

Survival Estimates Prediction with Survey Data

Bocheng Jing, Northern California Institute for Research and Education;

L. Grisell Diaz-Ramirez, W. John Boscardin, Alexandra K. Lee, Sei J. Lee, University of California, San Francisco.

ABSTRACT

The SAS® PHREG procedure includes a BASELINE statement that allows users to easily obtain the survival predictions, standard error, and confidence interval from a survival model. In the setting of complex survey design, such as stratification and multistage sampling from clusters, SAS SURVEYPHREG procedure is needed to appropriately incorporate survey weights. However, SURVEYPHREG currently does not support the BASELINE statement. Hence, it is difficult to obtain survival predictions, standard errors, and confidence intervals from survival models derived from complex survey data. In this paper, we explore the threshold behind survival prediction using survey data and provide possible solutions to bypass this obstacle.

INTRODUCTION

Survival predictions and its inference helps clinicians assess risk factors for patients. It is especially helpful to quantify the risk uncertainty when making medical decisions. For example, our team studied the one-year mortality for geriatric patients who have had major surgery including Abdominal Aortic Aneurysm Repair, Coronary Artery Bypass Grafting, and Colectomy. The survival predictions of patients who had functional declines or cognitive impairment could inform future patients considering surgery, providing them with estimates of how likely they would experience functional decline or cognitive impairment if they proceeded with surgery. In the field of population health, cohort and covariates are often collected through complex survey data. Complex survey design creates non-independence and disproportional probability of selection in the sample units. Not counting for the complex survey design in statistical analysis would lead to errors when estimating the variances of parameters. SAS has many useful procedures to analyze survey data, including SURVEYPHREG for survival analysis. However, this procedure does not provide the option for survival prediction in Cox model.

The semi-parametric Cox proportional hazard model is widely used in survival analysis. One fascinating fact about the Cox model is that it includes a random baseline hazard function which is eliminated when we estimate the parameters via the partially likelihood function. Unlike other regression methods, the baseline hazard function becomes a myth for prediction. That is, we could not directly plug in patients' information to draw

survival prediction. Fortunately, Breslow (1973) and Kalbfleisch and Prentice (1980) provided two different methods to evaluate the baseline hazard functions. Breslow estimator analogizes the Nelson-Aalen estimator by utilizing exponential function as baseline cumulative hazard function, while Kalbfleisch-Prentice extends Kaplan-Meier estimator with covariates utilizing discrete hazard model. When specifying the BASELINE statement in SAS PHREG procedure, we can easily obtain survival predictions through either method. Another useful feature of the BASELINE statement is that it prevents the interval approximation from over-shooting $[0, 1]$ bound. SAS applies the asymptotic normality to transform survival estimates to bound the confidence interval bounded under $[0, 1]$, which provides sensible estimates for risk assessment. However, the SURVEYPHREG procedure does not support the BASELINE statement, which increases the difficulty for survival prediction with survey data.

In the following sections, we will explore the reasons why survey design prevents the accurate estimations for survival prediction, standard error, and confidence interval. Even though a theoretical solution was proposed by Boudreau and Lawless (2006), no statistical package has implemented a solution for this. Therefore, we propose two methods to address this issue by using some of the options already available in PHREG and its BASELINE statement to obtain accurate survival prediction estimations with approximate standard errors.

THE THRESHOLD OF THE SURVEY DATA

Although SAS SURVEYPHREG currently does not support the BASELINE statement, we can still calculate the survival estimates prediction after fitting the Cox model. It is the standard error of the predictions that we cannot accurately estimate due to the complex survey design. To explore this issue, let us start with the survey design. The most common survey design is to select samples with stratification, clustering, and unequal probability weights. To not complicate notations, we will use the same notation as SAS does. To build a Cox model with survey data, a subject is denoted with a set of matrices (w, t, Δ, Z) , where

- w =sampling weights
- t =event time
- Δ =event indicator
- Z = covariate matrix

For the survey design, denote h as stratum, $1 \leq h \leq H$; i as cluster, $1 \leq i \leq n_h$; j as the subject index within cluster i of stratum h , $1 \leq j \leq m_{hi}$. Denote t as the survival time, following by

- n =total number of observations in the sample; $n = \sum_{h=1}^H \sum_{i=1}^{n_h} m_{hi}$.
- $Y_{hij}(t) = I(t_{hij} \geq t)$, the indicator for whether subject is at risk at time t_{hij}
- $n_{hij}(t) = I(t_{hij} \leq t)$, the indicator for the number of events before time t_{hij}

When we run the Cox model in SURVEYPHREG, the parameters ($\hat{\beta}$ s) are estimated through weighted Breslow partial likelihood function:

$$L(\beta) = \prod_{i=1}^n \prod_{t \geq 0} \left\{ \frac{Y_i(t) \exp(\beta'_i Z_i(t))}{\sum_{j=1}^n w_j Y_j(t) \exp(\beta' Z_j(t))} \right\}^{w_i \Delta N_i(t)} \quad (1)$$

Therefore, we can incorporate the weight variable into the Breslow's baseline accumulative hazard function, which is required for the survival prediction:

$$\hat{\Lambda}_0(t) = \sum \left\{ \frac{\Delta N(t)}{\sum_{bij} w_{bij} Y_{bij}(t) \exp(\beta' z_{bij}(t))} \right\}$$

The OUTPUT statement can deliver all the required variables to calculate manually survival prediction. Fortunately, the SASPHREG procedure can also incorporate sampling weights into parameter estimation. Since we can obtain the predictions from the BASELINE statement, it is much easier to use PHREG procedure than manual calculation.

On the other hand, estimating the standard error of the prediction and drawing inferences is much more complicated. The crux is to order the event times of each subject within each stratum of clusters. For example, when the event time is at 3rd month, 5th month, 7th month in stratum A, and 2nd month, 4th month, 6th month, and 8th month in stratum B, both stratum A and stratum B belong to cluster 1 or primary sample unit (PSU) 1. When we count the events at 3rd month from stratum A, we must consider estimating the survival from both stratum A and B because 3rd month could occur beyond stratum A of cluster 1 in the population. When we draw standard error estimation for 3rd month survival prediction, we must consider the structure of all the strata and the clusters. Ignoring this structure would lead to a biased standard error estimation. Nevertheless, it is a byzantine process to count clusters, strata and sampling weight for standard error estimation and no universal solution has been applied to resolve this matrix. Boudreau and Lawless (2006) proposed a theoretical solution to this matter, that is, center the covariates of the data at Z^* , and then estimate the parameters and baseline hazard with the centered covariates. We show below the proposed solution described by Gardiner (2015) based upon Boudreau and Lawless's solution for the standard error of the survival prediction with survey design:

$$\begin{aligned} \text{var}(\hat{\mu}(t)) &= \sum_{i=1}^n \left\{ \sum_{j=1}^{n_i} \frac{N_{ij}}{\sum_{j=1}^n w_j Y_j(t) \exp(\beta' Z_j(t))} - \sum_{k=1}^{n_k} \sum_{j=1}^{n_k} \frac{Y_i(T_{kj} \leq t)}{\sum_{k=1}^{n_k} w_j Y_j(t) \exp(\beta' Z_j(t))^2} \right. \\ &\left. \left[\sum_{i=1}^n \sum_{j=1}^{n_i} \frac{Y_i(T_{kj} \leq t) w_i z_i \exp(\beta'_i Z_i(t))}{\sum_{j=1}^{n_i} w_j Y_j(t) \exp(\beta' Z_j(t))^2} \right] \hat{H}^{-1} \left[\int_0^\gamma U_i(\mu) dM_i(\mu) \right] \right\}^2, \end{aligned}$$

where

- μ =mean cumulative survival function.
- γ =time period which is independent of the event times.
- $dM_i(\mu)$ =martingale residuals
- \hat{H}^{-1} = Hessian matrix with weights incorporated.

$U_i(\mu)$ =Schoenfeld residuals.

This is a sophisticated theoretical solution, and no statistical package has adopted this method currently.

METHODS TO BYPASS THE THRESHOLD

We present here a preliminary attempt to estimate the standard errors of survival prediction, which requires further work before widespread adoption. Despite the lack of the BASELINE statement in SURVEYPHREG, this procedure provides us a golden standard for the parameter estimation ($\hat{\beta}$ s and variance of $\hat{\beta}$ s). If we can incorporate a pseudo survey structure that mimics the complex survey design into PHREG procedure while obtaining a close estimation for parameters and variance (compared to SURVEYPHREG), we can then use the BASELINE statement to obtain approximate survival prediction and inferences.

When estimating parameters with survey data, SURVEYPHREG uses weighted partially likelihood function (1) to calculate the maximum likelihood estimate, which is essentially the same as when we specify the WEIGHT statement in PHREG to estimate β s. However, the covariance matrix is estimated completely different. PHREG calculates the model-based covariance matrix of β s as $\left[-\frac{\partial^2 l(\hat{\beta})}{\partial \beta^2}\right]^{-1}$, which is also referred as Hessian matrix, $\hat{H}^{-1}(\hat{\beta})$. Note that the covariance matrix remains the same when WEIGHT statement is specified. SURVEYPHREG uses Taylor Series Linearization for covariance matrix estimation, denote as

$$\hat{V}(\hat{\beta}) = \hat{H}^{-1}(\hat{\beta}) G \hat{H}^{-1}(\hat{\beta}) \quad (2)$$

G is residual matrix, denoted as

$$G = \frac{n-1}{n-p} \sum_{b=1}^H (1 - f_h) \frac{n_b}{n_b-1} (e_{hi} - \bar{e}_h)(e_{hi} - \bar{e}_h)' \quad (3)$$

where

- $e_{hij} = w_{hij} r_{hij} Z_{hij}$
- $r_{hij} = Y_{hij} - Z'_{hij} \hat{\beta}$
- $e_{hi} = \sum_{j=1}^{m_{hi}} e_{hij}$
- $\bar{e}_h = \frac{1}{n_h} \sum_{i=1}^{n_b} e_{hi}$
- $f_h = \frac{n_b}{N_b}$

$\frac{n-1}{n-p}$ is a factor, which is always included in the survey procedure, to reduce bias when sample size is small.

Specifying the WEIGHT statement in PHREG accounts properly for the individual sample weights, which prevents certain groups from oversampling. However, it could potentially underestimate the standard errors. When the weight is much larger (the sample is less representative of the population), the effective sample size could be overestimated, resulting in small standard error and insignificant hypothesis testing results. To avoid this issue, we can apply normalized weight, which takes the selection probability into account. Denote w_{nw} as normalized weight,

$$w_{nw} = \frac{w_i}{\bar{w}}$$

Where $\bar{w} = \frac{\sum w_i}{n}$, the mean of sampling weights. By using normalized weight, we will obtain more accurate standard error estimation.

Another good option in PHREG procedure to estimate the covariance matrix is to use the Robust Sandwich Variance Estimation (RSVE), which is specified by SAS COVS(AGGREGATE) option. This method sums the score residuals from each distinct ID value, representing distinct clusters. When invoking RSVE method, the ID statement must be specified. We can use a linear combination of clusters and stratum to create a distinct ID value for the sample. RSVE is calculated as

$$\hat{V}_{RSV} = H^{-1}(\hat{\beta}) \left[\sum_{j=1}^n (w_j L_j(\hat{\beta})) \right] \left[\sum_{j=1}^n (w_j L_j(\hat{\beta})) \right]' H^{-1}(\hat{\beta}) \quad (4)$$

We can rewrite (4) in the same format as (3), with a new residual matrix

$$G_{RSV} = \left[\sum_{j=1}^n (w_j L_j(\hat{\beta})) \right] \left[\sum_{j=1}^n (w_j L_j(\hat{\beta})) \right]', \text{ which in the survey notation, can be written as}$$

$$G_{RSV} = \sum_{h'=1}^{H'} \sum_{i=1}^{n_{h'}} e_{h'i} e_{h'i}'.$$

h' represents the new cluster, a distinct linear combination of cluster and stratum. G_{RSV} mimics G matrix in (3), giving a comparable covariance matrix estimation.

In the next session, we will use the nationally representative Health and Retirement Study (HRS) survey data to test both RSVE and normalized weight methods.

HEALTH AND RETIREMENT STUDY DATASET

Functional impairment, such as difficulty with Activities of Daily Living (ADL), is commonly associated with adverse events including acute care hospitalization, nursing home admission, and death. In our example, we want to explore the association between long-time survival and functional impairment, depression, and physical activities for late middle age participants.

We created a nationally representative cohort of 5,650 community-dwelling seniors enrolled in the Health and Retirement Study (HRS) at age 50-56 years old in survey waves of 1992 or 1994 (HRS cohort), 1998 (War Babies Cohort-WB), and 2004 (Early Baby Boomers cohort-EBB), and who did not have ADL impairment at the time they entered the study sample. HRS is a longitudinal study that measures the health and economic circumstances changes within aging Americans. It is a nationally representative sample of participants over the age of 50. It started in 1992 and new participants are added in the study every 6 years so that the sample remains representative of the population over age 50. There are two Primary Sampling Units and 56 strata in HRS.

In our study, we built a Cox proportional hazard model for 20-year survival estimation. Subjects were censored after 20 years follow up. The primary predictors were ADL impairment, depression, and physical activities. The secondary predictors included socio-demographic variables like age, gender, race, marital status, education level, income, and net worth. As a

measure of health status, we also adjusted for the body mass index (BMI). We showed the estimation of the regression coefficients and the survival prediction for the primary predictors in the result section.

RESULTS

The main predictors, ADL functional impairment, depression, and physical activities, were all statistically significant in the Cox model (p-values<0.05). As expected, both functional impairment and depression were associated with an increased risk of death with hazard ratio (HR) 95% CI above 1, while physical activity has protective effect for survival with HR 95% CI below 1.

Table 1 Parameter Estimation

Methods	Parameters	ADL impairment	Depression	Physical activity
SurveyPhreg (golden standard)	Estimate	0.350	0.395	-0.532
	Standard error	0.1203	0.1250	0.0983
	P-value	0.0052	0.0025	<.0001
	95% CI for HR	1.42 (1.12, 1.81)	1.49 (1.16, 1.91)	0.59 (0.48, 0.72)
RSVE	Estimate	0.350	0.395	-0.532
	Standard error	0.1161	0.1223	0.0967
	P-value	0.0026	0.0012	<.0001
	95% CI for HR	1.42 (1.13, 1.78)	1.49 (1.19, 1.89)	0.59 (0.49, 0.72)
Normalized Weight	Estimate	0.350	0.395	-0.532
	Standard error	0.0933	0.0924	0.0882
	P-value	0.0002	<.0001	<.0001
	95% CI for HR	1.42 (1.18, 1.80)	1.49 (1.24, 1.78)	0.59 (0.49, 0.70)
Phreg	Estimate	0.338	0.445	-0.414
	Standard error	0.0853	0.0835	0.0768
	P-value	<.0001	<.0001	<.0001
	95% CI for HR	1.40 (1.20, 1.66)	1.56 (1.32, 1.84)	0.66 (0.57, 0.77)

The estimates and hazard ratios obtained from SURVEYPHREG, RSVE and Normalized weights are identical, which indicates that the maximum likelihood estimator of weighted partially likelihood function is consistent when the weight is specified, regardless of the survey design. The standard error estimation, however, tells quite a different story. Both RSVE and normalized weight methods underestimate the standard errors, which could lead to incorrect hypothesis testing. The RSVE method by far provides the closest estimation to the golden standard. Without any weights or design specification, the estimate and standard errors from PRREG procedure are both underestimated.

Even both RSVE and normalized weight methods try to mimic the survey design, they still treat the sample as simple randomly selected. That is, the nonindependence and disproportionality under distinct clusters are not counted when calculating the standard errors. Additionally,

despite all three variables were statistically significant in all four methods, the p-values from the golden standard are larger than the p-values from the rest of the methods. Thus, the result of the hypothesis testing with the survey design may be completely different than without considering complex survey design, leading to opposite conclusions.

For the predictions shown in table 2, RSVE and normalized weight gave almost identical survival prediction for 20-year follow-up. On the other hand, the standard error estimation had no clear patterns. The RSVE standard errors in the reference group (no functional impairment, no depression, no physical activities) were smaller than the standard errors using the normalized weight method, while in the risk group, the results were the opposite. When no weight was specified, predictions were smaller compared to weighted predictions.

Table 2 Survival Prediction

	ADL impairment		Depression		Physical activity	
	No	Yes	No	Yes	No	Yes
RSVE						
Survival	0.801	0.736	0.807	0.723	0.746	0.840
S.E.	0.011	0.025	0.011	0.024	0.015	0.012
95% CI	(0.78, 0.82)	(0.69, 0.79)	(0.79, 0.89)	(0.68, 0.77)	(0.72, 0.78)	(0.82, 0.86)
Normalized weight						
Survival	0.801	0.736	0.807	0.723	0.746	0.840
S.E.	0.013	0.022	0.013	0.023	0.017	0.013
95% CI	(0.78, 0.83)	(0.69, 0.78)	(0.78, 0.83)	(0.68, 0.77)	(0.72, 0.78)	(0.82, 0.89)
No weight						
Survival	0.784	0.715	0.792	0.692	0.736	0.815
S.E.	0.011	0.021	0.011	0.021	0.014	0.012
95% CI	(0.76, 0.81)	(0.68, 0.76)	(0.77, 0.81)	(0.65, 0.74)	(0.71, 0.77)	(0.79, 0.84)

DISCUSSION

Since there is no solution to estimate the standard errors of survey survival prediction, we tried 2 methods to bypass the issue. Our rationale was to utilize the BASELINE statement of PHREG for survival prediction to get the parameter estimation as close as they are from SURVEYPHREG. We applied both Robust Sandwich Variance Estimation and Normalized Weight methods to mimic the survey design effects. The parameters estimation indicated the same results among golden standard, RSVE and normalized weight, suggesting the consistency of weighted maximum likelihood estimator. On the other hand, the standard error estimations were different. Both RSVE and Normalized weight methods underestimated the standard error for the regression coefficients compared to SURVEYPHREG. Nevertheless, RSVE performed slightly better than the Normalized weight method given that the covariance matrix sandwiched – a G matrix – is comparable to the G matrix in SURVEYPHREG. Regardless of how we try to mimic

the survey design, the non-independence and disproportional probability of selection in the sample units cannot be replicated.

CONCLUSION

The variation of subjects among clusters and strata cannot be completely captured. The survival prediction obtained using the RSVE and Normalized weight methods were accurate when sampling weights were specified. On the contrary, the standard error estimation was incorrect, as indicated by SAS Technical Support in email exchange on this issue. It is difficult to count for events with survey design. There is one theoretical solution for the standard error estimation (Boudreau and Lawless, 2006), but no statistical packages have adopted it. Survival prediction provides invaluable information for researches especially clinicians. The inferences of predictions provide a vital way to assess risk uncertainty, which aids clinicians for decision making. As more clinicians use survey data as primary data resources, the needs for survey survival prediction is increasing. We hope that SAS could develop the BASELINE statement for the SURVEYPHREG procedure, so that correct inference for survival prediction can be drawn.

SAMPLE CODES

```
/*Create distinct cluster ID for RSVE. Create normalized weight*/
data outcome_ADL1;
    set outcome_ADL;
    clusterID=100*secu+stratum;
    norm_wt=sampling_weight/mean_wt;
run;

/*****Applying different models for parameter estimation*****/

*The golden standard method;
proc surveyphreg data=outcome_ADL1;
    class gender(ref='0') race(ref='0') BMI(ref='0')
marital_status(ref='0') education(ref='0') income(ref='0')
        networth(ref='0') ADL_diff(ref='0') depression(ref='0')
phy_acti(ref='0')/param=reference;
    model mortality*death(0)=age_at_baseline gender race BMI
marital_status education income
        networth ADL_diff depression phy_acti;
    cluster secu;
    strata stratum;
    weight sampling_weight;
run;

*The RSVE method;
```



```

proc phreg data=outcome_ADL1 covsandwich(aggregate);
class gender(ref='0') race(ref='0') BMI(ref='0')
marital_status(ref='0') education(ref='0') income(ref='0')
      networkth(ref='0') ADL_diff(ref='0') depression(ref='0')
phy_acti(ref='0')/param=reference;
  model mortality*death(0)=age_at_baseline gender race BMI
marital_status education income networkth ADL_diff depression phy_acti;
  id clusterID;
  weight sampling_weight;
run;

*The normalized weight method;
proc phreg data=outcome_ADL1 ;
  class gender(ref='0') race(ref='0') BMI(ref='0')
marital_status(ref='0') education(ref='0') income(ref='0')
      networkth(ref='0') ADL_diff(ref='0') depression(ref='0')
phy_acti(ref='0')/param=reference;
  model mortality*death(0)=age_at_baseline gender race BMI
marital_status education income networkth ADL_diff depression phy_acti;
  weight norm_wt;
run;

*NO weight method;
proc phreg data=outcome_ADL1 ;
  class gender(ref='0') race(ref='0') BMI(ref='0')
marital_status(ref='0') education(ref='0') income(ref='0')
      networkth(ref='0') ADL_diff(ref='0') depression(ref='0')
phy_acti(ref='0')/param=reference;
  model mortality*death(0)=age_at_baseline gender race BMI
marital_status education income networkth ADL_diff depression phy_acti;
run;

/***** Prediction for ADL function impairment*****/

** dicotomize all the categorical variables;
data outcome_ADL2;
  set outcome_ADL1;
  if race=1 then race1=1;else race1=0;
  if race=2 then race2=1;else race2=0;
  if race=3 then race3=1;else race3=0;
  if BMI=1 then BMI1=1;else BMI1=0;
  if BMI=2 then BMI2=1;else BMI2=0;
  if BMI=3 then BMI3=1;else BMI3=0;
  if income=1 then income1=1;else income1=0;
  if income=2 then income2=1;else income2=0;
  if income=3 then income3=1;else income3=0;
  if networkth=1 then networkth1=1;else networkth1=0;
  if networkth=2 then networkth2=1;else networkth2=0;
  if networkth=3 then networkth3=1;else networkth3=0;
run;

```

```

/*Creating a covariate set that contains all the mean values of each
predictor*/
proc sql;
  create table pred_adl as
  select distinct
    adl_diff,
    avg(age_at_baseline) as age_at_baseline,
    avg(gender) as gender,
    avg(race1) as race1,
    avg(race2) as race2,
    avg(race3) as race3,
    avg(marital_status) as marital_status,
    avg(education) as education,
    avg(BMI1) as BMI1,
    avg(BMI2) as BMI2,
    avg(BMI3) as BMI3,
    avg(income1) as income1,
    avg(income2) as income2,
    avg(income3) as income3,
    avg(networth1) as networth1,
    avg(networth2) as networth2,
    avg(networth3) as networth3,
    avg(depression) as depression,
    avg(phy_acti) as phy_acti
  from outcome_ADL2;
quit;

/*Use RSVE method to get prediction and sved it in adl_pred_RSVE*/
proc phreg data = outcome_ADL2 covsandwich(aggregate)
plots(overlay)=survival;
  model mortality*death(0) = age_at_baseline gender race1-race3
marital_status education BMI1-BMI3 income1-income3 networth1-networth3
                                depression phy_acti
adl_diff;
  id clusterID;
  weight sampling_weight;
  baseline out=adl_pred_RSVE covariates=pred_adl survival=_all_
/rowid=adl_diff;
run;

/*Getting the prediction at the end of follow-up period*/
proc sort data=adl_pred_rsve;by adl_diff mortality;run;

data adl_pred_rsve1 (keep=adl_diff mortality survival StdErrSurvival
LowerSurvival UpperSurvival);
  set adl_pred_rsve;
  by adl_diff mortality;
  if last.adl_diff;
run;

```

REFERENCE

1. Gardiner, JC., "Survival Analysis with Survey Data", Paper 2040-2015, SAS Global Forum 2015, Dallas, TX.
2. Harrell, FE., "Regression Modeling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis", Springer Series, June 15, 2001.
3. Gardiner JC., "Survival Analysis: Overview of Parametric, Nonparametric and Semiparametric approaches and New Developments", Paper 252-2010, SAS Global Forum 2010; Seattle, WA.
4. Fang X., Ning, J., Huang, X., "Empirical Comparison of the Breslow Estimator and the Kalbfleisch Prentice Estimator for Survival Functions", Journal of Biometrics and Biostatistics, Oct 25, 2018.
5. 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.
6. Lin, DY., Wei, LJ., "The Robust Inference for the Cox Proportional Hazard Model", Journal of American Statistical Association, Vol. 84, No. 408, Dec 1989.
7. Huang, X., Strawderman, RL., "A Note on the Breslow Survival Estimator", Journal of Nonparametric Statistics, 18:1, 45-56, 2006.
8. Kalbfleisch, JD., Prentice, RL., "Marginal Likelihoods Based on Cox's Regression and Life Model", Biometrika, Vol. 60, No.2, 267-278, Aug 1973.
9. Lin, DY., "On the Breslow Estimator", Lifetime Data Analysis, 13: 471-480, 2007.
10. Breslow, NE., "Discussion of the paper by D. R. Cox", Journal of the Royal Statistical Society, Series B 34:216-217, 1972.
11. Cox, DR., "Regression Models and Life-tables (with discussion)", Journal of the Royal Statistical Society, Series B 34: 187-220.
12. Hahs-Vaughn, DL., McWayne, CM., Bulotsky-Shearer, RJ., Wen, X., Faria, A., "Methodological Considerations in Using Complex Survey Data: AN Applied Example with the Head Start Family and Child Experiences Survey", SAGE Journals, 35(3) 269-303, 2011.
13. Thomas, SL., Heck, RH., "Analysis of Large-Scale Secondary Data in Higher Education Research: Potential Perils Associated with Complex Sampling Designs", Research in Higher Education Journal, Vol. 42, No. 5, 2001.
14. SAS Institute Inc. 2017. SAS/STAT® 14.3 User's Guide. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Bocheng Jing

Statistician

NCIRE-The Veterans Health Research Institute

4150 Clement St, Building 1, Office 210, San Francisco, CA94141

Email: bocheng.jing@ncire.org

L. Grisell Diaz-Ramirez

Senior Statistician

Department of Medicine

University of California, San Francisco

4150 Clement St, Building 1, Office 210, San Francisco, CA94141

Email: Grisell.diaz-ramirez@ucsf.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.