Paper 2040-2015

Survival Analysis with Survey Data

Joseph C. Gardiner, Division of Biostatistics, Department of Epidemiology and Biostatistics,

Michigan State University, East Lansing, MI 48824

Abstract

Surveys are designed to elicit information on population characteristics. A survey design typically combines stratification and multistage sampling of intact clusters, sub-clusters and individual units with specified probabilities of selection. A survey sample can produce valid and reliable estimates of population parameters at a fraction of the cost of carrying out a census of the entire population, with clear logistical efficiencies. For analyses of survey data SAS software provides a suite of procedures from SURVEYMEANS and SURVEYFREQ for generating descriptive statistics and conducting inference on means and proportions to regression-based analysis through SURVEYREG and SURVEYLOGISTIC. For longitudinal surveys and follow-up studies SURVEYPHREG is designed to incorporate aspects of the survey design for analysis of time-to-event outcomes based on the Cox proportional hazards model, allowing for time-varying explanatory variables. We review the salient features of the SURVEYPHREG procedure with application to survey data from the National Health and Nutrition Examination Survey (NHANES III) Linked Mortality File.

1. Introduction

Statistical analyses are based on samples drawn from a target population with the objective of making inference on characteristics of the population. Often the samples are assumed to drawn independently with equal probability of selection. The population of units is assumed to be infinite. Survey sampling comprises the design by which samples are selected from a finite population. The design may include stratification of the population, drawing sample of clusters at the first stage followed by subsequent stages of selection of subunits, with different probabilities of selection. A complex survey design refers to process of drawing multistage samples that include stratification, clustering and unequal probabilities of selection. The selection probabilities yield sampling weights that are used strategically in analyses. Using standard notation, survey data are represented by $\{(Y_{hij}, w_{hij}, \mathbf{x}_{hij}) : 1 \le h \le H, 1 \le i \le n_h, 1 \le j \le m_{hi}\}$ where Y_{hij} is the outcome variable of the *j*-th unit in the *i*-th cluster, in the *h*-th stratum. There are n_h clusters with m_{hi} units (usually individuals) in the cluster. The cluster is the primary sampling unit (PSU), $\mathbf{x}_{hij}$ are the covariates and w_{hij} is the sampling weight.

Several national studies use survey methodology to obtain information on health and nutrition, healthcare utilization and expenditures. They include the National Health and Nutrition Examination Survey (NHANES), Nationwide (renamed National in 2012) Inpatient Sample (NIS) of the Healthcare Utilization Project (HCUP) and Medical Expenditure Survey (MEPS). NHANES is conducted periodically in the non-institutionalized civilian US population to obtain estimates of the prevalence of diseases and risk factors for infectious diseases, monitor trends in risk behaviors, environmental exposures, diet and nutrition.

2. NHANES-III

For this article we assembled a data set from public use data files of NHANES-III conducted during 1988-1994 in adults aged ≥17 years (n=20,026). Mortality status of survey participants as of December 31, 2006 was ascertained from death certificates and linkage to files in the National Death Index (NDI). There are 49 strata and 2 PSUs per stratum assembled from the four-stage sampling procedure for selection of individuals from households in geographic areas (counties, city blocks, etc.). Details of the survey design can be found on the Centers for Disease Control & Prevention (CDC) website. Several subgroups were oversampled to increase the reliability and precision of estimates of health status indicators for these population subgroups. Of importance is the sampling weight assigned to each individual which reflects the (inverse) probability of selection. The sampling weight of a sampled person is a measure of the number of people in the population represented by that individual. It must be used in analyses to obtain unbiased population estimates and standard errors.

In the application to be presented later we use a small set of variables age, sex, smoking and diabetes status. Individuals with diabetes were identified by self-report, but women who had only gestational diabetes were not counted as cases. We excluded cases of diabetes where age of onset was <30 years whether or not they were insulin-dependent. Finally, we restricted the data set to individuals who were at least 40 years of age at interview which resulted in a sample of 11,337individuals. To protect confidentiality, survey participants of age 90 or older (n=195) are recorded in a single category (90+).

In survival analysis we use person years of follow-up from interview date to death (all-cause) with vital status determined from the NDI. A person is presumed to be alive, that is, right censored if the date of death was not in the NDI. There were 2,368 deaths with 461 among diabetics. Cause of death is not considered in our analyses. A boarder assessment of vital status is also available through linkage to files in the Social Security Administration and Centers for Medicare and Medicaid Services. There are several articles that have used NHANES-III for mortality analyses. Reddigan et al (2012), Brown et al (2013) and Doran et al (2014) are three recent papers reporting use of SAS software in their analyses. For a neat overview of SAS survey procedures see Berglund, 2011.

3. Regression Methods for Survey Data

Denote by $\mathbf{z}_l = (Y_l, \mathbf{x}_l)$ the outcome and covariate data of an individual $l \equiv (h, i, j)$ in the population and let $q(\mathbf{z}_l, \theta)$ be a function with parameters $\theta \in \Theta$. Our objective is to estimate θ via optimization of $E\big(q(\mathbf{z}_l, \theta)\big)$, that is $\hat{\theta} = \arg\min\big\{E\big(q(\mathbf{z}_l, \theta)\big) : \theta \in \Theta\big\}$. The selected sample from the population is described by a binary selection indictor s_l with selection probability $p_l(\mathbf{v}_l) = P[s_l = 1 \mid \mathbf{v}_l, \mathbf{z}_l]$ where $\mathbf{v}_l$ are variables employed in the selection. These are usually components of $\mathbf{x}_l$ and stratum indicators. From $E\big(s_l p_l^{-1} q(\mathbf{z}_l, \theta)\big) = E\big(q(\mathbf{z}_l, \theta) p_l^{-1} E(s_l \mid \mathbf{v}_l, \mathbf{z}_l)\big) = E\big(q(\mathbf{z}_l, \theta)\big)$ the estimator $\hat{\theta}$ if it exists is obtained as a solution the empirical estimating equations $\sum_{l=1}^{n} w_l \nabla_\theta q(\mathbf{z}_l, \hat{\theta}) = 0$ where n is the total sample size, $w_l = p_l^{-1}$ with indices re-numbered to retain only the selected observations (Wooldridge, 2010).

3.1 Linear regression

For linear regression, $q(\mathbf{z}_l, \beta) = \frac{1}{2}(Y_l - \mathbf{x}_l'\beta)^2$ where β is a $p\times 1$ vector of regression coefficients, including an intercept. The estimating equations are $\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\mathbf{x}_{hij}(Y_{hij} - \mathbf{x}_{hij}'\hat{\beta}) = 0$ giving the weighted least squares (WLS) estimator $\hat{\beta} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}$ where $\mathbf{Y}$ is the vector of all responses, $\mathbf{X}$ the matrix of covariates and $\mathbf{W}$ the diagonal matrix of sampling weights. The sample size is $n = \sum_{h=1}^{H}\sum_{i=1}^{n_h} m_{hi}$. Any of the SAS procedures for linear regression that has a weight statement, for example PROC REG, PROC GLM would serve to obtain $\hat{\beta}$. However, these procedures do not incorporate the features of stratification and clustering of the survey design in estimation of the variance matrix $\mathbf{V} = Var(\hat{\beta})$. PROC SURVEYREG computes the estimator $\hat{\mathbf{V}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{G}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$ by Taylor series linearization where the $p\times p$ matrix $\mathbf{G}$ is created from the residuals $r_{hij} = Y_{hij} - \mathbf{x}_{hij}'\hat{\beta}$ as

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

$\mathbf{e}_{hij} = w_{hij}r_{hij}\mathbf{x}_{hij}$, $\mathbf{e}_{hi} = \sum_{j=1}^{m_{hi}}\mathbf{e}_{hij}$, $\bar{\mathbf{e}}_h = \frac{1}{n_h}\sum_{i=1}^{n_h}\mathbf{e}_{hi}$. In survey procedures the factor $\frac{n-1}{n-p}$ is always included in the variance computation. It can be suppressed by the option VADJUST=none in the model statement. The sampling fraction $f_h = n_h / N_h$ in stratum h will be ignored unless information is supplied by the TOTAL= or RATE= options.

3.2 Logistic regression

A similar strategy is adopted by PROC SURVEYLOGISTIC for logistic regression. For a binary response the objective function in maximum likelihood estimation is $q(\mathbf{z}_l, \beta) = -\log L(\mathbf{z}_l, \beta)$ where $\log L(\mathbf{z}_l, \beta) = \sum_{l=1}^{n}\left(Y_l \log F(\mathbf{x}_l'\beta) + (1-Y_l)\log(1-F(\mathbf{x}_l'\beta))\right)$, $F(u) = (1+e^{-u})^{-1}$, $u \in (-\infty, \infty)$ is the logistic function. Incorporating the weights w_l, the MLE $\hat{\beta}$ of β is the solution to the estimating equations $\mathbf{U}(\beta) = \sum_{l=1}^{n}\mathbf{x}_l w_l\left(Y_l - F(\mathbf{x}_l'\beta)\right) = 0.$ Summation is over the triple index $l \equiv (h,i,j)$. The estimated variance of $\hat{\beta}$ is $\hat{\mathbf{V}}(\hat{\beta}) = \hat{\mathbf{H}}^{-1}(\hat{\beta})\mathbf{G}\hat{\mathbf{H}}^{-1}(\hat{\beta})$ where $\hat{\mathbf{H}}(\hat{\beta}) = \sum_{l=1}^{n} w_l\left(F(\mathbf{x}_l'\hat{\beta})(1-F(\mathbf{x}_l'\hat{\beta})\right)\mathbf{x}_l\mathbf{x}_l'$ and the matrix $\mathbf{G}$ is created from the residuals $r_{hij} = Y_{hij} - F(\mathbf{x}_{hij}'\hat{\beta})$ with exactly the same form as for linear regression. We may use PROC GENMOD or PROC LOGISTIC with the appropriate weight statement to obtain the Hessian $\hat{\mathbf{H}}(\hat{\beta})$ but not $\mathbf{G}$. PROC SURVEYLOGISTIC computes $\hat{\mathbf{V}}(\hat{\beta})$ which is used subsequently in estimation and hypothesis tests of linear functions $\mathbf{L}\beta$.

3.3 Proportional hazards regression: Time to event data

With right-censored data we observe $X = \min(T,U)$ and censoring indicator δ where $X = T$ is the survival (event) time if observed $(\delta = 1)$, or $X = U$ the follow-up time if censored $(\delta = 0)$. The data for an individual $l \equiv (h,i,j)$ at time t is comprised of the event indicator $N_l(t) = [T_l \le t, \delta_l = 1]$, the at-risk indicator $Y_l(t) = [X_l \ge t]$ and perhaps time-dependent covariates $\mathbf{z}_l(t)$ observed at a finite set of time points on $[0, X_l]$. Note the change in notation from the previous section with use of $Y_l(t)$.

The Cox-regression model for the hazard posits $h(t \mid \mathbf{z}(t)) = h_0(t)\exp\left(\mathbf{z}'(t)\beta\right)$ where the baseline hazard function $h_0(t)$ is unspecified. Given a random sample $\{(X_l, \delta_l, \mathbf{z}_l(t)) : 1 \le l \le n\}$ estimation of β is based on maximum partial likelihood: the Breslow likelihood has the form

$L(\beta) = \prod_{l=1}^{n}\prod_{t}\left\{\frac{Y_l(t)\exp(\mathbf{z}_l'(t)\beta)}{\sum_{\ell=1}^{n} Y_\ell(t)\exp(\mathbf{z}_\ell'(t)\beta)}\right\}^{\Delta N_l(t)}$. Since the likelihood requires evaluation only at event times, let $t_1 < t_2 < \ldots < t_K$ denote the be the K distinct survival times in the sample

- d_k = number of deaths at time t_k; D_k = the set of individuals who fail at t_k
- n_k = number at risk at time t_k; R_k = the set of individuals who are at risk at t_k.

The Breslow likelihood is rewritten as $L(\beta) = \prod_{k=1}^{K}\frac{\exp\left(\left(\sum_{l\in D_k}\mathbf{z}_i'(t_k)\right)\beta\right)}{\left\{\sum_{l\in R_k}\exp\left(\mathbf{z}_l'(t_k)\beta\right)\right\}^{d_k}}$ which can be derived from conditional probability arguments (Klein & Moeschberger, 2005, Lawless, 2003). Essentially, from the selected risk set R_k one selects the set D_k who fail. Hence with survey data the sampling weight w_{hij} will enter quite naturally in the construction. Recall that (h,i,j) refers to the *j*-th subject in the *i*-th cluster in the *h*-th stratum. Compress the notation as follows: $(h,i,j) \in A$ means the subject (h,i,j) belongs to the set A. For survey data we use the partial likelihood (Binder, 1992, Lin, 2000, Boudreau & Lawless, 2006):

$L(\beta) = \prod_{k=1}^{K}\frac{\exp\left(\left(\sum_{(h,i,j)\in D_k} w_{hij}\mathbf{z}_{hij}'(t_k)\right)\beta\right)}{\left\{\sum_{(h,i,j)\in\in R_k}\exp\left(w_{hij}\mathbf{z}_{hij}'(t_k)\beta\right)\right\}^{d_k}}$, where $d_k = \sum_{(h,i,j)\in D_k} w_{hij}$ is the 'estimate' of the count of events at time t_k. This is exactly the same as incorporating weights into the usual Breslow likelihood with the weight statement in PROC PHREG. The corresponding Hessian is also the same $\hat{\mathbf{H}}(\hat{\beta}) = -\frac{\partial^2 \log L(\hat{\beta})}{\partial\hat{\beta}\partial\hat{\beta}'}$ but the survey model-based variance of $\hat{\beta}$ is computed differently and not as $\hat{\mathbf{H}}^{-1}(\hat{\beta})$. Instead they are obtained as $\hat{\mathbf{V}} = \hat{\mathbf{H}}^{-1}(\hat{\beta})\mathbf{G}\hat{\mathbf{H}}^{-1}(\hat{\beta})$ where $\mathbf{G}$ is derived as described for linear and logistic regression

except that the score residuals vectors $\{\mathbf{L}_{hij}\}$ are used in $\mathbf{e}_{hij} = w_{hij}\mathbf{L}_{hij}$. For sake of completeness the expressions are below with SAS names in parentheses. Summations over the selected sample A are indicated by $\sum_{hij\in A} \bullet$. Let $\bar{\mathbf{Z}}(t,\hat{\beta}) = \dfrac{\sum_A w_{hij} Y_{hij}(t)\mathbf{z}_{hij}(t)\exp(\mathbf{z}'_{hij}(t)\beta)}{\sum_A w_{hij} Y_{hij}(t)\exp(\mathbf{z}'_{hij}(t)\beta)} \equiv \dfrac{S^{(1)}(t,\hat{\beta})}{S^{(0)}(t,\hat{\beta})}$. The Schoenfeld residuals (RESSCH) are $\mathbf{U}_{hij}(t) = \mathbf{z}_{hij}(t) - \bar{\mathbf{Z}}(t,\hat{\beta})$ and the score residuals (RESSCO) are

$$\mathbf{L}_{hij} = \delta_{hij}\mathbf{U}_{hij}(X_{hij}) - \sum_{l\in A} \frac{w_l Y_{hij}(X_l)\exp(\mathbf{z}'_{hij}(X_l)\beta)}{S^{(0)}(X_l;\hat{\beta})}\left\{\delta_l \mathbf{U}_{hij}(X_l)\right\} .$$

4. Application

We use the NHANES III data described in section (2) for survival analysis of person years (PERYRS) with sex, smoking and diabetes status as covariates. STATUS is the 0/1 right censoring indicator, with STATUS=0 for censoring. The following formats will be applied throughout to better display output; they are condensed in the macro string

```
%let covfmt=%str(format sex sex. smoker smoke. diabetes diabetes.;);

proc format;
value sex 1='male' 2='female';
value smoke 0='never' 1='current' 2='former';
value diabetes 1='yes' 2='no';
run;
```

Of importance are the STRATA, CLUSTER and WEIGHT statements describing the survey design. The variable names are those in the public use files of NHANES III.

```
%let svydsn=%str(strata SDPSTRA6;
cluster SDPPSU6;
weight wtpfqx6;);
```

The syntax to estimate the proportional hazards model with SEX, SMOKER and DIABETES as covariates is below with some output shown in Table 1.

```
proc surveyphreg data=SURV_SGF varmethod=taylor;
 &svydsn &covfmt
class sex(ref='male')smoker(ref='never')
           diabetes(ref='no')/param=ref ;
model PERYRS*STATUS(0)=sex smoker diabetes/rl ties=breslow
vadjust=none covb invhess;
run;
```

Summary information in the output shows that there are 49 strata and 98 clusters (2 PSU's per stratum). The 11,324 individuals in the data set represent an estimated 94.5 million individuals in the population.

Table 1: Analysis of Maximum Likelihood Estimates								
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Hazard Ratio**	**95% Hazard Ratio Confidence Limits**	
sex female	49	–0.132283	0.062507	–2.12	0.0394	0.876	0.773	0.993
smoker current	49	0.082340	0.085525	0.96	0.3404	1.086	0.914	1.289
smoker former	49	0.186373	0.079518	2.34	0.0232	1.205	1.027	1.414
diabetes yes	49	1.060252	0.091167	11.63	<.0001	2.887	2.404	3.468

Confidence intervals for hazards ratios (HR) are constructed from an approximate *t*-distribution for the log-hazard β-parameters. The degrees of freedom are v =#clusters $(\sum_{h=1}^{H} n_h)$ –#strata (H). Diabetes prevalence is associated with a higher mortality risk, HR=2.89 (95% CI: 2.40, 3.47). Females have lower risk than males (p-value=.039).

There are p=4 β-parameters in the model. The standard chi-square test computes, $Q_W = \hat{\beta}'\hat{\mathbf{V}}^{-1}\hat{\beta}$ with null distribution $Q_W \sim \chi^2(p)$. The estimated covariance matrix $\hat{\mathbf{V}} = \hat{\mathbf{H}}^{-1}(\hat{\beta})\mathbf{G}\hat{\mathbf{H}}^{-1}(\hat{\beta})$ of $\hat{\beta}$ is displayed by the **covb** option. The results are obtained under the DF=none option. The default option computes an F-statistic $F_{ADJ} = (v-p+1)Q_W / vp$. The null distribution is $F_{ADJ} \sim F(p,v)$, but under the DF=PARMADJ option, $F_{ADJ} \sim F(p, v-p+1)$. In general to test the hypothesis $H_0 : \mathbf{L}\beta = 0$ where $\mathbf{L}$ is a full row rank $r \times p$ matrix, the statistic is $F = (\mathbf{L}\hat{\beta})'\left(\mathbf{L}\hat{\mathbf{V}}\mathbf{L}'\right)^{-1}\mathbf{L}\hat{\beta}) / r$ with null distribution $F(r,v)$. For example, to test the association of smoking status with mortality use

```
estimate 'Smoker current' smoker 1,
         'Smoker former' smoker 0 1/joint;
```

The F(2, 49) test is not significant (p-value=.070) at the 5% level.

The **invhess** option in the MODEL statement displays $\hat{\mathbf{H}}^{-1}(\hat{\beta})$. This is the same as the estimated covariance matrix from PROC PHREG with the same sampling weight but ignoring the stratification and clustering features of the survey design. We also get the same parameter estimates but their standard errors are grossly underestimated. PROC PHREG also provides a robust (sandwich) covariance matrix using aggregated residuals in the calculation of $\mathbf{G} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\mathbf{e}_{hi}\mathbf{e}'_{hi}$. Use an ID statement with the stratum and cluster variables.

```
proc phreg data=SURV_SGF covs(aggregate);
id SDPSTRA6 SDPPSU6;
weight wtpfqx6;
```

However with the survey design $\mathbf{G} = \frac{n-1}{n-p}\sum_{h=1}^{H}(1-f_h)\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(\mathbf{e}_{hi}-\bar{\mathbf{e}}_h)(\mathbf{e}_{hi}-\bar{\mathbf{e}}_h)'$. The robust variance from PHREG does not demean the $\mathbf{e}_{hi}$ in the computation of **G.**

Table 2 displays the ratio of standard errors from the survey design relative to those obtained from the covariance matrix $\hat{\mathbf{H}}^{-1}(\hat{\beta})$; and next ignoring stratification, clustering and dropping the sampling fraction, the covariance matrix is $\hat{\mathbf{V}} = \hat{\mathbf{H}}^{-1}(\hat{\beta})\mathbf{G}\hat{\mathbf{H}}^{-1}(\hat{\beta})$ with $\mathbf{G} = \{n/(n-1)\}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}\mathbf{e}_{hij}\mathbf{e}'_{hij}$. Other than for the outside multiplier this is the same as the robust standard errors from COVS option in PROC PHREG. The **SERATIO=all** option in the model statement will produce the results.

	Table 2: Analysis of Maximum Likelihood Estimates						
Parameter	**DF**	**Estimate**	**STDERR**	$diag\{\hat{\mathbf{H}}^{-1}(\hat{\beta})\}^{1/2}$	**ROBUST STDERR†**	**SERATIO Model**	**SERATIO IND**
sex female	49	–0.132283	0.062507	0.0005938	0.06140	105.260	0.945
smoker current	49	0.082340	0.085525	0.0007708	0.08461	110.957	0.981
smoker former	49	0.186373	0.079518	0.0006766	0.08230	117.528	1.078
diabetes yes	49	1.060252	0.091167	0.0007406	0.08836	123.096	1.081

† From PHREG with **covs(aggregate) and id SDPSTRA6 SDPPSU6;**

4.1 Scaled weights or NLOPTIONS

The sampling weights in this application range from 220 to 120,000. When several observations have large weights, instabilities in estimation might occur. Often this is noticeable when partial maximum likelihood parameter estimates are reported with DF=0. An easy fix is to rescale the weights, or modify the optimization using NLOPTIONS. Consider our previous hazard model with an additional covariate DUR for duration of disease (diabetes). DUR (in years) is continuous and only relevant for individuals with diabetes. We might rescale the sampling weight to **anlwgt=wtpfqx6/120000;** The scaled weight must be in the data set, and not created via programming statements within SURVEYPHREG. The preferred alternative uses the original weights and specifies quasi-Newton for optimization. The parameter estimates from model statement below are in Table 3.

```
model PERYRS*STATUS(0)=sex smoker diabetes diabetes*dur/rl
ties=breslow vadjust=none;
nloptions tech=quanew;
dur=dur-5;
```

Table 3: Analysis of Maximum Likelihood Estimates (SURVEYPHREG)								
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Hazard Ratio**	**95% Hazard Ratio Confidence Limits**	
sex female	49	–0.129331	0.062227	–2.08	0.0429	0.879	0.775	0.996
smoker current	49	0.098398	0.087890	1.12	0.2684	1.103	0.925	1.317
smoker former	49	0.195686	0.081922	2.39	0.0208	1.216	1.032	1.434
diabetes yes	49	0.897529	0.108983	8.24	<.0001	2.454	1.971	3.054
dur*diabetes yes	49	0.034061	0.007533	4.52	<.0001	1.035	1.019	1.050

One unit increase in DUR of disease among diabetics has hazard ratio 1.035 (95% CI: 1.019, 1.050).

For interpretation DUR is centered at 5. Uncannily, but not altogether surprising are the results from PHREG with robust standard errors in Table 4. The aforementioned HR and CI are computed by

```
hazardratio dur/diff=ref cl=wald;
```

Table 4: Analysis of Maximum Likelihood Estimates (PHREG)									
Parameter		**DF**	**Parameter Estimate**	**Standard Error**	**Chi–Square**	**Pr > ChiSq**	**Hazard Ratio**	**95% Hazard Ratio Confidence Limits**	
sex	**female**	1	–0.12933	0.06144	4.4309	0.0353	0.879	0.779	0.991
smoker	**current**	1	0.09840	0.08631	1.2998	0.2542	1.103	0.932	1.307
smoker	**former**	1	0.19568	0.08349	5.4930	0.0191	1.216	1.033	1.432
diabetes	**yes**	1	0.89753	0.10217	77.1674	<.0001	2.454	2.008	2.998
dur*diabetes	**yes**	1	0.03412	0.00730	21.7741	<.0001	1.035	1.020	1.050

4.2 Domain Analysis

In survey data analyses it is very common to conduct analyses of outcomes in subgroups. These subgroups might not be related to the original survey design. Formation of a domain would typically lead to a random sample size for the domain. A domain variable is always categorical. It must not be declared in the CLASS statement or present in the MODEL, STRATA or CLUSTER statements.

For illustration we carry out a domain analysis of the hazard model with the covariates SEX, SMOKER, and DIABETES in subpopulations defined by marital status (MARSTAT). There are 4 levels: Divorced includes Separated. A few observations (n=49) are dropped because of missing information.

```
value marstat4x 1,3='Married/Cohabit' 2,5,6='Divorced'
                7='Never married' 4='Widowed';
```

For a domain D, let $I_D(h,i,j) = [(h,i,j) \in D]$ denote the indicator for membership of unit (h, i, j). The sampling weights are now $v_{hij} = w_{hij} I_D(h,i,j)$. The domain analysis correctly estimates the parameters of the hazard model at each category of marital status from the full data set. Analyzing the subgroups by creating subsets through BY or WHERE statements is incorrect because the subsets might not have the original information on f_h, n_h to calculate the component **G** of the covariance matrix. Worse still some strata and clusters could get eliminated by sub-setting.

```
ods output parameterestimates=parms_d;
proc surveyphreg data=SURV_SGF varmethod=taylor;
&svydsn &covfmt
class sex(ref='male') smoker(ref='never')
                  diabetes(ref='no')/param=ref ;
format marstat marstat4x.;
model PERYRS*STATUS(0)=sex smoker diabetes/rl ties=breslow
vadjust=none;
domain marstat;
run;
```

The DOMAIN statement triggers estimation of the hazard model at each level of MARSTAT. By default results from the overall model are reproduced. `ods output parameterestimates=parms_d;` saves the results in a data set. The *t*- distribution has DF=49.

Table 5: Analysis of Maximum Likelihood Estimates (Domain Analysis)

Parameter	Estimate	StdErr	tValue	Probt	HR	Lower 95% CL	Upper 95% CL	Domain
sex female	–0.133699	0.063256	–2.11	0.0397	0.875	0.770	0.993	
smoker current	0.078040	0.087894	0.89	0.3789	1.081	0.906	1.290	
smoker former	0.183688	0.079133	2.32	0.0245	1.202	1.025	1.409	
diabetes yes	1.059486	0.091540	11.57	<.0001	2.885	2.400	3.468	
sex female	–0.540110	0.178928	–3.02	0.0040	0.583	0.407	0.835	Divorced
smoker current	0.634078	0.249873	2.54	0.0144	1.885	1.141	3.115	Divorced
smoker former	0.405296	0.266812	1.52	0.1352	1.500	0.877	2.564	Divorced
diabetes yes	0.881969	0.279827	3.15	0.0028	2.416	1.377	4.239	Divorced
sex female	–0.387090	0.079741	–4.85	<.0001	0.679	0.578	0.797	Married/Cohabit
smoker current	0.298151	0.142495	2.09	0.0416	1.347	1.012	1.794	Married/Cohabit
smoker former	0.406308	0.104401	3.89	0.0003	1.501	1.217	1.852	Married/Cohabit
diabetes yes	1.203025	0.119300	10.08	<.0001	3.330	2.620	4.232	Married/Cohabit
sex female	–0.136197	0.366365	–0.37	0.7117	0.873	0.418	1.822	Never married
smoker current	–0.100243	0.353589	–0.28	0.7780	0.905	0.444	1.841	Never married
smoker former	–0.136284	0.454831	–0.30	0.7657	0.873	0.350	2.177	Never married
diabetes yes	1.276625	0.376285	3.39	0.0014	3.585	1.683	7.635	Never married
sex female	–0.713464	0.127268	–5.61	<.0001	0.490	0.379	0.633	Widowed
smoker current	–0.429608	0.157523	–2.73	0.0088	0.651	0.474	0.893	Widowed
smoker former	–0.002637	0.111687	–0.02	0.9813	0.997	0.797	1.248	Widowed
diabetes yes	0.553289	0.124610	4.44	<.0001	1.739	1.354	2.234	Widowed

Consider the category of 'Married/Cohabit'. Three summaries in the output will show that of 11,324 observations that were read, 6803 are in this category with 1137 deaths observed (STATUS=1). They represent 63.46 million and 62.95 deaths. This information could be seen from

```
proc surveyfreq data=SURV_SGF varmethod=taylor;
&svydsn &covfmt
tables marstat*status/row cl nocellpercent;
format marstat marstat4x.;
run;
```

Table 6: MARSTAT by STATUS								
MARSTAT	STATUS	Frequency	Weighted Frequency	95% Confidence Limits for Percent		Row Percent	95% Confidence Limits for Row Percent	
Married/ Cohabit	0	5666	57168446	58.7503	62.6322	90.0805	88.8614	91.2996
	1	1137	6295294	5.8886	7.4779	9.9195	8.7004	11.1386
	Total	6803	63463740	65.6430	69.1059	100.000		
Divorced	0	1280	11068675	10.6370	12.8645	88.5831	86.3890	90.7771
	1	245	1426573	1.2190	1.8099	11.4169	9.2229	13.6110
	Total	1525	12495248	12.1131	14.4173	100.000		
Widowed	0	1430	9153522	8.8202	10.6150	70.5283	66.9084	74.1481
	1	863	3824994	3.4890	4.6324	29.4717	25.8519	33.0916
	Total	2293	12978515	12.7516	14.8049	100.000		
Never married	0	543	4628132	4.1678	5.6588	88.0197	84.3878	91.6515
	1	111	629933	0.4559	0.8816	11.9803	8.3485	15.6122
	Total	654	5258066	4.7949	6.3693	100.000		
Total	0	8919	82018774	85.7632	88.3825			
	1	2356	12176795	11.6175	14.2368			
	Total	11275	94195569					
Frequency Missing = 49								

Consider the category of 'Never Married'. It has a relatively small number of subjects. If we carry out a subgroup analysis by sub-setting with a WHERE clause

```
proc surveyphreg data=SURV_SGF(where=(marstat=7));
```

we will find that two strata have a single cluster (PSU). A warning is issued to this effect in the SAS-log with elimination of these strata in variance estimation. The SAS-log includes a note recommending the use of a DOMAIN statement. In Table 7 notice DF=47 and different standard errors instead of the correct (DF=49) results in Table 5.

Table 7: Analysis of Maximum Likelihood Estimates (Subset 'Never-married')								
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio	95% Hazard Ratio Confidence Limits	
sex female	47	–0.136197	0.365940	–0.37	0.7114	0.873	0.418	1.822
smoker current	47	–0.100243	0.352898	–0.28	0.7776	0.905	0.445	1.840
smoker former	47	–0.136284	0.454748	–0.30	0.7657	0.873	0.350	2.178
diabetes yes	47	1.276625	0.368877	3.46	0.0012	3.585	1.707	7.529

4.3 Estimation of the Survival Curve

After fitting a proportional hazards model we might want to display survival curves estimated from the model. For a specified fixed covariate profile **z** we have $S(t\,|\,\mathbf{z}) = \exp(-H_0(t)\exp(\mathbf{z}'\beta))$ where $H_0(t)$ is the baseline cumulative hazard. Therefore from estimators $\hat{\beta}$ and $\hat{H}_0(t)$ we obtain the survival estimates $\hat{S}(t\,|\,\mathbf{z}) = \exp(-\hat{H}_0(t)\exp(\mathbf{z}'\hat{\beta}))$. The BASELINE statement in PROC PHREG is set up for calculation of $\hat{S}(t\,|\,\mathbf{z})$ at covariate profiles specified in a COVARIATES= data set (Gardiner, 2010). With weights incorporated, the estimated event count is $\hat{N}(t) = \sum_{A} w_{hij}[T_{hij} \le t, \delta_{hij} = 1]$ and number at risk is $\hat{Y}(t) = \sum_{A} w_{hij}[X_{hij} \ge t]$ with summation over subjects $l \equiv (h,i,j)$. Using notation described in section (3.3), $\hat{H}_0(t) = \sum_{u \le t}\left(\frac{\Delta\hat{N}(u)}{S^{(0)}(u,\hat{\beta})}\right)$ where $S^{(0)}(t,\hat{\beta}) = \sum_{hij} w_{hij} Y_{hij}(t)\exp(\mathbf{z}'_{hij}(t)\beta)$. The estimator $\hat{S}(t\,|\,\mathbf{z})$ is a step function which changes only at failure times across the whole data set, not just at the failure times observed for the profile **z.** When covariates are absent, effectively setting $\mathbf{z} = 0$, we obtain the Nelson-Aalen estimator $\hat{S}(t) = \exp(-\hat{H}_0(t)) = \exp\left(-\sum_{u \le t}\frac{\Delta\hat{N}(u)}{\hat{Y}(u)}\right)$ and Kaplan-Meier estimator $\hat{S}_{KM}(t) = \prod_{u \le t}\left(1 - \frac{\Delta\hat{N}(u)}{\hat{Y}(u)}\right)$ by simple approximation $\exp(-x) \approx 1 - x, 0 \le x < 1$.

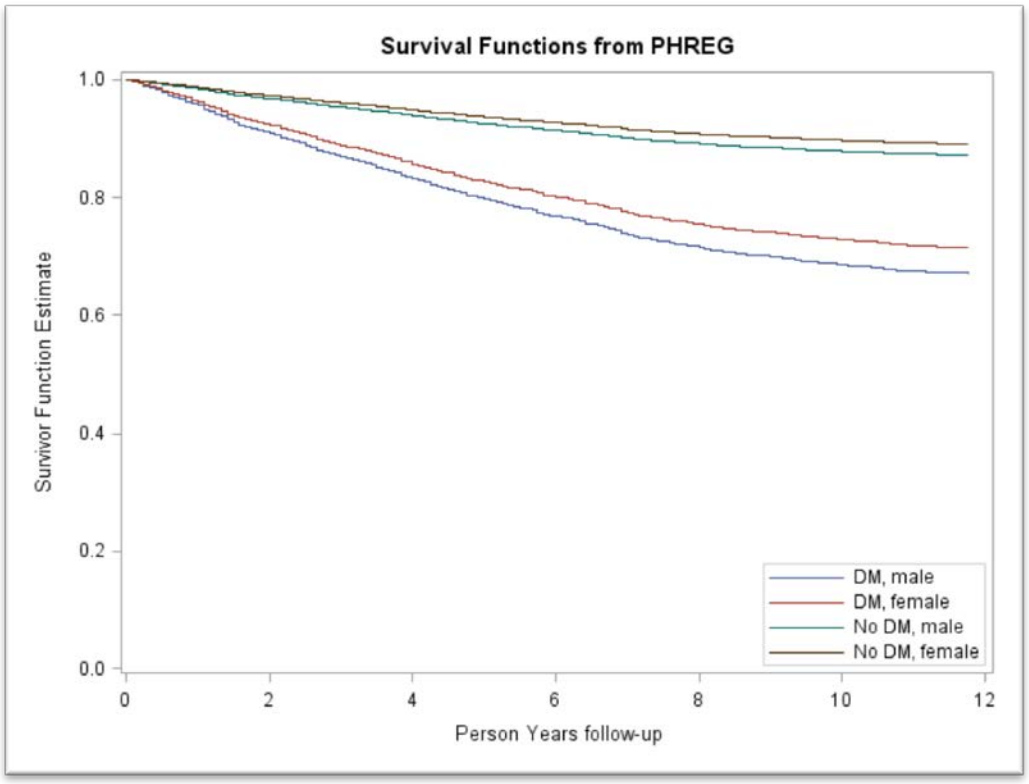
PROC SURVEYPHREG does not support currently the features that PROC PHREG has for survival curve estimation. However, the OUTPUT statement will create a data set with $\hat{Y}(t)$ called WTATRISK and martingale residuals $\hat{M}_{hij}$ (RESMART) defined as $\hat{M}_{hij} = \delta_{hij} - \exp(\mathbf{z}'_{hij}\hat{\beta})\hat{H}_0(t_{hij})$. Calculations are at the observed time $t = t_{hij}$, censored or not, for each record in the data set. Hence either procedure could be used to obtain survival estimates $\hat{S}(t\,|\,\mathbf{z})$. For illustration we define four profiles in the data set COVAR below. PROC SURVEYPHREG would require some extra data steps to get estimates for each covariate profile at all failure times. Hence the invocation of PROC PHREG is simpler.

Obs	ID	sex	diabetes
1	DM, male	male	yes
2	DM, female	female	yes
3	No DM, male	male	no
4	No DM, female	female	no

```
proc phreg data=SURV_SGF; /*plots(overlay)=survival*/
weight wtpfqx6;
class diabetes(ref='no') sex(ref='male')/param=ref ;
&covfmt
model PERYRS*STATUS(0)=diabetes sex/ties=breslow ;
baseline out=survival1 covariates=covar survival=survival/method=ch
rowid=Id;
run;
```

Instead of the default plots, some additional annotations are possible with PROC SGPLOT.

```
proc sgplot data=survival1;
step x=peryrs y=survival/group=ID;
yaxis values=(0 to 1 by 0.2);
label peryrs='Person-years follow-up';
title "Survival Functions from PHREG";
keylegend /location=inside position=bottomright across=1;
run;
```



So far so good. But, what about standard errors? Could we exploit the analogy with the proportional rates/means model for recurrent events analysis in PROC PHREG? For recurrence data the individual i is the cluster supplying multiple data elements via *time-ordered* recurrence times $\{T_{ij} : 1 \le j \le n_i\}$ observed in say $(0, \tau_i]$ where τ_i is independent of events times. The event count up to time t is $N_i(t)$, the at risk indicator is $Y_i(t) = [\tau_i \ge t]$ and n_i is the number of recurrences. The mean cumulative function (MCF) at a fixed covariate $\mathbf{z}$, $\mu(t \mid \mathbf{z}) = E\left(N_i(t) \mid \mathbf{z}\right) = \int_0^t \rho(u \mid \mathbf{z})du$ is estimated under the proportional rates assumption, $\rho(t \mid \mathbf{z}) = \rho_0(t)\exp(\mathbf{z}'\beta)$. For data on recurrences, time varying covariates are allowed, but should be exogenous (Gardiner, 2014; Cook & Lawless, 2007). The baseline MCF $\mu_0(t) = \int_0^t \rho_0(u)du$ is analogous to the baseline cumulative hazard function $H_0(t)$. The estimator $\hat{\mu}_0(t)$ of $\mu_0(t)$ is the same $\hat{H}_0(t)$ presented here but weights have been incorporated.

With survey data, the index i corresponds to a *cluster* of individuals $j = 1,...,m_{hi}$ (in stratum h). Maintaining the structure of stratified cluster samples from the survey design in PROC SURVEYPHREG would entail some serious modifications of the estimate $Var(\hat{\mu}_0(t)) = \sum_{i=1}^{n} \hat{\Psi}_i^2(t)$,

$$\hat{\Psi}_i(t) = \sum_{j=1}^{n_i} \frac{[T_{ij} \le t, \delta_{ij} = 1]}{S^{(0)}(T_{ij}, \hat{\beta})} - \sum_{l=1}^{n} \sum_{j=1}^{n_l} \frac{Y_i(T_{lj})[T_{lj} \le t, \delta_{lj} = 1]}{\{S^{(0)}(T_{lj}, \hat{\beta})\}^2} \exp(\mathbf{z}_i'(T_{lj})\hat{\beta})$$
$$- \left[\sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{[T_{ij} \le t, \delta_{ij} = 1]}{\{S^{(0)}(T_{ij}, \hat{\beta})\}^2} S^{(1)}(T_{ij}, \hat{\beta}) \right] \hat{\mathbf{H}}^{-1} \left[\int_0^{\tau} \mathbf{U}_i(u) dM_i(u) \right]$$

where $dM_i(u) = dN_i(u) - \exp(\mathbf{z}_i'(u)\hat{\beta}) d\hat{\mu}_0(u)$, and $\mathbf{U}_i(t) = \mathbf{z}_i(t) - \overline{\mathbf{Z}}(t, \hat{\beta})$. It is not possible to time order the event times in the cluster to obtain the structure needed of recurrent event analysis. However, our attempt is premature since future enhancements to PROC SURVEYPHREG would likely incorporate options to estimate survival functions and standard errors. For relevant articles see Lin (2000) and Boudreau & Lawless (2006).

4.4 Testing the proportional hazards assumption

The standard approach of including a defined time-dependent function of the covariate to be evaluated can be applied with programming statements within the SURVEYPHREG procedure. Write the model as $h(t \mid \mathbf{z}) = h_0(t)\exp(\beta \mathbf{z}^0 + \beta_p z_p)$, where z_p is the variable that we wish to investigate for non-proportionality in hazard. The other covariates are denoted by $\mathbf{z}^0$. We replace $\beta_p z_p$ by $\beta_p z_p + \gamma z_p g(t)$ where g is a non-decreasing function of t, usually taken as $g(t) = t - t^*$ or $g(t) = \log(t / t^*)$, t^* being a constant. For example, t^* may be taken as the mean survival time or median survival time. Now assess the significance of γ in the model $h(t \mid \mathbf{z}) = h_0(t)\exp(\beta \mathbf{z}^0)\exp(\beta_p z_p + \gamma z_p g(t))$.

For the model described in Table 1 we evaluate z_p =DIABETES. The syntax for Table 1 is modified as

```
proc surveyphreg data=SURV_SGF varmethod=taylor;
 &svydsn &covfmt
class  diabetes(ref='no') sex(ref='male')
          smoker(ref='never')/param=ref ;
model PERYRS*STATUS(0)=sex smoker diabetes dm_t /rl ties=breslow
vadjust=none;
dm_t=diabetes*Log((peryrs+.1)/13);
run;
```

The scaling at 13 is roughly the median follow up time. The addition of 0.1 avoids evaluating the logarithm of very small values. The test of significance for dm_t has p-value=0.790 based on the t-distribution (DF=49). We conclude that our original model without the time-dependent term is tenable.

4.5 Other variance estimation methods

PROC SURVEYPHREG provides two alternatives to the linearized Taylor series method for estimation of $Var(\hat{\beta})$. Balanced repeated replication (BRR) requires a stratified sample with 2 PSU per stratum. For example, the NHANES III data set has H=49 strata with 2 clusters per stratum. Hence $n_h = 2$ for all $h = 1,\ldots,H$. We obtain a half-sample by selecting one PSU in each stratum—there are 2^H possible choices.

- *Form replicate sample.* Let R be the number of replicates (REPS= option). The default is the smallest multiple of 4 that is >H—which here is 52. Each of the R =52 replicates is formed by choosing the "first" PSU (--label "1") or the "second" PSU(--label "−1") is the stratum: a replicate would look like a string of "1" and "−1" of total length H. The ensemble matrix of R rows and H columns is called the Hadamard matrix
- *A replicate gets a new weight.* The chosen PSU gets 2×original weight; the other PSU gets zero weight and hence will be discarded in the analysis
- *Perform the analysis using the replicate sample.* Get the estimator $\hat{\beta}_r$ for the r-th replicate (half sample), $r = 1,\ldots,R$ using the replicate weights. The BRR variance estimator is $Var(\hat{\beta}_{BRR}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_r - \hat{\beta})(\hat{\beta}_r - \hat{\beta})'$ where $\hat{\beta}$ is the original full sample estimator. The option CENTER=REPLICATES replaces $\hat{\beta}$ by the average of the replicates
- The degrees of freedom associated with $\hat{\beta}_{BRR}$ is H

A modification of BRR is VARMETHOD=BRR(FAY=ε) where 0≤ε<1 modifies the replicate weights as follows: if ε is not specified the default value is 0.5.

(1) if the r-th row h-th column of the Hadamard matrix is 1, the first PSU weight is multiplied by ε and the second PSU weight by 2−ε,

(2) if the r-th row h-th column of the Hadamard matrix is −1, the first PSU weight is multiplied by 2−ε and the second PSU weight by ε. Then

$Var(\hat{\beta}_{BRR}) = \frac{1}{R(1-\varepsilon)^2}\sum_{r=1}^{R}(\hat{\beta}_r - \hat{\beta})(\hat{\beta}_r - \hat{\beta})'$ where $\hat{\beta}$ is the original full sample estimator and the degrees of freedom associated with $\hat{\beta}_{BRR}$ is H.

The Fay modification allows all the PSUs to be used in getting the estimate $\hat{\beta}_r$ from the replicate samples.

PROC SURVEYPHREG provides another option—the jackknife which deletes one PSU at a time from the full sample to obtain R replicates. See SAS/STAT documentation for details and additional options.

In our application there is very little difference in the standard errors of the log(hazard) ratios computed from these methods. Our results are summarized in Table 8.

Table 8: Standard Errors of log(hazard) ratios from VARMETHOD options

Parameter	Estimate	BRR	BRR(FAY=.5)	JACKKNIFE	Taylor	PHREG†
sex female	–0.132283	0.062811	0.062565	0.062526	0.062507	0.06140
smoker current	0.082340	0.086239	0.085638	0.085566	0.085525	0.08461
smoker former	0.186373	0.080523	0.079811	0.079541	0.079518	0.08230
diabetes yes	1.060252	0.090594	0.090703	0.091199	0.091167	0.08836

†From PHREG with weighting, `covs(aggregate) and id SDPSTRA6 SDPPSU6;`

5. Discussion

In this article we described some of the capabilities of PROC SURVEYPHREG to analyze right censored time to event data from complex survey samples. The procedure incorporates the design features of stratification, clustering and unequal probabilities of selection—through sampling weights, to estimate the Cox proportional hazard model. The crux of the analysis is the appropriate estimation of standard errors of the estimator $\hat{\beta}$ of regression coefficients β (log-hazard ratios). The variance matrix of $\hat{\beta}$ has the familiar sandwich form $\hat{\mathbf{V}} = \hat{\mathbf{H}}^{-1}(\hat{\beta})\mathbf{G}\hat{\mathbf{H}}^{-1}(\hat{\beta})$ where $\hat{\mathbf{H}}^{-1}(\hat{\beta})$ is the inverse Hessian matrix from the partial log likelihood. PROC PHREG with a weight statement for the sampling weights estimates $\hat{\mathbf{H}}^{-1}(\hat{\beta})$ but ignores the stratification and clustering features of the survey design. Requesting `covs(aggregate)` with aggregation by the stratum and cluster variables comes very close to, but not fully mimicking the matrix $\mathbf{G}$ in $\hat{\mathbf{V}}$.

Survival curves for specified covariate profiles are not too difficult to obtain using the statistics computed in the OUPUT statement. An alternative is the use the plotting options in PROC PHREG. However, standard errors need another computation because an estimator of the variance of the baseline cumulative hazard is not currently available.

The standard approach of testing non-proportionality in a covariate is by defining a time-dependent function of the covariate through programming statements and including it in the model. If its coefficient in not significant, it gives some assurance that proportionality is tenable. Currently, PROC SURVEYPHREG does not support the counting process style input in the model statement (as does PROC PHREG). This is a simple way of having the updated time-dependent covariate values in the input data set. When only a few updates are required, the programming approach within the procedure is relatively easier and has the same functionality.

In our application we used person-years follow-up as the time metric in the proportional hazards model. Another metric is the subject's age itself which has been advocated for use in longitudinal surveys (Korn et al, 1997, Graubard & Korn, 1999) especially where period and cohort effects might be appreciable. Another point to notice is that our analyses are conditional on individuals being alive at interview. It is useful to have an option to incorporate delayed entry (left truncation) into survey data time to event analyses.

References

Berglund PA. An Overview of Survival Analysis using Complex Sample Data, Paper 338-2011. Paper presented at: SAS Global Forum 2011; Las Vegas, NV.

Binder DA. Fitting Cox's proportional hazards models from survey data. *Biometrika.* 1992;79(1):139-147.

Boudreau C, Lawless JF. Survival analysis based on the proportional hazards model and survey data. *Canadian Journal of Statistics-Revue Canadienne De Statistique.* 2006;34(2):203-216.

Brown RE, Riddell MC, Macpherson AK, et al. All-cause and cardiovascular mortality risk in US adults with and without type 2 diabetes: Influence of physical activity, pharmacological treatment and glycemic control. *Journal of Diabetes and Its Complications.* 2014;28(3):311-315.

Cook RJ, Lawless JF. *The Statistical Analysis of Recurrent Events.* New York, NY: Springer-Verlag; 2007.

Doran B, Guo Y, Xu JF, et al. Prognostic value of fasting versus nonfasting low-density lipoprotein cholesterol levels on long-term mortality insight from the National Health and Nutrition Examination Survey III (NHANES-III). *Circulation.* 2014;130(7):546-553.

Gardiner JC. Survival Analysis: Overview of Parametric, Nonparametric and Semiparametric approaches and New Developments, Paper 252-2010. Paper presented at: SAS Global Forum 2010; Seattle, WA.

Gardiner JC. Regression Analysis of Duration and Severity Data –New Capabilities with SAS Software, Paper 1502-2014. Paper presented at: SAS Global Forum 2014; Washington, DC.

Graubard BI, Korn EL. Predictive margins with survey data. *Biometrics.* 1999;55(2):652-659.

Heeringa SG, West BT, Berglund PA. *Applied Survey Data Analysis.* Boca Raton, FL: Chapman & Hall/CRC Press; 2010.

Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data, 2nd Edition.* New York, NY: Springer-Verlag; 2005.

Korn EL, Graubard BI, Midthune D. Time-to-event analysis of longitudinal follow-up of a survey: Choice of the time-scale. *Am. J. Epidemiol.* 1997;145(1):72-80.

Lawless JF. *Statistical Models and Methods for Lifetime Data, 2nd Edition.* Hoboken: John Wiley & Sons; 2003.

Lin DY. On fitting Cox's proportional hazards models to survey data. *Biometrika.* 2000;87(1):37-47.

Reddigan JI, Riddell MC, Kuk JL. The joint association of physical activity and glycaemic control in predicting cardiovascular death and all-cause mortality in the US population. *Diabetologia.* 2012;55(3):632-635.

Wooldridge JM. *Econometric Analysis of Cross Section and Panel Data, 2nd Edition.* Cambridge, MA: MIT Press; 2010.

ACKNOWLEDGMENTS

We thank Dr. Jean Kerver, PhD for assistance in interpreting the design of NHANES III.

CONTACT INFORMATION

We welcome your comments and questions. Please contact

Joseph C. Gardiner
Division of Biostatistics
Department of Epidemiology and Biostatistics
B629 West Fee Hall
Michigan State University
East Lansing, MI 48824
jgardiner@epi.msu.edu