey data, and it is the default variance estimation method used in PROC SURVEYPHREG. This method uses a sandwich variance estimator that incorporates the score residuals (Binder 1990).

An issue with the Taylor series method is that, if the survey you study involves stratified or clustered sampling, then you must provide this information to PROC SURVEYPHREG in order to compute variances by this method. However, because of privacy concerns, public-use survey data often does not include such stratum or cluster identification. Instead, a set of replicate weights is provided. You can use these replicate weights in PROC SURVEYPHREG to compute variances with replication-based methods. Even if stratum or cluster information is available, replication-based methods might be preferred over the Taylor series method—for example, in order to account for poststratification or unit nonresponse adjustment. See Mukhopadhyay et al. (2008) for details about replication-based variance estimation methods in SAS/STAT software. In addition to Taylor series linearization and replicate weights, you can use delete-1 jackknife and balanced repeated replication (BRR) in PROC SURVEYPHREG.

The following example shows how to use the delete-1 jackknife method in PROC SURVEYPHREG. The syntax for specifying the variance estimation method is the same as in other SAS/STAT survey analysis procedures. The VARMETHOD=JACKKNIFE option in the PROC SURVEYPHREG statement requests the delete-1 jackknife method. Most other statements are the same as in the previous section, except that the PARAM=REF option in the CLASS statement explicitly specifies a reference parameterization for the classification variables. By default, the procedure uses the GLM parameterization. See Chapter 19, “Shared Concepts and Topics” (*SAS/STAT User’s Guide*), for information about class variable parameterizations.

```
proc surveyphreg data = nhefs varmethod = jk;
   class gender bloodchol / param = ref;
   weight analysisweight;
   strata stratum;
   cluster psu;
   model age*heartattack(-7 -8 -9 -1 2) = gender bloodchol income;
run;
```

Figure 4 displays the results of this analysis. The "Variance Estimation" table shows that the jackknife method is used for variance estimation. A total of 644 replicate samples are obtained by deleting one PSU at a time, but one replicate sample could not be used due to nonconvergence. The "Analysis of Maximum Likelihood Estimates" table shows that the estimated regression coefficient for Gender is 0.308 with a standard error of 0.140, the estimated regression coefficient for BloodChol is –0.308 with a standard error of 0.141, and the estimated regression coefficient for Income is –1.31E–6 with a standard error of 2.37E–6. The estimated hazard for units without high blood cholesterol is 0.74 times the estimated hazard for units with high blood cholesterol. The denominator degrees of freedom for the *t* tests are 604, which equal the number of replicate samples that are used for variance estimation (643) minus the number of strata (35). Alternatively, you can use the DF= method-option for the VARMETHOD=JACKKNIFE option to specify a different value for the degrees of freedom.

The estimated proportional hazards regression coefficients in this example are the same as the estimated proportional hazards regression coefficients in the previous example, but the estimated variances are different. This to be expected, because the VARMETHOD= option only specifies a variance estimation method and has no impact on point estimates. Also, in this example, the delete-1 jackknife method produces higher estimated variances for the proportional hazards regression coefficients than the estimated variances for the Taylor series method.

Figure 4 Global Tests and Parameter Estimates for the Jackknife Method

The SURVEYPHREG Procedure

Variance Estimation

Method	Jackknife
Number of Replicates	644
Number of Replicates Used	643

Testing Global Null Hypothesis: BETA=0

Test	Test Statistic	Num DF	Den DF	p-Value
Likelihood Ratio	1.1277	3	Infty	0.7704
Wald	4.1207	3	608	0.0066

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
gender 1	608	0.307914	0.139664	2.20	0.0279	1.361
bloodchol 0	608	-0.308293	0.141326	-2.18	0.0295	0.735
income	608	-0.000001306	0.000002366	-0.55	0.5811	1.000

PROC SURVEYPHREG uses an iterative algorithm to maximize the weighted partial likelihood to obtain parameter estimates for the full sample and all replicate samples (subsamples). The iterative algorithm might not converge for some replicate samples even if it converges for the full sample. Hence, replicate estimates are not available for these replicate samples due to nonconvergence. It is also possible that one or more parameters are not estimable in some replicate samples even if they are estimable for the full sample. Estimability and nonconvergence are the two most common reasons why some replicate samples cannot be used for replicate variance estimation. In the preceding example, the procedure could not use one replicate sample due to nonconvergence. By default, the procedure uses only those replicate samples for which replicate estimates are available. If you use the jackknife method or the REPWEIGHT statement, you can specify the VADJUST=AVGREPSS option in the MODEL statement to use the average sum of squares for the invalid replicate samples. If you use the BRR method, you can specify the FAY= method-option for the VARMETHOD=BRR option to request Fay's BRR method (Fay 1984, 1989).

POSTPROCESSING AND RESIDUAL ANALYSIS

A statistical model must be appropriate for a given data set before it can be used for inferential purposes. PROC SURVEYPHREG produces summary statistics such as Akaike's information criterion, and the Wald test for the null model to assess the overall model. In addition, the procedure computes several observation-level statistics and residuals that you can use to assess the fitness of the model and to perform predictions at the model level. Those statistics include the linear predictors and their standard errors, the (weighted) number of observations that are at risk at the observation time, score residuals, martingale residuals, deviance residuals, and Schoenfeld residuals. See Kalbfleisch and Prentice (2002) for definitions of these residuals.

For example, the following SAS statements again request a proportional hazards regression of Age on Gender, BloodChol, and Income with HeartAttack as the censor indicator. The OUTPUT statement requests an output data set of observation-level statistics. The OUT= option names the output data set. The keywords name the statistics to include: the keywords BETA and STDXBETA request the linear predictor scores $\mathbf{z}\boldsymbol{\beta}$ and their standard errors, respectively; the keyword ATRISK requests the number of units at risk; and the keywords RESMART, RESDEV, and RESSCO request the martingale, deviance, and score residuals, respectively.

```
proc surveyphreg data = nhefs;
   class gender bloodchol;
   weight analysisweight;
   strata stratum;
   cluster psu;
   model age*heartattack(-7 -8 -9 -1 2) = gender bloodchol income;
   output out = FitStatistics xbeta stdxbeta atrisk resmart resdev ressco;
run;
```

When the model is appropriate for the data, then it can be used to draw inference about the population. For example, in the following statements, the ESTIMATE statement requests estimation of the difference in hazards for males without high blood cholesterol and females with high blood cholesterol at a common value of the continuous covariate Income:

```
proc surveyphreg data = nhefs;
   class gender bloodchol;
   weight analysisweight;
   strata stratum;
   cluster psu;
   model age*heartattack(-7 -8 -9 -1 2) = gender bloodchol income;
   estimate "Males Low vs Females High" gender 1 -1 bloodchol 1 -1;
   output out = FitStatistics xbeta stdxbeta atrisk resmart resdev ressco;
run;
```

The following example also includes the NLOPTIONS statement. The SURVEYPHREG procedure uses a nonlinear optimizer to maximize the weighted partial likelihood for the Cox model. The NLOPTIONS statement specifies different options for the optimizer. The GCONV= option in the NLOPTIONS statement specifies a relative gradient convergence criterion, and the PHISTORY option displays the optimization history. By default, the procedure uses the Newton-Raphson ridge optimization technique. You can use the TECHNIQUE= option in the NLOPTIONS statement to specify a different optimization technique.

```
proc surveyphreg data = nhefs;
   class gender bloodchol;
   weight analysisweight;
   strata stratum;
   cluster psu;
   model age*heartattack(-7 -8 -9 -1 2) = gender bloodchol income;
   nloptions gconv = 1E-11 phistory;
   estimate "Males Low vs Females High" gender 1 -1 bloodchol 1 -1;
   output out = FitStatistics xbeta stdxbeta atrisk resmart resdev ressco;
run;
```

Figure 5 displays some selected results that PROC SURVEYPHREG produces for this analysis. The regression coefficient for Gender is estimated as 0.308 with a standard error of 0.136, the regression coefficient for BloodChol is estimated as –0.308 with a standard error of 0.138, and the regression coefficient for Income is estimated as –1.318E–6 with a standard error of 1.88E–6. The "Estimate" table shows the estimated difference for hazards is –3.80E–4 with a standard error of 0.2148. The estimate has a t value of 0.00 with 609 degrees of freedom. Apparently, being female lowers the risk of a heart attack just about as much as having high cholesterol raises it.

Figure 5 Parameter Estimates and Tests

The SURVEYPHREG Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
gender 1	609	0.307914	0.135731	2.27	0.0236	1.361
gender 2	609	0	.	.	.	1.000
bloodchol 0	609	-0.308293	0.137616	-2.24	0.0254	0.735
bloodchol 1	609	0	.	.	.	1.000
income	609	-0.000001306	0.000001882	-0.69	0.4879	1.000

Estimate

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Males Low vs Females High	-0.00038	0.2148	609	-0.00	0.9986

Figure 6 displays the iteration history table and the optimization results that are requested by the PHISTORY option in the NLOPTIONS statement. The iteration history table displays information such as the iteration number, the value of the objective function, and the maximum of the absolute gradient elements at every iteration. The "Optimization Results" table displays summary information such as the number of iterations, the number of function calls, the value of the objective function, and the value of the maximum gradient element after the last iteration.

Figure 6 Iteration History

Maximum Likelihood Iteration History

Iter	Restarts	Function Calls	Active Constraints	Objective Function	Objective Function Change	Max Abs Gradient Element	Ridge	Ratio Between Actual and Predicted Change
1	0	4	0	-86.12280	0.5608	9677.5	0	0.979
2	0	6	0	-86.11973	0.00307	388.3	0	1.017
3	0	8	0	-86.11973	3.403E-6	0.8017	0	1.001
4	0	10	0	-86.11973	1.48E-11	3.513E-6	0	1.002

Optimization Results

Iterations	4	Function Calls	12
Hessian Calls	0	Active Constraints	0
Objective Function	-86.11973078	Max Abs Gradient Element	3.5130806E-6
Ridge	0	Actual Over Pred Change	1.0015000068

PROC SURVEYPHREG has a full complement of other postfitting analysis statements. In addition to the OUTPUT and ESTIMATE statements, it also shares with many other procedures the TEST statement for testing specific effects and the LSMEANS, LSMESTIMATE, and SLICE statements for analyzing group LS-means. It also incorporates the STORE statement, new in many SAS/STAT 9.22 procedures, for saving the fitted model. You can subsequently restore this fitted model for additional postfitting analysis in the new PLM procedure; see Tobias and Cai (2010).

DOMAIN ANALYSIS

You often need separate estimates for one or more specific subpopulations (domains). For example, you might want to estimate the proportional hazards model parameters separately for each race. Most frequently, the domain identification is not known at the design stage. For example, the race of a person might not be known before the person is interviewed. This means that domain sample sizes are not fixed, and this randomness needs to be taken into account. Even if the domain identification is known at the design stage, it is not always possible to allocate fixed sample sizes for every domain. Domain analysis for survey data accounts for this randomness by using the entire sample in estimating the variance of domain estimates.

The following SAS statements request a separate proportional hazards regression for each level of the DOMAIN variable race. Also, the TIES= option in the MODEL statement requests the Efron likelihood for handling ties.

```
proc surveyphreg data = nhefs;
   class gender;
   weight analysisweight;
   strata stratum;
   cluster psu;
   domain race;
   model age*heartattack(-7 -8 -9 -1 2) = gender / ties = efron;
run;
```

Figure 7, Figure 8, Figure 9, and Figure 10 display some selected results from the analysis. First the procedure performs an overall analysis for the entire population, and then it performs separate analyses for race. The estimated regression coefficient for Gender is 0.390 (Figure 7) with a standard error of 0.154 for the overall analysis, but it is estimated as –0.173 (Figure 8) with a standard error of 0.396 for the domain Race=1. The estimated regression coefficient for Gender is 0.459 (Figure 9) with a standard error of 0.171 for the domain Race=2, and the regression coefficient for Gender is estimated as 8.64 (Figure 10) with a standard error of 1.39 for the domain Race=3. Note that there are only 50 observation units and four events in the sample for the Race=3 domain; this sample might be too small for valid inference.

Figure 7 Parameter Estimates for the Overall Analysis

The SURVEYPHREG Procedure

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
4676	214	4462	95.42

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
gender 1	609	0.389562	0.153966	2.53	0.0117	1.476
gender 2	609	0	.	.	.	1.000

Figure 8 Parameter Estimates for Domain Race=1

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
445	31	414	93.03

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
gender 1	609	-0.172598	0.395761	-0.44	0.6629	0.841
gender 2	609	0	.	.	.	1.000

Figure 9 Parameter Estimates for Domain Race=2

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
4181	179	4002	95.72

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
gender 1	609	0.458713	0.171476	2.68	0.0077	1.582
gender 2	609	0	.	.	.	1.000

Figure 10 Parameter Estimates for Domain Race=3

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
50	4	46	92.00

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
gender 1	609	8.641462	1.392036	6.21	<.0001	5661.598
gender 2	609	0	.	.	.	1.000

CONCLUSION

The new SURVEYPHREG procedure is similar to the PHREG procedure in the sense that both procedures fit the proportional hazards model. However, whereas you use the PHREG procedure for likelihood-based inference (Kalbfleisch and Prentice 2002), you use the SURVEYPHREG procedure for model-assisted inference (Särndal, Swensson, and Wretman 1992). The SURVEYPHREG procedure is useful when your data are from complex surveys that involve stratification, clustering, and unequal weights. The point estimates of the regression coefficients from the SURVEYPHREG procedure are often the same as the point estimates of the regression coefficients from the PHREG procedure with the WEIGHT statement. But the estimated variances, confidence intervals, and *p*-values are not necessarily the same. Both procedures support a STRATA statement, but the statement has different meanings in the two procedures. The STRATA statement in the PHREG procedure fits a stratified hazards model (Kalbfleisch and Prentice 2002), but the STRATA statement in the SURVEYPHREG procedure estimates the variance for a stratified sample design.

With the introduction of the new SURVEYPHREG procedure, SAS/STAT users can now perform survival analysis for survey data. The SURVEYPHREG procedure is similar to the PHREG procedure and other SAS/STAT survey analysis procedures. More features, such as time-dependent covariates and the counting process style of input (Therneau 1994), will be added to PROC SURVEYPHREG in the future.

APPENDIX

This section discusses different forms of the weighted partial likelihood and the variance estimators that are used in PROC SURVEYPHREG. The materials discussed in this section can be found in Chapter 87, "The SURVEYPHREG Procedure" (*SAS/STAT User's Guide*). Without loss of generality, the rest of this section uses indices for stratified clustered designs. For a stratified clustered sample design, observations are represented by a matrix

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

where

- **w** denotes the vector of sampling weights
- **t** denotes the event-time variable
- $\mathbf{\Delta}$ denotes the event indicator
- **Z** denotes the $n \times p$ matrix of auxiliary information
- $h = 1, 2, \ldots, H$ is the stratum index
- $i = 1, 2, \ldots, n_h$ is the cluster index within stratum h
- $j = 1, 2, \ldots, m_{hi}$ is the unit index within cluster i of stratum h
- p is the total number of parameters
- $n = \sum_{h=1}^{H} \sum_{i=1}^{n_h} m_{hi}$ is the total number of observations in the sample
- $y_{hij}(t) = I(t_{hij} \geq t)$, where $I(\cdot)$ is an indicator function
- $n_{hij}(t) = I(t_{hij} \leq t)$, where $I(\cdot)$ is an indicator function

Let $\sum_B = \sum_{(h,i,j)\in B}$ denote the summation over the set of indices such that the observation unit j in PSU i and stratum h belongs to the index set B. Typically, B is the set of all population indices that are in the sample, the risk set, or the set of all units with a failure.

The first-stage sampling rate (fraction of PSUs selected for the sample) is denoted by f_h. The first-stage sampling rate is used in Taylor series variance estimation. If the RATE= option or the TOTAL= option is not specified in the PROC SURVEYPHREG statement, then the procedure assumes that the stratum sampling rates f_h are negligible and does not use a finite population correction when computing variances.

Weighted Partial Likelihoods for the Cox Model

Let $t_{(1)} < t_{(2)} < \ldots < t_{(K)}$ denote the K distinct, ordered event times. Let d_k denote the multiplicity of failures at $t_{(k)}$; that is, d_k is the size of the set $\mathcal{D}_k$ of individuals that fail at $t_{(k)}$. Let w_{hij} be the weight associated with the jth observation unit in the ith cluster in stratum h. Let $\mathcal{R}_k$ denote the risk set just before the kth ordered event time $t_{(k)}$.

Using this notation, the pseudo-likelihood functions used in PROC SURVEYPHREG to estimate $\boldsymbol{\beta}_N$ are described as follows. The Breslow likelihood is expressed as

$$L_{\text{Breslow}}(\boldsymbol{\beta}) = \prod_{k=1}^{K} \frac{\exp\left(\boldsymbol{\beta}' \sum_{\mathcal{D}_k} w_{hij}\mathbf{z}_{hij}\right)}{\left\{\sum_{\mathcal{R}_k} w_{hij}\, \exp(\boldsymbol{\beta}'\mathbf{z}_{hij})\right\}^{\sum_{\mathcal{D}_k} w_{hij}}}$$

The Efron likelihood is expressed as

$$L_{\text{Efron}}(\boldsymbol{\beta}) = \prod_{k=1}^{K} \frac{\exp\left(\boldsymbol{\beta}' \sum_{\mathcal{D}_k} w_{hij}\mathbf{z}_{hij}\right)}{\{\phi(\boldsymbol{\beta}, \mathbf{Z}, \mathbf{w}, k)\}^{\frac{1}{d_k}\sum_{\mathcal{D}_k} w_{hij}}}$$

where

$$\phi(\boldsymbol{\beta}, \mathbf{Z}, \mathbf{w}, k) = \prod_{l=1}^{d_k} \left\{\sum_{\mathcal{R}_k} w_{hij}\, \exp\left(\boldsymbol{\beta}'\mathbf{z}_{hij}\right) - \frac{l-1}{d_k}\sum_{\mathcal{D}_k} w_{hij}\, \exp\left(\boldsymbol{\beta}'\mathbf{z}_{hij}\right)\right\}$$

Note that if $d_k = 1$ for all k, then the two likelihoods are identical.

Taylor Series (Linearized) Variance Estimator

The Taylor series method is the default variance estimation method used by PROC SURVEYPHREG. Let

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

where $r = 0, 1$. Let

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

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

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

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

For TIES=EFRON, the computation of the score residuals is modified to comply with the Efron partial likelihood.

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

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

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

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

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

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

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

REFERENCES

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

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

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

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

Chambless, L. E. and Boyle, K. E. (1985), "Maximium Likelihood Methods for Complex Sample Data: Logistic Regression and Discrete Proportional Hazards Models," *Communications in Statistics, Part A: Theory and Methods*, 14(1), 1377–1392.

Cox, D. R. (1972), "Regression Models and Life Tables," *Journal of the Royal Statistical Society, Series B*, 20, 187–220, with discussion.

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

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

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

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

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

Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M. P. (1989), *Panel Surveys*, New York: John Wiley & Sons.

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

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

Lin, D. Y. (2000), "On Fitting Cox's Proportional Hazards Models to Survey Data," *Biometrika*, 87(1), 37–47.

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

Lohr, S. L. (2009), *Sampling: Design and Analysis*, Second Edition, Pacific Grove, CA: Duxbury Press.

Mukhopadhyay, P. K., Anthony, B. A., Tobias, R. D., and Watts, D. L. (2008), "Try, Try Again: Replication-Based Variance Estimation Methods for Survey Data Analysis in SAS 9.2," in *Proceedings of the SAS Global Forum 2008 Conference*, Cary, NC: SAS Institute Inc.

National Center for Health Statistics (2010), "NHANES I Epidemiologic Followup Study (NHEFS)," http://www.cdc.gov/nchs/nhanes/nhefs/nhefs.htm, last accessed February 19, 2010.

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

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

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

Tobias, R. and Cai, W. (2010), "Introducing PROC PLM and Postfitting Analysis for Very General Linear Models in SAS/STAT 9.22," in *Proceedings of the SAS Global Forum 2010 Conference*, Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Pushpal K Mukhopadhyay
SAS Institute Inc.
100 SAS Campus Drive
Cary, NC, 27513
919-531-2123
pushpal.mukhopadhyay@sas.com