Paper 338-2011

An Overview of Survival Analysis using Complex Sample Data

Patricia A. Berglund, Institute For Social Research-University of Michigan, Ann Arbor, Michigan

ABSTRACT

This paper presents practical guidance on conducting survival analysis using data derived from a complex sample survey. Survival curves, Cox models, and discrete-time logistic regression are demonstrated through use of PROC LIFETEST, PROC SGPLOT, PROC SURVEYPHREG and PROC SURVEYLOGISTIC. The analytic techniques presented can be used on any operating system and are intended for an intermediate level audience.

INTRODUCTION

The primary objective of this paper is to provide guidance for the analyst performing survival analysis using SAS® v9.2 with complex sample data. A short overview of survival analysis including theoretical background on time to event techniques is presented along with an introduction to analysis of complex sample data. These introductory sections are followed by a typical analytic progression of descriptive and inferential survival analyses using appropriate SAS SURVEY procedures.

The analysis examples include survival curves using the Kaplan-Meier method and regression models predicting onset of the event of interest using common covariates such as age at interview, race/ethnicity and gender. Cox Proportional Hazards and discrete-time logistic regression models are demonstrated and contrasted.

The descriptive examples focus on the use of PROC LIFETEST with ODS graphics to produce survival plots as well as plot generation using PROC SGPLOT with an output data set from the LIFETEST procedure. The modeling examples demonstrate the use of PROC SURVEYPHREG and PROC SURVEYLOGISTIC with selected options such as reference category specification, estimate and class statements, and model link options. Where possible, the analysis examples include use of the survey design variables and weights to correctly account for the complex sample design.

OVERVIEW OF SURVIVAL ANALYSIS

EVENT HISTORY DATA

Event history data is common in many disciplines and at its core, is focused on time. Analysis of event history data or survival analysis is used to refer to a statistical analysis of the time at which the event of interest occurs (Kalbfleisch and Prentice, 2002 and Allison, 1995). Event history data can be categorized into broad categories: 1. longitudinal data, 2. administrative follow-up data, and 3. retrospective event history data.

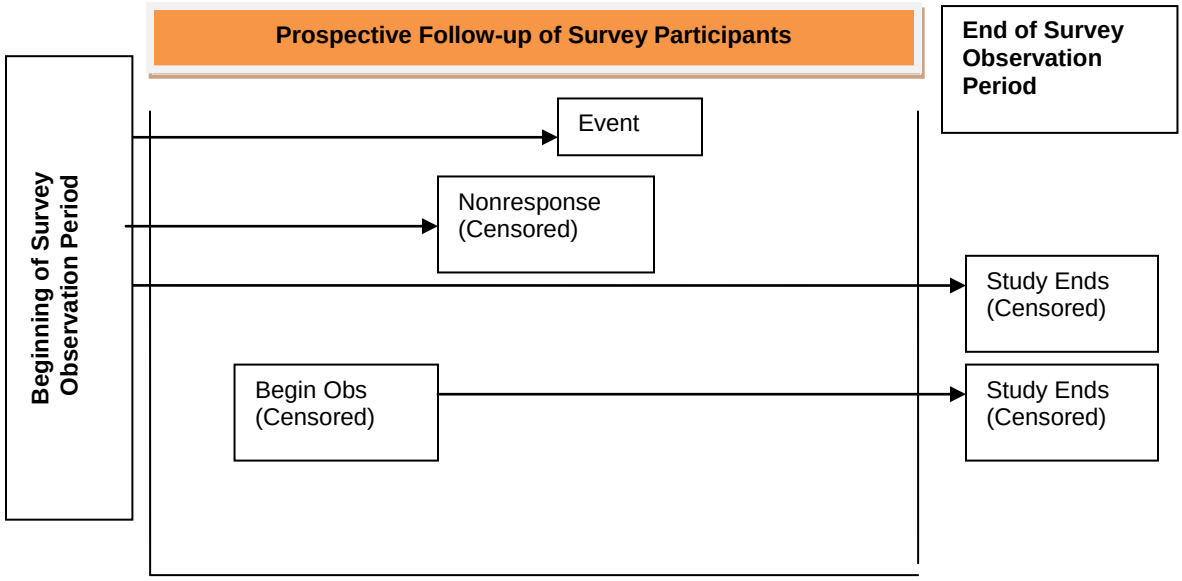
Longitudinal data is prospectively collected on individuals followed over time. One example is the Panel Study for Income Dynamics, an ongoing US panel study focused on income dynamics and related topics (http://psidonline.isr.umich.edu/).

Administrative follow-up data comes from a study that collects administrative records and additional survey data for a sample of respondents and then prospectively follows those individuals to a key event such as death by linking to another data source. An example of this type of data might be a medical claims data set that is linked to a mortality data set using respondent Social Security Numbers. The linked files would provide an opportunity to study time to death using a survival analysis approach. An example of this type data is the NHANES III linked mortality file (http://cdc.gov/nchs/data/datalinkage).

The third category is retrospective event history data where respondents are asked to recall details about an event of interest which occurred at some point in the past. An example of this type of data is the National Comorbidity Survey-Replication survey (http://www.hcp.med.harvard.edu/ncs/) which contains retrospective data on mental illness and related physical conditions.

FEATURES OF SURVIVAL ANALYSIS

Survival analysis centers on analysis of time to an event of interest, denoted as (T), given the event occurred, or time to censoring, denoted as (C). If an individual is right censored, the respondent does not experience the event of interest before follow-up ends and it is unknown if the event occurs after censoring. Left censoring means that follow-up began after the beginning of data collection. See Figure 1 for a graphic presentation of the common types of timelines. Time and censoring are key pieces of information used in statistical analysis of event history data.



(From *Applied Survey Data Analysis,* p. 306)

Figure 1. Prospective View of Event History Survey Data

Time can be regarded as continuous or discrete and this basic distinction affects the analytic approach selected. For example, an analysis of the time in milliseconds to the event of interest (i.e. particle explosion) would be handled using a continuous time assumption while an analysis of age of onset of alcohol abuse measured in 2 year increments is a discrete time approach since age is measured in coarse time units.

DEFINITIONS

Key definitions used in survival analysis are presented in this section. Probability density functions, cumulative distribution functions and the hazard function are central to the analytic techniques presented in this paper. For statistical details, please refer to the SAS/STAT *Introduction to Survival Analysis Procedures* or a general text on survival analysis (Hosmer et al., 2008).

The probability density function for the event time is denoted by $f(t)$, and is defined as the probability of the event at time t (for continuous time), or by π_m, denoting the probability of failure in the interval $(m, m+1)$ for discrete time.

The corresponding cumulative density functions are defined in the standard fashion:

$$F(t) = \int_0^t f(t)dt \text{ for continuous } t; \text{ or}$$

$$F(m) = \sum_{k \le m} \pi(k) \text{ for } t \text{ measured in discrete intervals of time.}$$

The CDFs for survival time measure the probability that the event occurs at or before time t (continuous) or before the close of time period m (for discrete time).

The survivor function or survivorship function, $S(t)$, is the complement to the CDF and is defined as follows:

$$S(t) = 1 - P(T \le t) = 1 - F(t) \text{ for continuous time; or}$$
$$S(m) = 1 - F(m)$$

The value of the survivor function for an individual is the probability that the event has not yet occurred at time *t* (continuous) or prior to the close of observation period *m* (discrete time).

The concept of a hazard or hazard function plays an important role in the interpretation of survival analysis models. A hazard is essentially a conditional probability. For continuous time models, the hazard is $h(t) = f(t)/S(t)$ or the conditional probability that the event will occur at time *t* given that it has not occurred prior to time *t*. In discrete time models, this same conditional probability takes the form $h(m) = \pi(m)/S(m)$ (Heeringa, West and Berglund, 2010).

SURVIVAL ANALYSIS MODELS

Analytic models for survival analysis can be categorized into four general types: 1. parametric models 2. nonparametric models, 3. semi-parametric models and 4. discrete time. Analysis examples of all but the parametric model technique are presented in this paper. This is primarily due to the lack of a SURVEY procedure to estimate parametric models in the current version of SAS.

Parametric models assume an underlying distribution for the probability function. For example, a common type of parametric model is the exponential distribution. As previously noted, these models are not yet programmed in a SAS SURVEY procedure and thus, are omitted from this presentation. For simple random sample data, however, use of the LIFEREG procedure is appropriate. See the SAS/STAT documentation for details.

Nonparametric models include no assumptions regarding the probability density function and use observed data to describe survivor functions and hazards. Although there are limitations to PROC LIFETEST regarding the incorporation of complex sample adjusted variance estimation and integer weights, this procedure still has merit for descriptive analysis and tests of the proportional hazards assumption. Use of PROC LIFETEST to compute Kaplan-Meier estimates and survival/failure curves is presented in Example 1.

Semi-parametric models do not have strong assumptions about the underlying probability function but do include an assumption of proportional hazards among model covariates. The proportional hazard assumption can be evaluated through examination of survival curves or by use of model diagnostics where available. Use of PROC SURVEYPHREG to fit a Cox model with sample survey data is demonstrated and discussed in Example 2.

Models such as the logit and complementary log-log are popular choices for discrete time survival analysis. Key features of this type of analysis are a properly structured data set with multiple records per respondent, appropriate model links to define the model, and design corrected variance estimates and hypothesis tests, all available via data step programming and PROC SURVEYLOGISTIC. Use of PROC SURVEYLOGISTIC to fit a discrete time logistic model with complex sample data is presented in Example 3.

OVERVIEW OF COMPLEX SAMPLE DATA

The analyst faced with the task of performing survival analysis with complex survey data must consider some basic issues and questions. What changes when analyzing complex sample data instead of simple random sample data? What SAS procedures are appropriate for the analysis at hand? How does SAS incorporate the complex sample information and correctly calculate the statistics?

In short, variance estimates and hypothesis tests (and associated degrees of freedom) require incorporation of the design features and probability weights for correct estimation. This can be accomplished in SAS via use of the SURVEY procedures in general, and for survival analysis via PROC SURVEYPHREG and PROC SURVEYLOGISTIC. For more information on complex sample data analysis, see the SAS "*Introduction to Survey Sampling and Analysis Procedures"* of the SAS/STAT documentation or a text such as *Applied Survey Data Analysis* (Heeringa, West and Berglund, 2010).

All analysis examples presented in this paper use data from the National Comorbidity Survey-Replication, a public release, nationally representative sample based on a stratified, multi-stage area probability sample of the United States population (Kessler et al, 2004 and Heeringa, 1996). The NCS-R data set provides variables that allow

analysts to incorporate the complex survey design into variance estimation computations. Use of the **sestrat** (strata) and **seclustr** (Sampling Error Computing Unit or cluster) variables along with the probability weights (**ncsrwtsh, ncsrwtlg)** is required. All NCS-R analyses should account for the complex sample and be correctly weighted in order to produce statistically correct variance estimates. Additional references on this topic are Kish (1965), and Rust, (1985).

The event of interest in each of the upcoming examples is first onset of Major Depressive Episode, a DSM-IV diagnosis. Data indicating a diagnosis of MDE and the associated age of onset of MDE were obtained from retrospective survey questions in the NCS-R. For those respondents that never had an onset of MDE, the age at interview is used to represent right censoring. Selected demographic variables such as age at interview, gender, and race/ethnicity are included as covariates.

In summary, three analytic approaches are presented: Kaplan-Meier survival/failure curves using PROC LIFETEST and PROC SGPLOT, the Cox Proportional Hazards model using PROC SURVEYPHREG, and the discrete time logistic model using PROC SURVEYLOGISTIC. Additional procedure options such as ODS graphics, estimate and test statements, link options, and ODS output data sets for further analysis are also demonstrated.

ANALYSIS APPLICATIONS

EXAMPLE 1 - KAPLAN-MEIER ESTIMATION OF THE SURVIVOR FUNCTION

The first example demonstrates the use of PROC LIFETEST for estimation of the survivorship (and failure) functions. The event of interest is onset of MDE. A total of four curves are presented in this example. Figures 2 and 3 present the survival and failure curves for time to onset of MDE in the total sample. Figures 4 and 5 present survival and failure curves for onset of MDE stratified by race/ethnicity groups (Asian/Other, Hispanic, Black and White). The stratified curve has a dual function in that it can be used to check for differences in onset curves within race/ethnicity and also a way to visually inspect the proportional hazards assumption (underlying the Cox model).

DATA CONSTRUCTION

Prior to analysis, data step coding is used to create 1. an integer version of the probability weight, **ncsrwtsh**, and 2. a variable representing time to either onset of MDE or age at interview (right censoring if no event occurred), **ageevent.** Note that the probability weight provided in the NCS-R data is non-integer and is multiplied by 100 for use with the **freq** statement of PROC LIFETEST, which only accepts integer weights. Furthermore, the weight used in the examples is the Part 1 weight, see the NCS-R documentation for details on weights. Given that the primary use of PROC LIFETEST is for descriptive analysis only, this presents no particular problem for basic graphing and visual examination. The binary indicator of an MDE diagnosis, **mde,** is coded 1 for a diagnosis of MDE and 0 for all others.

```
data ncsr;
  set ncsr_input;
  ncsrwtsh100=ncsrwtsh*100;
  if mde=1 then AgeEvent=mde_ond;
  else AgeEvent=age;
run;
```

KAPLAN-MEIER ESTIMATES

With the necessary variables constructed, the K-M curve can be estimated. The following code requests a survival plot without a plot symbol indicating censored cases. The default is to include the censored cases in the plot line. Other features used here are: **ods graphics on** for ODS graphics output, the **time** statement, **time ageevent*mde(0),** and the **freq** statement for a weighted analysis.

```
ods graphics on;
proc lifetest data=ncsr plots=(survival(nocensor));
        time AgeEvent*mde(0);
        freq ncsrwtsh100;
run;
```

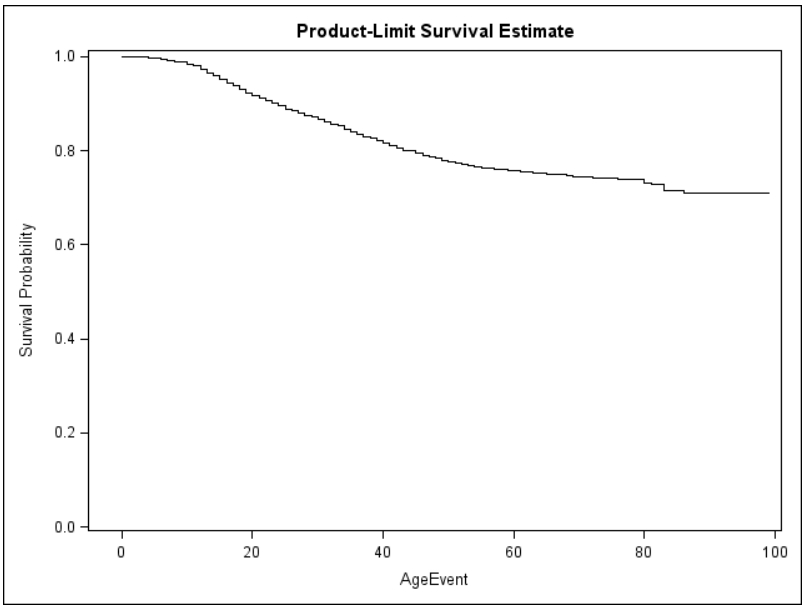


Figure 2. K-M Estimates of the Survivor Functions for MDE

Figure 2 shows the survival curve for onset of MDE. It indicates an overall survival rate of about 72%, meaning that about 72% of the sample "survived" by not having an onset of MDE. Or, about 28% of the sample will eventually have an onset of MDE. In addition, all onsets of MDE have occurred by about age 80 with censoring continuing until age 99 (the oldest age interviewed in this data set). This can be verified by examining the numeric output from PROC LIFETEST (not included here). Other interesting features of this curve show that the steepest section of the curve occurs between about 18 and 50 years of age, showing that the majority of MDE onsets occur between the late teens and middle age.

An alternate presentation is a plot of failure which can be specified in PROC LIFETEST using a slightly different plot request. For example, use of the **failure** (bold red) option in the plots statement will produce a mirror image of the survival curve.

```
ods graphics on;
proc lifetest data=ncsr plots=(survival(failure nocensor));
       time AgeEvent*mde(0);
       freq ncsrwtsh100;
run;
```

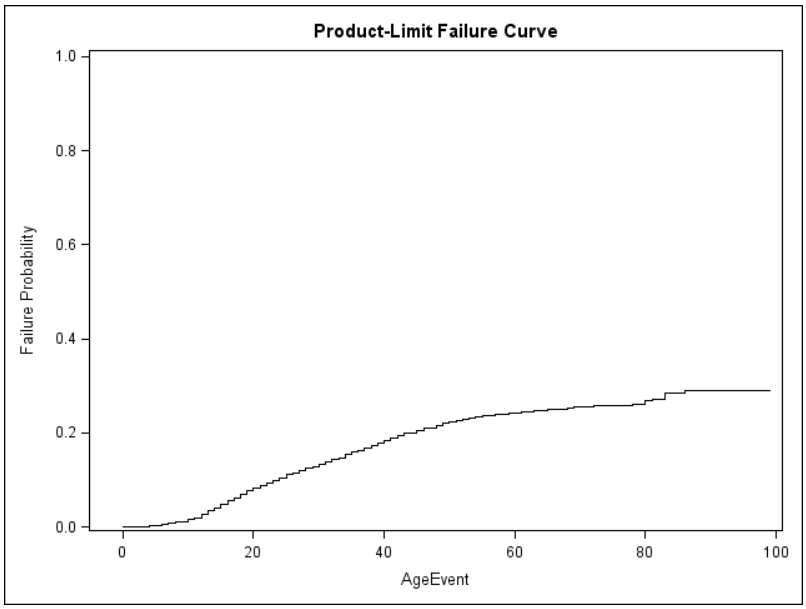


Figure 3. K-M Estimates of the Failure Functions for MDE

Figure 3 illustrates the "mirror image" of the survival curve, the failure curve.

The next section of SAS code produces Figure 4, a stratified analysis of onset of MDE by race/ethnicity groups. It also produces an output data set called **surv_est,** used in subsequent analyses. This curve provides a descriptive graphic of the survival curves for four categories of race/ethnicity: Asians/Other Race, Hispanic, Black and White. Another potential use of this curve is as a visual test of the assumption of proportional hazards, a key assumption for the Cox proportional hazards model.

Though the standard errors and hypothesis tests from PROC LIFETEST are not design corrected and are also distorted due to the inflated weight, the curves still provide important descriptive information such as similarities and differences in survivorship functions.

Use of the **strata racegroup** statement requests a stratified analysis of age to onset of MDE by race/ethnicity. The other parts of this code are similar to previous examples.

.

```
ods output productlimitestimates=surv_est;
proc lifetest data=ncsr plots=(survival(nocensor));
       time AgeEvent*mde(0);
       strata RaceGroup;
       freq ncsrwtsh100;
       format RaceGroup rf.;
run ;
```

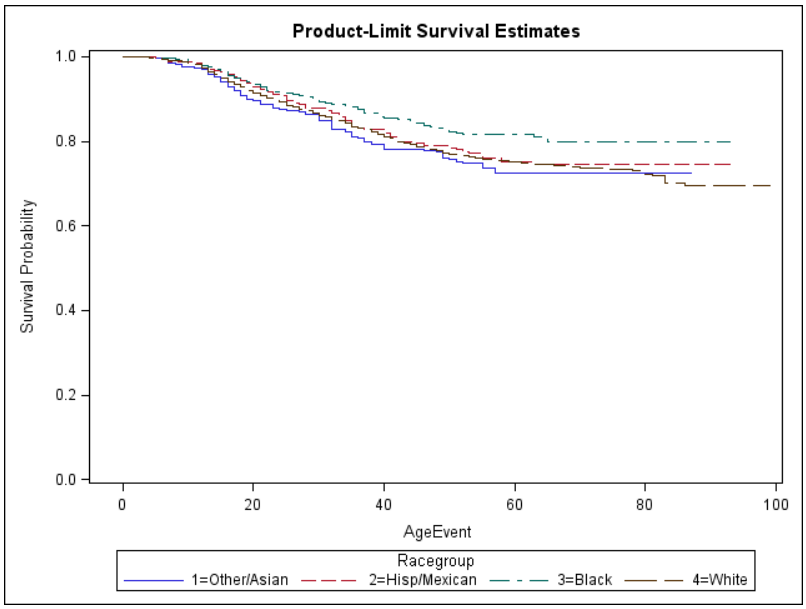


Figure 4. K-M Estimates of the Survivor Functions for MDE by Race/Ethnicity

Figure 4 highlights differences in survival between race/ethnicity groups. For example, Blacks consistently have the lowest number of individuals with onset of MDE (or highest survival proportion) in the life-course. In addition, by about age 80, at least 20% of each group will have an onset of MDE with the majority of onsets occurring between the ages of 18 and 55. The "flat" section on the right tail of each line represents cases that are censored at age of interview. Note that the four lines are roughly parallel, indicating that the proportional hazards assumption underlying the upcoming Cox model is not violated.

FAILURE CURVE WITH OUTPUT SURVIVAL DATA SET AND PROC SGPLOT

The next example presents an alternative method of producing failure plots via use of PROC SGPLOT for statistical graphing. In this example, a failure curve stratified by race/ethnicity is produced from the output data set, **surv_est**, which was created by the **ods output** statement from the previous LIFETEST run. Use of the **series** statement with a **/group=** option requests a plot with four overlaid lines in the graph. The **x** and **y** syntax specifies the variables for the two axes (**ageevent** and **failure)**. Note how the use of a different y axis scale spreads out the graph lines, making interpretation easier.

```
proc sgplot data=surv_est;
       series x=AgeEvent y=failure / group=RaceGroup;
       format RaceGroup rf. ;
       label AgeEvent='MDE Age of Onset' failure='Failure';
       title "Kaplan-Meier Curve Representing Onset of MDE" ;
run ;
```

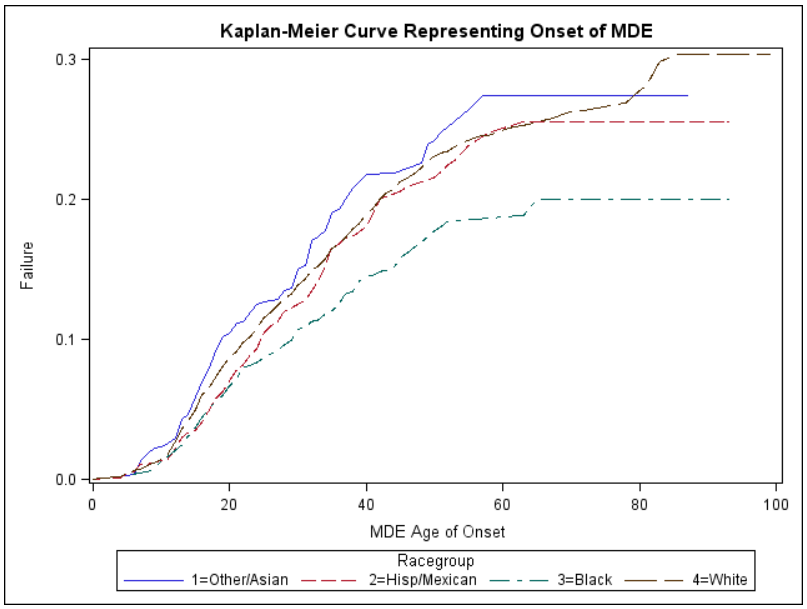


Figure 5. K-M Estimates of the Failure Functions for MDE by Race/Ethnicity

EXAMPLE 2 - THE COX PROPORTIONAL HAZARDS MODEL

The second analysis example demonstrates use of PROC SURVEYPHREG for fitting a Cox Proportional Hazards model (Cox, 1972) using complex sample data. This model is considered semi-parametric and assumes continuous time with proportional hazards among covariates.

For the Cox model approach, the data set is again structured as a one record per person file with the **ageevent** variable representing either the age at onset of MDE or the age at interview, if censored. As a reminder, this is the same dependent variable used in Example 1. Variance estimates and hypothesis tests will be correctly estimated using the default Taylor Series Linearization method of PROC SURVEYPHREG. Because SURVEYPHREG is new in SAS 9.2 and model diagnostics for survey data analysis are relatively undeveloped at this time, a test of the proportional hazards assumption is currently omitted from this procedure. As an alternative test, a log transformed plot from PROC LIFETEST is demonstrated (post model fitting) for evaluation of this assumption.

FITTING THE COX MODEL WITH PROC SURVEYPHREG

Example 2 demonstrates fitting of the Cox model with PROC SURVEYPHREG. The outcome of interest, time to onset of MDE, is predicted by race/ethnicity, age at interview, and gender.

The SURVEYPHREG procedure, by default, outputs the hazard ratio, which is the ratio of the probability that an event will occur at time t, given that it has not yet occurred. This ratio is represented by the following equation:

$$\hat{HR}_j = \frac{h_0(t)\exp(\hat{B}_1 x_1 + \cdots \hat{B}_j(x_j + 1) + \cdots \hat{B}_p(x_p)}{h_0(t)\exp(\hat{B}_1 x_1 + \cdots \hat{B}_j(x_j) + \cdots \hat{B}_p(x_p)}$$
$$= \exp(\hat{B}_j)$$

The formula shows that the hazard ratio for a given predictor represents the multiplier that a one unit change in that predictor will have on the expected hazard. For categorical predictors, the one unit change in a predictor is compared to the omitted reference category.

The following SAS code illustrates use of PROC SURVEYPHREG with **strata**, **cluster** and **weight** statements to identify the complex sample design variables (**sestrat** and **seclustr**) and weight (**ncsrwtsh**).

Additional options include the **class** statement to declare categorical or classification variables, the **(ref=)** option to declare the omitted category used, and the **/param=ref** option for reference group parameterization. The model statement has the **/risklimit** option to obtain 95% confidence limits for the hazard ratios. Note that the dependent variable is specified as the age of the event * an indicator of the event occurring or censor: **ageevent*mde(0)**. Finally, use of the **ods rtf** statement produces output in rich text format, directly readable in Word or other text processing software.

```
ods rtf;
proc surveyphreg data=ncsr;
  weight ncsrwtsh;
  strata sestrat;
  cluster seclustr;
  class racegroup (ref=first) sex (ref=last) / param=ref;
  model AgeEvent*mde(0) = racegroup sex age / risklimit;
run;
ods rtf close;
```

SURVEYPHREG OUTPUT

Table 1.1

Model Information		
Data Set	WORK.NCSR	
Dependent Variable	AgeEvent	
Censoring Variable	Mde	
Censoring Value(s)	0	
Weight Variable	NCSRWTSH	NCSR sample part 1 weight
Stratum Variable	SESTRAT	SAMPLING ERROR STRATUM
Cluster Variable	SECLUSTR	SAMPLING ERROR CLUSTER
Ties Handling	BRESLOW	

Table 1.2

Design Summary	
Number of Strata	42
Number of Clusters	84

Table 1.3

Variance Estimation	
Method	Taylor Series

Table 1.1 provides basic information about the variables used in the analysis including the dependent variable, censoring variable and value, the complex sample variables and weight as well as the method used by default for handling ties (Breslow). For more information, see the SAS/STAT SURVEYPHREG documentation.

Table 1.2 informs about the number of strata and clusters in the NCS-R data set. These variables are used to account for the complex sample design and are used for degrees of freedom calculations.

Table 1.3 states that the default method for complex sample variance estimation, the Taylor Series Linearization method, was used. Other variance estimation options include repeated replication methods: Balanced Repeated Replication (BRR) and Jackknife Repeated Replication (JRR), see the SAS/STAT documentation for details.

Table 2

Summary of the Weighted Number of Event and Censored Values			
Total	**Event**	**Censored**	**Percent Censored**
9282	1779.464	7502.536	80.83

Table 2 provides a table of the weighted number of events that occurred (MDE=1=1779.464) as well as the weighted number of censored cases (MDE=0=7502.536).

Table 3

Testing Global Null Hypothesis: BETA=0				
Test	**Test Statistic**	**Num DF**	**Den DF**	**p-Value**
Likelihood Ratio	785.9019	5	Infty	<.0001
Wald	106.9944	5	42	<.0001

Table 3 presents tests of the global null hypothesis that all of the model predictors are equal to zero. The results are highly significant and the null hypothesis is rejected. Note the use of numerator and denominator degrees of freedom, 5 and 42 indicating 5 degrees of freedom for the model parameters and 42 degrees of freedom used in the denominator, to reflect the complex sample. The degrees of freedom for complex sample data analysis are calculated using the "fixed" rule which is equal to ((number of strata*number of clusters) - number of clusters) or ((42*2)-42)=42 for this sample. For more information on the rule, see *Applied Survey Data Analysis,* p. 62-63 .

Table 4

Analysis of Maximum Likelihood Estimates								
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Hazard Ratio**	**95% Hazard Ratio Confidence Limits**	
Hispanic	42	-0.230892	0.133957	-1.72	0.0921	0.794	0.606	1.040
Black	42	-0.406011	0.150585	-2.70	0.0100	0.666	0.492	0.903
White	42	0.078095	0.119649	0.65	0.5175	1.081	0.849	1.377
Male	42	-0.493506	0.061470	-8.03	<.0001	0.610	0.539	0.691
Age at Interview	42	-0.046409	0.002142	-21.67	<.0001	0.955	0.951	0.959

Omitted groups are Asian/Other and Females.

Table 4 is the Analysis of Maximum Likelihood Estimates table including degrees of freedom, parameter estimates, standard errors and t tests, hazard ratios and 95% hazard ratio confidence limits. Before proceeding with final interpretation of the Cox model results, a test of the proportional hazards assumption for the model considered is recommended. As a reminder, Example 1 illustrated inspection of a survivorship function for parallel lines but an alternate test of this assumption underlying the Cox model is the log transform: ln(-ln(survival probability)) versus ln(t). This type of curve can be produced by PROC LIFETEST and serves as a way to check the assumption of the Cox model.

CHECK OF PROPORTIONAL HAZARDS ASSUMPTION

The two sets of syntax below specify another type of plot request, **plots=(loglogs),** which is the log of the negative log of the survivor functions. This is done for race/ethnicity and sex in separate plots. The remaining parts of the SAS code are identical to the previous LIFETEST examples.

```
ods graphics on;
proc lifetest data=ncsr plots=(loglogs);
       time AgeEvent*mde(0);
       strata racegroup;
       freq ncsrwtsh100;
       format racecat rf.;
run;
```

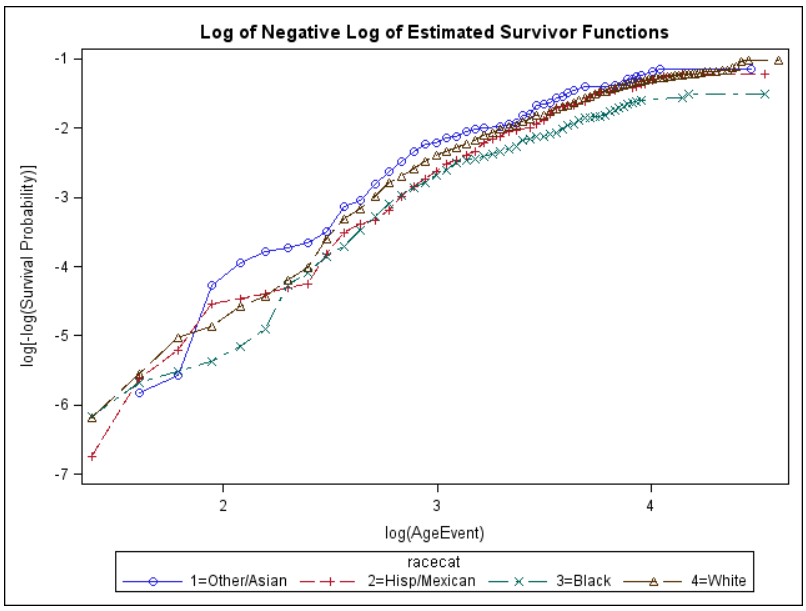


Figure 6. Log-Negative Log of Estimated Survivor Functions by Race/Ethnicity

```
ods graphics on;
proc lifetest data=ncsr plots=(loglogs);
       time AgeEvent*mde(0);
       strata sex;
       freq ncsrwtsh100;
       format sex sexf.;
run;
```

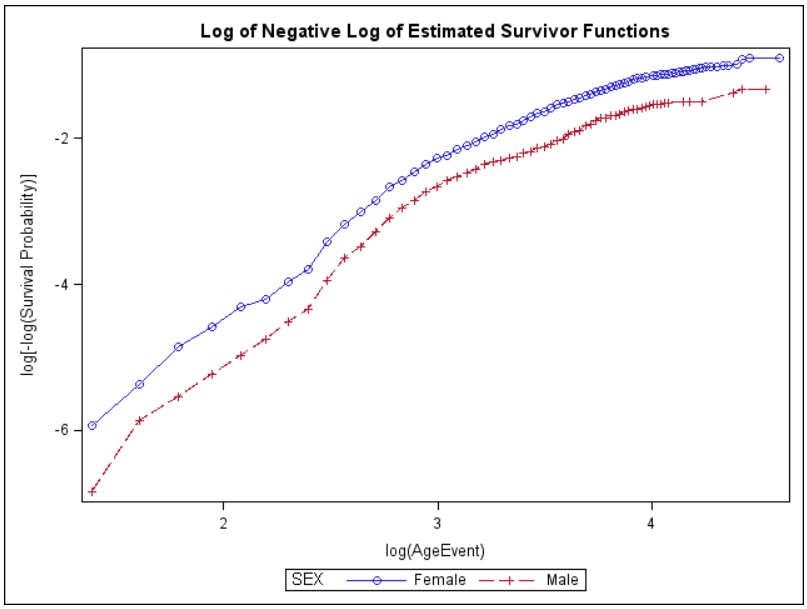


Figure 7. Log-Negative Log of Estimated Survivor Functions by Gender

Figures 6 and 7 show minimal or no evidence of lines crossing and thus, the proportional hazards assumption underlying the Cox model is not violated for these covariates. If that had not been the case and evidence of violation of the assumption was seen, consideration of a time-varying interaction of time and the variable could be considered. For example of this approach, see Allison, *Survival Analysis Using The SAS System, A Practical Guide,* Chapter 5.

With the proportional hazard assumption verified, interpretation of the model results proceeds. Table 4 is repeated here for ease of review.

Table 4

Analysis of Maximum Likelihood Estimates								
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Hazard Ratio**	**95% Hazard Ratio Confidence Limits**	
Hispanic	42	-0.230892	0.133957	-1.72	0.0921	0.794	0.606	1.040
Black	42	-0.406011	0.150585	-2.70	0.0100	0.666	0.492	0.903
White	42	0.078095	0.119649	0.65	0.5175	1.081	0.849	1.377
Male	42	-0.493506	0.061470	-8.03	<.0001	0.610	0.539	0.691
Age at Interview	42	-0.046409	0.002142	-21.67	<.0001	0.955	0.951	0.959

The hazard ratios represent the impact on the expected hazard of a one unit change in a given predictor variable, holding all other covariates constant. For example, being Hispanic (a one unit change in a categorical variable) multiplies the hazard of onset of MDE by .79, compared to the omitted reference group (Asian/Other) and while holding all other predictors constant. In other words, being Hispanic decreases the hazard of having an onset of MDE by about 21%, as compared to Asian/Other, holding all other covariates constant. Similarly, being Black decreases the hazard by 33% while being White increases the hazard by about 8%, both as compared to the Asian/Other group and holding other covariates constant. The hazard ratio for Hispanics is marginally significant and significant for Blacks. For Whites, the slightly elevated hazard ratio is not statistically significant.

Males have statistically significant decreased hazards (39%) of onset of MDE, as compared to Females. Finally, a year increase in Age at Interview results in a statistically significant decreased hazard of MDE onset (4% decrease),

again while holding all other predictors constant. These results are consistent with the survival curves examined in the Example 2.

HYPOTHESIS TESTS

Hypothesis tests for model covariates can be done with the **estimate** statement. This optional statement will calculate a design corrected F test of the null hypothesis that all of the specified parameters in the estimate statement are equal to zero. In this example, two **estimate** statements are specified. The first tests the hypothesis that all of the race/ethnicity parameters from the Cox model are equal to zero and the second tests the hypothesis that the race/ethnicity and gender parameters together are equal to zero.

These tests are done using the **estimate** statement with a series of specifications for each level of the classification variable. The use of the **/ joint** option requests a joint parameter test. Note that the **estimate** statements must be placed after the model statement as they are calculated using the model results. See the SAS/STAT PROC SURVEYPHREG documentation for more details on this topic.

```
proc surveyphreg data=ncsr;
  weight ncsrwtsh;
  strata sestrat;
  cluster seclustr;
  class racegroup (ref=first) sex (ref=last) / param=ref;
  model AgeEvent*mde(0) = racegroup sex age / risklimit;
  estimate 'Race' racegroup 1 0 0, racegroup 0 1 0, racegroup 0 0 1 / joint;
  estimate 'Race and Gender' racegroup 1 0 0, racegroup 0 1 0, racegroup 0 0 1,
sex 1 /joint;
run;
```

Table 4.1

F Test for Estimates				
Label	Num DF	Den DF	F Value	Pr > F
Race	3	42	11.17	<.0001

Table 4.2

F Test for Estimates				
Label	Num DF	Den DF	F Value	Pr > F
Race and Gender	4	42	28.28	<.0001

Tables 4.1 and 4.2 show that race/ethnicity (F value=11.17 with 3 df, p value=<.0001) and race/ethnicity and gender (F value=28.28, 4 df, p value=<.001) are significantly different from zero and thus, the null hypothesis is rejected in both tests.

Additional interesting features of PROC SURVEYPHREG not presented here are domain analysis for subpopulations, inclusion of time-varying covariates, use of the **nomcar** option for handling missing data, interactions between covariates, and use of other variance estimation methods such as JRR and BRR. See the SAS/STAT SURVEYPHREG documentation for details.

EXAMPLE 3 - DISCRETE TIME SURVIVAL ANALYSIS

Survival models for discrete time are needed when time is measured in a non-continuous or "coarsened" manner. Given a correctly structured data set with properly created variables, executing a discrete time logit model using complex survey data is easily accomplished using PROC SURVEYLOGISTIC. As with all SAS SURVEY procedures, PROC SURVEYLOGISTIC incorporates the complex sample design through use of the **strata**, **cluster** and **weight** statements and correctly calculates variance estimates and associated hypothesis tests.

THE DISCRETE TIME LOGISTIC MODEL

The choice of the discrete time logit model for modeling hazard functions with discrete outcomes is common. Many excellent references are available to the analyst interested in more details, see Allison, (1995) or Singer and Willett, (1993). The definition of the logit model is as follows:

$$\ln\left(\frac{h_{i,m}}{1-h_{i,m}}\right) = B_{0,m} + \boldsymbol{x}_{i,m}\boldsymbol{B}$$
$$= B_{0,m} + B_1 x_{1m} + \cdots + B_p x_{pm}$$

where $h_{i,m}$ refers to the hazard of failure at discrete time m for respondent i, $B_{0,m}$ is a time-specific intercept term that applies to all individuals at time m, $\boldsymbol{x}_{i,m}$ is a row vector of values on covariates (possibly time-varying) for respondent i, and $\boldsymbol{B}$ is the vector of regression parameters.

The model above defines the logit of the individual hazard and the transformation to obtain the estimated hazard is the inverse logit transformation(Heeringa, West and Berglund, 2010):

$$\hat{h}_{i,m} = \frac{\exp(\hat{B}_{0,m} + \boldsymbol{x}_{i,m}\hat{\boldsymbol{B}})}{1+\exp(\hat{B}_{0,m} + \boldsymbol{x}_{i,m}\hat{\boldsymbol{B}})}$$

DATA STRUCTURE FOR DISCRETE TIME ANALYSIS

The data set used in the discrete time model should be structured to include multiple records for each discrete time unit up to and including the person-time record of the event of interest or censoring. This data structure directly models the effect of time and also permits the use survey data regression tools provided by PROC SURVEYLOGISTIC.

To clarify the structure required, consider two types of respondents: the first a person who experienced the event of interest at a certain age of life and the second who never had the event of interest and was right censored at the age of interview. In this hypothetical example, person one had the event of interest at age 5 and would contribute 5 records from ranging from 1 to 5 to the data set. Person two was interviewed at age 50, would have 50 person-time records, ranging from 1 to 50. This person is considered right censored at age 50. Furthermore, the data set should include a binary indicator of event occurrence with a 1 indicating the person-time record in which the event of interest occurred and a zero in every other person-time record in the file. This indicator serves as a time-varying indicator of the event occurrence.

An additional consideration for complex sample survey data is that each person-time record should include the weight(s), strata and cluster variables along with the key variables described above. The hypothetical example in Figure 8 uses a time-invariant weight but some data sets, i.e. longitudinal surveys, often include time-varying weights and these should be allowed to vary in the discrete time data set. Example 3, however, uses the NCS-R data set which has time-invariant weights.

ID Variable	Indicator of Event	Age of Event	Binary Indicator of Time of Event	Person Time Units	Weight	Strata Code	Cluster Code
120	0	3	0	1	0.90	2	1
120	0	3	0	2	0.90	2	1
120	**0**	**3**	**0**	**3**	**0.90**	**2**	**1**
121	1	5	0	1	0.76	1	2
121	1	5	0	2	0.76	1	2
121	1	5	0	3	0.76	1	2
121	1	5	0	4	0.76	1	2
121	**1**	**5**	**1**	**5**	**0.76**	**1**	**2**

Figure 8. Data Structure for Discrete Time Survival Models

Figure 8 presents a hypothetical data structure including person-time records for two individuals: the first, ID=120, who never experienced the event of interest and was censored at person-time unit 3 and the second, ID=121, who did experience the event of interest at person-time unit 5. Therefore, ID=120 has 3 records all with a zero for the binary indicator of time of event and ID=121 with 5 records with a zero for the binary indicator for records 1-4 and a 1 for record number 5 (in bold red). This figure also includes the complex design variables and weight to illustrate how they are repeated for each person-time record in the discrete time data set.

APPLICATION OF THE DISCRETE TIME LOGIT MODEL

Example 3 demonstrates the entire discrete time logistic modeling process including data preparation, model estimation and evaluation of results. The dependent variable used throughout this paper, onset of MDE (measured in age in years), is again predicted by age at interview, gender and race/ethnicity but now modeled as a discrete time logistic model. Use of the same model as in Example 2 allows a side-by-side comparison of the Cox model and discrete time logistic model results.

DATA PREPARATION

The initial step is preparation of a person-year data set for use in the upcoming discrete time survival analysis. The NCS-R data includes a variable with age of onset of MDE and thus, time is measured in years (considered discrete in this example). As previously explained, a data set with multiple person-year records per respondent is required. For those that experienced the event of interest, age of onset of MDE is used and records ranging from age 1 to the onset of MDE are included while for those without onset of MDE, records range from age 1 to age at interview.

The following SAS code produces an output data set that meets the specified requirements. Use of the **do** loop with an **output** statement outputs one record for each person-year ranging from age 1 to age of the event (either the event age of onset or age at interview/censor).

As a reminder, **ageevent**, was created for the survival curve analysis in Example 1 and was used as the dependent variable for the Cox model in Example 2. The purpose of this data step is to "expand" the data set from the one record per person structure to multiple records per person, representing each person's time in years to the event or censoring.

```
data personyear;
  set ncsr;
  do personyear=1 to AgeEvent;
    output;
  end;
run;
```

The second data step reads in the data set called **personyear** and creates a key analysis variable, **mdetv**, which is a binary indicator with a 1 indicating the year in which onset of MDE occurred (obtained from the variable **mde_ond**), while every other person-year is assigned to zero. This simple coding approach creates the time-varying dependent variable, **mdetv,** to be used in the logistic model.

```
data ncsrpersonyear;
  set personyear;
  if personyear=mde_ond then mdetv=1 ; else mdetv=0 ;
run ;
```

As a double check of the NCS-R data structure and variable creation, Figure 9 includes a printout of the two types of person-year records, a set of records for a respondent who did have MDE (Caseid=3800) and another set for Caseid=56, a respondent without MDE and censored at age 20 (partial set of records). Note that the only person year with a 1 is the year of onset of MDE: person-year=8 for CaseID=3800. (SAS code not shown here).

CaseID	MDE	Age of Event	MDE_TV	Person Year
3800	1	8	0	1
3800	1	8	0	2
3800	1	8	0	3
3800	1	8	0	4
3800	1	8	0	5
3800	1	8	0	6
3800	1	8	0	7
3800	**1**	**8**	**1**	**8**
56	0	20	0	1
56	0	20	0	2
56	0	20	0	3
56	0	20	0	4-19
56	**0**	**20**	**0**	**20**

Figure 9. Sample Records from the NCS-R Person-Year Data Set

DISCRETE TIME LOGISTIC REGRESSION

With data preparation complete, discrete time logistic modeling can now proceed. Use of PROC SURVEYLOGISTIC with the default **link=logit** for logistic regression is presented. Note some features of the syntax used for this analysis: the familiar **strata/cluster/weight** statements represent the complex sample design and the **class** statement which specifies classification variables along with custom reference categories and reference parameterization. The **model** statement uses the **mdetv (event='1')= ,** syntax which requests that SAS model the probability of **mdetv=1**. The **link=logit** model option is not required as this is the default but is included for clarity. Use of the **test** statement provides a joint test of the race/ethnicity and gender variables together. This is the equivalent of the second **estimate** statement presented in PROC SURVEYPHREG. Note that the default output from PROC SURVEYLOGISTIC includes a Wald test for each classification variable, i.e. for **racegroup** with 3 df.

Example 3 specifies the same model as used in Example 2 with one key exception. The variable, **personyear,** which represents person-time is the additional covariate included in the logistic model. This is due to the way time is used to define the data structure for discrete time modeling.

The form (continuous, categorical, or indicator) of the person-year variable is important. In this example, **personyear** is used as a linear predictor. One advantage of this approach is that the model is more parsimonious but if a non-

linear relationship is suspected, a categorical/indicator version of person-year could be considered. For a discussion of use of categorical/indicator person-time variables and inclusion/exclusion of the intercept, see Allison, (1995).

```
proc surveylogistic data=ncsrpersonyear;
strata sestrat;
cluster seclustr;
weight ncsrwtsh;
class racegroup (ref=first) sex (ref=last) / param=ref;
model mdetv (event='1') =racegroup sex personyear age / link=logit;
testrace_sex: test racegroup2, racegroup3, racegroup4, sex1;
run;
```

SURVEYLOGISTIC OUTPUT

Table 5.1

Response Profile			
Ordered Value	**Mdetv**	**Total Frequency**	**Total Weight**
1	0	383867	385086.58
2	1	1829	1779.46

Table 5.2

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	973.9629	6	<.0001
Score	860.7735	6	<.0001
Wald	699.3262	6	<.0001

Table 5.3

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Race Group	3	32.8935	<.0001
Sex	1	64.5315	<.0001
Person Year	1	244.9812	<.0001
Age at Interview	1	575.7922	<.0001

Tables 5.1-5.3 include the weighted MDE response profile, the Global Null Hypothesis tests and the Analysis of the Effects tables. This set of tables indicates that there are 1779.46 weighted respondents with MDE, the test of the null hypothesis that all of the predictors in the model are equal to zero is rejected, and that race/ethnicity, gender, and age at interview, as groups of effects, are significant predictors of MDE onset.

Table 6.1

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.4877	0.1346	671.4694	<.0001
Hispanic	1	-0.2454	0.1335	3.3779	0.0661
Black	1	-0.4105	0.1503	7.4563	0.0063
White	1	0.0716	0.1203	0.3538	0.5520
Male	1	-0.4927	0.0613	64.5315	<.0001
Person Year	1	0.0326	0.00208	244.9812	<.0001
Age at Interview	1	-0.0538	0.00224	575.7922	<.0001

Table 6.2

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Hispanic	0.782	0.602	1.016
Black	0.663	0.494	0.891
White	1.074	0.849	1.360
Male	0.611	0.542	0.689
Person Year	1.033	1.029	1.037
Age at Interview	0.948	0.943	0.952

Tables 6.1 and 6.2 present the Maximum Likelihood Estimates and Odds Ratio Estimates for the discrete time logistic model. Interpretation of Table 6.2 is that Hispanics have (1-.78) or 22% lower odds of having a first onset of MDE, as compared to the omitted Asian/Other group, while holding all other covariates including person-year constant. Blacks have 34% lower odds of first onset of MDE and Whites have 7% higher odds of onset of MDE, both compared to Asian/Other and holding all other variables constant. Males have 39% lower odds of first onset of MDE, as compared to females and a one year increase in Age at Interview results in a 5% reduction in the odds of first onset of MDE, again holding all covariates constant. Blacks, Males, and Age at Interview are all statistically significant predictors of onset of MDE while Hispanics are marginally significant.

Table 6.3

Linear Hypotheses Testing Results				
	Label	**Wald Chi-Square**	**DF**	**Pr > ChiSq**
	Race/Ethnicity, Sex	112.3411	4	<.0001

Table 6.3 includes the requested linear hypothesis test of the null hypothesis that race/ethnicity and sex are equal to zero in the discrete time logistic model. This hypothesis is rejected given the highly significant p value, <.0001.

COMPARISON OF THE COX MODEL AND THE DISCRETE TIME LOGISTIC MODEL

Table 7 presents a comparison of the discrete time logistic model (Example 3) and the Cox model (Example 2) results. Odds Ratios and Hazard Ratios with the associated 95% Confidence Limits are included.

Table 7

	Discrete Time Logistic Odds Ratio Estimates			Cox Model		
Effect	**Point Estimate**	**95% Wald Confidence Limits**		**Hazard Ratios**	**95% Hazard Ratio Confidence Limits**	
Hispanic	0.782	0.602	1.016	0.794	0.606	1.040
Black	0.663	0.494	0.891	0.666	0.492	0.903
White	1.074	0.849	1.360	1.081	0.849	1.377
Male	0.611	0.542	0.689	0.610	0.539	0.691
Person Year	1.033	1.029	1.037	--	--	--
Age at Interview	0.948	0.943	0.952	0.955	0.951	0.959

Table 7 shows very similar results overall with the same predictors (Blacks, Males and Age at Interview) being significant in both models. This finding suggests that the two approaches, which differ in how time is handled and underlying model assumptions, result in similar estimates of the impact of key socio-demographic variables on first onset of MDE at any age during the life course studied.

Other useful features of PROC SURVEYLOGISTIC not presented here include the **clog-log link** for the complementary log-log model, domain analysis for subpopulations, use of the **nomcar** option for missing data, and repeated replication for variance estimation. See the SAS/STAT PROC SURVEYLOGISTIC documentation for details.

CONCLUSION

Survival analysis using data derived from complex sample surveys can be carried out using PROC LIFETEST, PROC SURVEYPHREG, and PROC SURVEYLOGISTIC. These procedures cover nearly all common event history descriptive and modeling techniques (with the exception of parametric survival models for complex sample data) while incorporating the design features and weights for correct variance estimates and hypothesis tests.

REFERENCES

Allison, P.D., *Survival Analysis Using the SAS System*: *A Practical Guide*, SAS Institute, Cary, NC, SAS Institute, Inc., 1995.

Berglund, P. (2008) "Getting the Most out of the SAS® Survey Procedures: Repeated Replication Methods, Subpopulation Analysis, and Missing Data Options in SAS® v9.2", SAS Global Forum 2008.

Cox, D.R., Regression models and life tables, *Journal of the Royal Statistical Society-B*, 34, 187-220, 1972.

Heeringa, S. (1996) "National Comorbidity Survey (NCS): Procedures for Sampling Error Estimation".

Heeringa, S., West, B.T., Berglund, P.A, "Applied Survey Data Analysis", Chapman Hall CRC Press, 2010.

Hosmer, D.W., Lemeshow, S., and May, S., *Applied Survival Analysis: Regression Modeling of Time to Event Data* (2nd ed.), John Wiley & Sons, Hoboken, NJ, 2008.

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

Kessler, R.C., Berglund, P., Chiu, W.T., Demler, O., Heeringa, S., Hiripi, E., Jin, R., Pennell, B-E., Walters, E.E., Zaslavsky, A., Zheng, H. (2004). The US National Comorbidity Survey Replication (NCS-R): Design and field procedures. *The International Journal of Methods in Psychiatric Research*, 13(2), 69-92.

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

Rust, K. (1985). Variance Estimation for Complex Estimation in Sample Surveys. Journal of Official Statistics, Vol 1, 381-397. (CP)

Singer, J.D. and Willett, J.B., It's about time: Using discrete-time survival analysis to study duration and the timing of events, Journal of Educational and Behavioral Statistics*, 18, 155-195, 1993.*

CONTACT INFORMATION

Patricia Berglund
Institute for Social Research
University of Michigan
426 Thompson St.
Ann Arbor, MI 48106
pberg@umich.edu