HCHS/SOL Analysis Methods - Visit 2

July 2020

Version 2

Prepared by the

HCHS/SOL Coordinating Center

Collaborative Studies Coordinating Center
UNC Department of Biostatistics

Jianwen Cai
Daniela Sotres-Alvarez
Donglin Zeng
Pedro Baldoni
Nicole Butera
Beibo Zhao
Franklyn Gonzalez II
Marston Youngblood

Analysis Methods for HCHS/SOL Visit 2

Table of Contents

i. FOREWORD 3

Note to Users of these Analysis Methods Guidelines 3

MAIN Updates in Version 2.0 (July 2020) 4

1. INTRODUCTION 5

1.1. Sampling Weights for Visit 2 5

1.2. Comparison of Estimates for Baseline Characteristics Using Data from Visits 1 and 2 6

2. Linear Regression Models for Change in Continuous Measures 11

2.1. Linear Regression Model for the Difference between Visit 2 and Visit 1 11

2.1.1. SAS 11

2.1.2. SUDAAN 12

2.1.3. R 13

2.1.4. STATA 14

2.2. Linear Regression Model for the Rate of Change 15

2.2.1. SAS 16

2.2.2. SUDAAN 16

2.2.3. R 17

2.2.4. Stata 18

3. Logistic Regression for Visit 2 Binary Outcome 19

3.1. SAS 19

3.2. SUDAAN 20

3.3. R 22

3.4. Stata 23

4. Poisson Regression with Robust Variance 24

4.1. SUDAAN 24

4.2. R 27

4.3. Stata 28

5. Survival Analysis for Right Censored Incident Event Time Data 30

5.1. Diabetes Definitions and the Outcome Variables for Right Censored Incident Event Time Data 30

5.2. Complex Survey Procedures for Cox Regression Model 34

5.2.1. SAS.....35

5.2.2. SUDAAN 39

5.2.3. R 41

5.2.4. Stata 42

5.2.5. Mplus 44

5.3. Weighted Regression Analysis Approach for Cox Regression Model 46

5.3.1. SAS.....47

5.3.2. R 49

5.3.3. Stata 50

REFERENCES 53

i. FOREWORD

Note to Users of these Analysis Methods Guidelines

- This Guide is for illustration purposes in working with the HCHS/SOL visit 1 and visit 2 datasets and has been developed using participants who attended both visit 1 and visit 2 (n=11,623).
- Included on the HCHS/SOL visit 2 examination datasets with INV3 extension are three sampling weight variables (weight_norm_overall_v2, weight_norm_center_v2, and weight_expanded_v2). All weights were calibrated to the age, gender and Hispanic/Latino background distributions from the 2010 US Census for the four study field centers based on participants' visit 1 age. Go to HCHS/SOL Analyses Methods at Baseline to understand the differences between these and their proper use.
- The document is not intended for direct citation.
- Statistical program output used in the examples in this Guide has been modified and/or formatted for presentation and clarity.
- Additional documentation for SAS 9.4 can be found at https://support.sas.com/documentation/onlinedoc/stat/ for SAS 9.2 at: http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#titlepage.htm and for SAS 9.3 at: http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#titlepage.htm

MAIN Updates in Version 2.0 (July 2020)

- Uses data from most updated PART_DERV_V2_INV3 (July 2020; N=11,623).
- NEW Chapter 5 illustrates how to analyze right censored incident event time data in HCHS/SOL, using DIABETES definition #5 as an example.

MAIN Updates in Version 1.1 (March 2018)

- HCHS/SOL Visit 2 Database Version 2 (March 2018; N=11,623) with final sampling weight variable (weight_norm_overall_v2) is used rather than HCHS/SOL Visit 2 Database Version 1 (November 2016; N=9,329) with weight_norm_overall_v2 derived for the interim data release.
- Chapter 1 is updated to provide information on final sampling weights. Section 1.2 is added for comparison of visit 1 and visit 2 data releases.

1. INTRODUCTION

The purpose of this document is to present a set of statistical procedures to analyze longitudinal data from HCHS/SOL collected at the first two visits of the study. Because the HCHS/SOL cohort was selected through a stratified multi-stage area probability sample design (LaVange, Kalsbeek, et al., 2010), study design specifications will be included in all the analysis. For more details of the sampling design, sampling weights, study design specification, and analysis methods for cross-sectional analysis, please refer to *HCHS/SOL Analysis Methods at Baseline*. This document will focus on analysis methods for longitudinal data with two visits. Specifically, we will provide guidelines for analyzing changes in continuous measures using linear regression models, and changes in binary outcomes using logistic Regression and Poisson Regression Models. For these analyses, examples and SAS/SUDAAN/R/STATA program code will be presented using Body Mass Index (BMI) as a continuous outcome and incidence of Chronic Kidney Disease (CKD) as a discrete outcome. Study design specifications will be included in all the analysis presented. Sampling weights are adjusted for non-responses for visits 1 and 2, trimmed, and calibrated to the age, gender and Hispanic/Latino background distributions from the 2010 US Census for the four study field centers based on participants' visit 1 age.

1.1. Sampling Weights for Visit 2

As in any complex survey design, and as was done for the HCHS/SOL baseline (visit 1), sampling weights account for non-responders. One important and big difference between non-response at visit 1 and visit 2 is that at visit 1 all we knew from non-responders was their age and sex (from screening) whereas at visit 2 we know all their baseline data. The calculation of the sampling weights for visit 2 is based on the sampling weights for visit 1 and accounting for the participant nonresponse for visit 2.

To identify baseline factors that are associated with the probability of returning to attend visit 2, a classification tree approach (R package rpart) was used. The advantage of the classification tree approach is that it took interactions among the baseline factors into consideration and it also provides estimates for the cutpoints for continuous variables. The baseline factors that we considered are Hispanic/Latino Background, Gender, Strata, Education, Income, Mental Health, Physical Health, Alcohol Use, Cigarette Use, Diabetes Status, Employment Status, Physical Activity, Prevalent Hypertension, Prevalent MI, Health Insurance, Prevalent Stroke, Born in Mainland US, Years Lived in US, and AFU refusal for categorical variables. For continuous variables we considered Age, BMI, Cardiac Risk Ratio, eGFR, Log-Distance to Field Center, Triglycerides, HDL, LDL, Glucose, Creatinine, Urine Creatinine, Urine Micro albumin, Albumin/Creatinine Ratio, Cystatin, Height, Weight, and Insulin. The classification tree identified AFU

refusal, Log-Distance to Field Center, Hispanic/Latino Background, eGFR, Gender, Strata, and Education to be associated with the probability of returning to visit 2.

Visit 1 nonresponse adjustment was stratified on field center, gender and 6-level age groups. Based on the classification tree results for visit 2 nonresponse adjustment and building on the strata formed by field center, gender and age groups, we formed finer strata based on AFU refusal, log-distance to field center (cutpoints: 4.35 and 4.67), Hispanic/Latino background, eGFR (cutpoints 103 and 110), strata, and education. The smallest number of participants in strata formed by field center, gender and age groups is 90, hence we required the number of participants to be at least 90 to form a finer stratum in order to obtain a reliable nonresponse rate. The nonresponse rate for visit 2 is then calculated for each stratum. The sampling weights are calculated based on visit 1 nonresponse adjusted sampling weights and these nonresponse rates for visit 2. The sampling weights are then trimmed, calibrated to the age, gender and Hispanic/Latino background distributions from the 2010 US Census for the four study field centers based on participants' visit 1 age, and normalized (weight_norm_overall_v2).

1.2. Comparison of Estimates for Baseline Characteristics Using Data from Visits 1 and 2

The sampling weights that are released for visit 1 data (weight_final_norm_overall) and for visit 2 data (weight_norm_overall_v2) are both for inferences in the HCHS/SOL target population. Due to the trimming of the sampling weights, which is a necessary step to control the variability of the non-response rate, the estimates for the target population based on these two sampling weights could be slightly different. We compared the estimates for some baseline characteristics using visit 1 sampling weights (weight_final_norm_overall) with data from visit 1 to those using visit 2 sampling weights (weight_norm_overall_v2) with data from visit 2. The SAS code that produced the estimates as well as the table that summarizes the results are provided below.

```
data sol;
      merge inv1.part_derv_inv4(keep=id strat psu_id
weight_final_norm_overall age education_c3) inv2.part_derv_v2_inv2(keep=id
weight_norm_overall_v2 consent_v2 in=inpart2);
      by id;
      *VISIT2 is an indicator that the participant attended Visit 2;
      if inpart2 & consent_v2=1 then VISiT2=1;
      else VISIT2=0;
      label VISIT2='Participant in Visit 2';
run;

proc sort data=sol;
      by strat PSU_ID;
run;

********** Example Code for Continuous Variable **********;
```

```
* For Visit 1 Target Population (N=16415, weight=WEIGHT_FINAL_NORM_OVERALL);
proc descript data=sol filetype=sas design=wr /* notsorted */;
      nest strat PSU_ID / NOSORTCK;
      weight WEIGHT_FINAL_NORM_OVERALL;
      var AGE;
run;

* For Visit 2 Target Population (N=11623, weight=WEIGHT_NORM_OVERALL_V2);
proc descript data=sol filetype=sas design=wr /* notsorted */;
      nest strat PSU_ID / NOSORTCK;
      subpopn VISIT2=1;
      weight WEIGHT_NORM_OVERALL_V2;
      var AGE;
run;

********** Example Code for Categorical Variable **********;
* For 1 Target Population (N=16415, weight=WEIGHT_FINAL_NORM_OVERALL);
proc descript data=sol filetype=sas design=wr /* notsorted */;
      nest strat PSU_ID / NOSORTCK;
      subgroup EDUCATION_C3;
      levels 3;         *number of levels for the categorical variable;
      weight WEIGHT_FINAL_NORM_OVERALL;
      var EDUCATION_C3 EDUCATION_C3 EDUCATION_C3; *the variables listed on
the VAR statement correspond to the levels listed on the CATLEVEL statement;
      catlevel 1 2 3; *specify the categories for which percents are
requested;
run;

* For Visit 2 Target Population (N=11623, weight=WEIGHT_NORM_OVERALL_V2);
proc descript data=sol filetype=sas design=wr /* notsorted */;
      nest strat PSU_ID / NOSORTCK;
      subgroup EDUCATION_C3;
      subpopn VISIT2=1;
      levels 3; *number of levels for the categorical variable;
      weight WEIGHT_NORM_OVERALL_V2;
      var EDUCATION_C3 EDUCATION_C3 EDUCATION_C3; *the variables listed on
the VAR statement correspond to the levels listed on the CATLEVEL statement;
      catlevel 1 2 3; *specify the categories for which percents are
requested;
run;
```

To compare the results, we examined the absolute differences, defined as value_at_v2-value_at_v1, and the relative differences, defined as (value_at_v2-value_at_v1)/ value_at_v1. Comparing the results, we note that these estimates all have the absolute value of the absolute difference less than 1.6 and the absolute value of the relative difference less than 12%.

Characteristics of HCHS/SOL Target Population using Data from Visit 1 (Baseline) and Visit 2 (Follow-up)

Characteristic[a]	Visit 1 Target Population (N=16415)				Visit 2 Target Population (N=11623)					
	N	Mean or %	Low 95%	Up 95%	N	Mean or %	Low 95%	Up 95%	Absolute Difference	Relative Difference
Age (years)	16415	41.06	40.6	41.5	11623	41.11	40.6	41.6	0.05	0.00
Gender (%)										
Male	6583	47.88	46.8	48.9	4281	47.88	46.6	49.1	0.01	0.00
Female	9832	52.12	51.1	53.2	7342	52.12	50.9	53.4	-0.01	0.00
Education (%)										
Less than high school	6207	32.35	31.0	33.8	4358	32.18	30.6	33.8	-0.17	-0.01
High school graduate	4180	28.20	27.1	29.3	2900	27.68	26.4	29.0	-0.52	-0.02
Greater than high school	5937	39.46	37.9	41.1	4322	40.14	38.4	41.9	0.69	0.02
Hispanic/Latino background(%)										
Cuban	2348	20.02	16.9	23.5	1645	20.03	17.2	23.3	0.02	0.00
Dominican	1473	9.94	8.6	11.4	1021	9.93	8.6	11.5	-0.01	0.00
Mexican	6472	37.37	34.2	40.6	4806	37.28	34.2	40.5	-0.09	0.00
Puerto Rican	2728	16.15	14.7	17.8	1801	15.96	14.4	17.6	-0.19	-0.01
Central American	1732	7.40	6.4	8.6	1207	7.58	6.4	9.0	0.17	0.02
South American	1072	4.98	4.4	5.6	795	4.85	4.2	5.6	-0.13	-0.03
Other	503	4.13	3.6	4.7	313	4.36	3.7	5.1	0.23	0.05
Annual family income(%)										
<$20,000	7207	41.85	40.2	43.6	5070	42.75	40.9	44.6	0.90	0.02
$20,000-$50,000	6119	36.88	35.6	38.2	4424	36.60	35.0	38.3	-0.28	-0.01
>$50,000	1601	11.70	10.3	13.3	1156	11.24	9.9	12.7	-0.46	-0.04
Not reported	1488	9.57	8.8	10.4	973	9.40	8.5	10.3	-0.16	-0.02
Marital status(%)										
Single	4522	34.64	33.3	36.0	2890	34.98	33.3	36.7	0.34	0.01

Characteristic[a]	Visit 1 Target Population (N=16415)				Visit 2 Target Population (N=11623)				Absolute Difference	Relative Difference
	N	Mean or %	Low 95%	Up 95%	N	Mean or %	Low 95%	Up 95%		
Married or living with partner	8436	48.82	47.3	50.4	6253	48.82	46.9	50.7	0.00	0.00
Seprated divorced, or widowed	3369	16.54	15.6	17.6	2438	16.20	15.1	17.3	-0.34	-0.02
Health insurance(%)	7920	50.54	48.7	52.4	5589	50.95	49.0	52.9	0.41	0.01
US residence >= 10 Years(%)	3805	27.66	25.8	29.6	2629	28.08	26.1	30.2	0.41	0.01
Language preference(%)										
Spanish	13119	74.86	73.0	76.6	9517	75.51	73.6	77.3	0.65	0.01
English	3296	25.14	23.4	27.0	2106	24.49	22.7	26.4	-0.65	-0.03
Systolic BP (mmHg)	16401	119.92	119.4	120.4	11616	119.62	119.1	120.1	-0.30	0.00
Diastolic BP (mmHg)	16394	72.19	71.9	72.5	11611	72.10	71.7	72.5	-0.09	0.00
Hypertension (%)	4937	24.19	23.0	25.4	3684	24.17	22.9	25.5	-0.03	0.00
Treated for hypertension(%)[b]	3464	79.78	77.9	81.5	2661	80.17	78.0	82.2	0.39	0.00
Total cholesterol(mg/dL)	16248	194.32	193.2	195.4	11533	194.68	193.4	195.9	0.36	0.00
LDL-cholesterol(mg/dL)	15918	119.74	118.8	120.7	11308	120.19	119.1	121.3	0.45	0.00
HDL-cholesterol(mg/dL)	16246	48.48	48.2	48.8	11533	48.49	48.1	48.9	0.01	0.00
eGFR	16131	106.92	106.3	107.5	11457	107.34	106.7	108.0	0.42	0.00
Treated for hypercholesterolemia(%)[c]	1629	34.64	32.4	37.0	1629	33.57	31.3	35.9	-1.08	-0.03
BMI kg/m^2	16344	29.36	29.2	29.5	11584	29.40	29.2	29.6	0.04	0.00
Obesity Status (%)										
Underweight (BMI<18.5 kg/m^2)	130	1.16	0.9	1.5	73	1.11	0.8	1.5	-0.05	-0.04
Normal (BMI 18.5-25 kg/m^2)	3191	22.07	21.1	23.1	2133	22.01	20.8	23.3	-0.06	0.00
Overweight (BMI 25-30 kg/m^2)	6116	37.19	36.0	38.4	4398	36.87	35.5	38.2	-0.32	-0.01
Obese (BM>=30 kg/m^2)	6907	39.58	38.3	40.9	4980	40.01	38.6	41.4	0.43	0.01

Characteristic[a]	Visit 1 Target Population (N=16415)				Visit 2 Target Population (N=11623)					
	N	Mean or %	Low 95%	Up 95%	N	Mean or %	Low 95%	Up 95%	Absolute Difference	Relative Difference
Fasting glucose(mg/dL)	16220	102.20	101.4	103.0	11519	102.26	101.3	103.2	0.06	0.00
Diabetes (%)	3218	14.88	14.1	15.7	2392	15.07	14.2	16.0	0.18	0.01
Treated for diabetes(%)[d]	1836	61.76	59.1	64.3	1380	62.13	59.0	65.2	0.38	0.01
Waist circumference (cm)	16349	97.37	96.9	97.8	11590	97.48	97.0	97.9	0.11	0.00
Current Smoker (%)	3166	21.37	20.3	22.5	2066	19.83	18.6	21.1	-1.55	-0.08
Asthma (%)	2637	17.37	16.4	18.4	1858	17.74	16.6	19.0	0.38	0.02
COPD (%)	488	2.78	2.4	3.2	354	2.75	2.4	3.2	-0.02	-0.01
CVD (%)	858	4.72	4.2	5.3	607	4.44	3.9	5.0	-0.29	-0.06
MI (%)	384	2.34	2.0	2.7	274	2.08	1.7	2.5	-0.26	-0.12
Hearing Loss (%)	2799	15.06	14.2	16.0	2031	14.74	13.8	15.7	-0.33	-0.02

Abbreviations: BMI: body mass index; BP: blood pressure; LDL: low density lipoprotein; HDL: high density lipoprotein; COPD: chronic obstructive pulmonary disease; CVD: cardiovascular disease; MI: myocardial infarction.

[a]All values (except N) weighted for study design and non-response.

[b]Denominator is restricted to participants with hypertension (Unweighted Visit 1: N=4937, Visit 2: N=3684).

[c]Denominator is restricted to participants with hypercholesterolemia (Unweighted Visit 1: N=5332, Visit 2: N=5332).

[d]Denominator is restricted to participants with diabetes (Unweighted Visit 1: N=3384, Visit 2: N=2511).

2. Linear Regression Models for Change in Continuous Measures

For a continuous measure, change from visit 1 to visit 2 can be described in two ways: (1) the difference between visit 2 and visit 1; and (2) the rate of change from visit 1 to visit 2. Throughout this section, BMI will be used as the outcome of interest for illustration purposes. In the examples provided, we examine the effect of baseline age (AGE) on the change in BMI after adjusting for gender (GENDER) and baseline BMI (BMI).

2.1. Linear Regression Model for the Difference between Visit 2 and Visit 1

In this section, we model the difference in BMI between visit 2 and visit 1, denoted as BMI_V2V1 and defined as BMI_V2-BMI. Because the length between visit 1 and visit 2 varies among participants, we will adjust for the time elapsed between visit 1 and visit 2 (YRS_BTWN_V1V2) in the model. Note: the default option when incorporating the study design for SAS and R is sampling with replacement (WR), while for SUDAAN, the option `design= “wr” ’ needs to be specified.

2.1.1. SAS

The procedure SURVEYREG is used to produce linear regression estimates while accounting for the study design of the HCHS/SOL. Design variables are specified through the statements *strata*, *cluster*, and *weight*.

The following example code creates the analysis dataset that will be used throughout this document. Note the creation of the two derived variables BMI_V2V1 (difference in BMI between visit 1 and visit 2) and RBMI_V2V1 (rate of change in BMI between visit 1 and visit 2).

```
data worklib.sol;
      merge Inv2.Part_derv_v2_inv2(keep=ID BMI_V2 YRS_BTWN_V1V2
WEIGHT_NORM_OVERALL_V2 CKD2 CKD2_V2 in=inv2)
               Inv1.Part_derv_inv4;
      by ID;
      BMI_V2V1 = BMI_V2 - BMI;
      RBMI_V2V1 = BMI_V2V1/YRS_BTWN_V1V2;
      CKD2_V2_SUDAAN = CKD2_V2;
      if CKD2_V2 = 0 then CKD2_V2_SUDAAN = 2;
      label CKD2_V2_SUDAAN='1=Yes, 2=No';*Recoding as SUDAAN models the last
category as reference;
      KEEP_DATA_CKD = (CKD2 = 0); *Keep only those CKD free at baseline for
incidence modelling;
      if inv2;
run;
```

By default, SURVEYREG will set the last category of each of the class variables as the reference level; for example, for gender, GENDER=M (Male) will be the reference level. In order to change the reference level of a class variable in this procedure, one might choose to either change the format of a variable or recode a variable, with the former being preferable. If we are interested in making inference on a particular subpopulation, we need to use the domain statement, for example, domain KEEP_DATA, where KEEP_DATA is a variable indicating the subpopulation of interest.

```
proc surveyreg data=worklib.sol; /* DEFAULT: order=formatted */
   strata STRAT; cluster PSU_ID; weight WEIGHT_NORM_OVERALL_V2;
   *domain KEEP_DATA;
   class GENDER;
   model BMI_V2V1 = AGE GENDER BMI YRS_BTWN_V1V2 / solution;
run;
```

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	4.6318398	0.54560688	8.49	<.0001
AGE	-0.0357911	0.00341966	-10.47	<.0001
GENDER F	0.1999512	0.09440903	2.12	0.0346
GENDER M	0.0000000	0.00000000	.	.
BMI	-0.1152792	0.01155684	-9.97	<.0001
YRS_BTWN_V1V2	0.1089852	0.06404904	1.70	0.0893

This result indicates that after adjusting for gender, baseline BMI, and years elapsed between visits, a one-year increment in age at baseline is associated with a decrease of 0.03 kg/m^2 in the change in BMI. In other words, if BMI increases over time, then older age at baseline is associated with smaller increase in BMI from visit 1 to visit 2.

2.1.2. SUDAAN

The next code invokes PROC REGRESS to produce the equivalent model fitted in SAS by using PROC SURVEYREG. Because SUDAAN cannot handle non-numeric categorical covariates, such as GENDER that assumes values ‘M’ and ‘F’, the variable GENDERNUM that takes values ‘1’ and ‘2’ will be used. Note that results produced from SUDAAN and SAS are very similar, after rounding the results to one decimal place. Also, note that SUDAAN requires the dataset to be sorted with respect to the variables specified in the NEST statement; to avoid sorting the dataset manually, the option NOTSORTED can be used in the main statement, which automatically sorts the dataset internally. If interest lies on making inference for a specific subpopulation, one might specify an additional variable, for example, KEEP_DATA=1, where KEEP_DATA is a variable indicating the subpopulation of interest, in the SUBPOPN statement.

```
proc regress data=worklib.sol filetype=sas design=wr notsorted;
   nest strat PSU_ID;
   weight WEIGHT_NORM_OVERALL_V2;
   class GENDERNUM;
   *subpopn KEEP_DATA=1;
   model BMI_V2V1 = AGE GENDERNUM BMI YRS_BTWN_V1V2 ;
   reflevel GENDERNUM=1; /* reference: Male */
   setenv decwidth=4;
run;
```

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable BMI_V2V1: BMI_V2V1
by: Independent Variables and Effects.
```

Independent Variables and Effects	Beta Coeff.	SE Beta	Lower 95% Limit Beta	Upper 95% Limit Beta	T-Test B=0	P-value T-Test B=0
Intercept	4.6318	0.5455	3.5606	5.7031	8.4906	0.0000
Age	-0.0358	0.0034	-0.0425	-0.0291	-10.4681	0.0000
Gender (0=Female, 1=Male)						
0	0.2000	0.0944	0.0146	0.3853	2.1184	0.0345
1	0.0000	0.0000	0.0000	0.0000	.	.
BMI (kg/m2)	-0.1153	0.0116	-0.1380	-0.0926	-9.9757	0.0000
Elapsed time between visits 1 and 2 (yrs)	0.1090	0.0640	-0.0168	0.2347	1.7019	0.0893

2.1.3. R

In order to fit linear regression models (or generalized linear models) in R, one needs to specify the study design by invoking the *svydesign* function and storing it in a variable that will be used later on. The function *svydesign* requires the user to specify the variables for the Primary Sampling Unit (argument 'id'), the strata (argument 'strata'), the weights (argument 'weights'), and, finally, the dataset to be analyzed. Note that, during the process of model fitting or any computation that involves the study design, only the variable storing the study design will be used; therefore, if one creates an additional variable, for example, during the pipeline of the analysis, a new call of *svydesign* will be needed considering the updated dataset.

After specifying the study design, the user can proceed with the model fitting. In the code below, we invoke the function *svyglm*, which fits generalized linear models and takes into account the study design through the input argument 'design'. The model

itself is specified as a regular model following the pattern of the well-known function *glm*. If we are interested in making inference for a specific subpopulation, we need to subset the original full dataset by making use of the 'subset' argument and the condition KEEP_DATA ==1, where KEEP_DATA is a variable indicating the subpopulation of interest. Finally, because we want to fit a linear regression model, we specify the Gaussian family with identity link through the 'family' argument.

```
sol.design = svydesign(id=~PSU_ID, strata=~STRAT, weights=~WEIGHT_NORM_OVERALL_V2, data=sol)

model.diff = svyglm(BMI_V2V1 ~ AGE + GENDER + BMI + YRS_BTWN_V1V2, design =
sol.design,subset=KEEP_DATA==1,family=gaussian(link='identity'))

summary(model.diff)
Call:
svyglm(formula = BMI_V2V1 ~ AGE + GENDER + BMI + YRS_BTWN_V1V2, design = sol.design, subset =
KEEP_DATA == 1, family = gaussian(link = "identity"))

Survey design:
svydesign(id = ~PSU_ID, strata = ~STRAT, weights = ~WEIGHT_NORM_OVERALL_V2, data = sol)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.631840   0.545527   8.491 < 2e-16 ***
AGE           -0.035791   0.003419  -10.468  < 2e-16 ***
GENDERF        0.199951   0.094386   2.118    0.0345
BMI           -0.115279   0.011556  -9.976  < 2e-16 ***
YRS_BTWN_V1V2  0.108985   0.064038   1.702    0.0893
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 9.150868)
Number of Fisher Scoring iterations: 2
```

2.1.4. STATA

In Stata, the analysis dataset first needs to be loaded into working memory. This can be done using the *use* command for Stata datasets (with a ".dta" file extension) or the *import* command if the dataset is in a different format (e.g., CSV files, Excel files, SAS XPORT Transport files). Then any variable in the loaded dataset can be referenced by its variable name. The *fvset* command can be used to change the reference level of any "factor" (categorical) variables for all subsequent analyses; for example, the command *fvset base last gendernum diabetes2_indicator bkgrd1_c7* changes the reference level for the variables GENDERNUM, DIABETES2_INDICATOR, and BKGRD1_C7 from the lowest category (the default) to the highest category.

The survey design can be specified for the analysis dataset using the *svyset* command. The *svyset* command requires the user to specify the primary sampling unit (psu_id), sampling weight (weight_norm_overall_v2), and strata (strat). This command only needs to be run once at the beginning of the program (after loading the analysis dataset, but before running any statistical analyses).

After specifying the survey design, the linear regression can be fit using the *regress* command with the usual syntax. The prefix *svy* should be used with the *regress* command to ensure that the linear regression accounts for the complex survey design specified using the *svyset* command. Note that adding the characters "*i.*" to a predictor variable when specifying the regression model (e.g., *i.gendernum* in the example below) indicates that the variable is a "factor" (categorical) variable.

```
. import delimited "H:\PATH\sol.csv", clear
(273 vars, 16,415 obs)

. fvset base last gendernum diabetes2_indicator bkgrd1_c7
. svyset psu_id [pw=weight_norm_overall_v2], strata(strat)
      pweight: weight_norm_overall_v2
          VCE: linearized
  Single unit: missing
     Strata 1: strat
         SU 1: psu_id
        FPC 1: <zero>

. svy: regress bmi_v2v1 age i.gendernum bmi yrs_btwn_v1v2
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =        20              Number of obs     =     11,212
Number of PSUs     =       648              Population size   = 11,120.631
                                            Design df         =        628
                                            F(   4,    625)   =      93.77
                                            Prob > F          =     0.0000
                                            R-squared         =     0.0872

------------------------------------------------------------------------------
             |             Linearized
    bmi_v2v1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0357911   .003419   -10.47   0.000    -.0425052   -.0290769
  0.gendernum |   .1999512   .0943922     2.12   0.035     .0145887    .3853137
         bmi |  -.1152792   .0115548    -9.98   0.000    -.1379699   -.0925885
yrs_btwn_v1v2 |   .1089852   .0640376     1.70   0.089    -.0167686     .234739
       _cons |    4.63184   .5455095     8.49   0.000     3.560596    5.703083
------------------------------------------------------------------------------
```

2.2. Linear Regression Model for the Rate of Change

For the models fitted below, the outcome of interest is the rate of change in BMI between the first two visits, denoted as RBMI_V2V1 and defined as the ratio between BMI_V2-BMI and the time between the two visits YRS_BTWN_V1V2. The rate of change has already taken the varying length of time between the two visits into consideration in the outcome variable, therefore we do not need to additionally adjust for it in the model.

2.2.1. SAS

The code provided below invokes the procedure SURVEYREG in SAS to produce parameter estimates for the desired model. As before, statements and options specified are the same to the ones presented for the model that fits the difference in BMI between visits 1 and 2.

```
proc surveyreg data=worklib.sol; /* DEFAULT: order=formatted */
   strata STRAT; cluster PSU_ID; weight WEIGHT_NORM_OVERALL_V2;
   *domain KEEP_DATA;
   class GENDER;
   model RBMI_V2V1 = AGE GENDER BMI / solution;
run;
```

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.8736091	0.04837530	18.06	<.0001
AGE	-0.0058929	0.00055809	-10.56	<.0001
GENDER F	0.0323073	0.01542131	2.09	0.0366
GENDER M	0.0000000	0.00000000	.	.
BMI	-0.0191282	0.00183703	-10.41	<.0001

The results indicate that after adjusting for gender and baseline BMI, a one-year increment in age at baseline is associated with a decrease of 0.006 kg/m^2 in the annual rate of change in BMI. In other words, if BMI increases over time, then older age at baseline is associated with slower annual increase in BMI.

2.2.2. SUDAAN

The same type of model can be fitted in SUDDAN by invoking the procedure REGRESS. The same statements and options are used as before for modeling the difference in BMI between the two visits.

```
proc regress data=worklib.sol filetype=sas design=wr notsorted;
   nest STRAT PSU_ID;
   weight WEIGHT_NORM_OVERALL_V2;   class GENDERNUM;
   *subpopn KEEP_DATA=1;
   model RBMI_V2V1 = AGE GENDERNUM BMI;
   reflevel GENDERNUM=1; /* reference: Male */
   setenv decwidth=4;
run;
```

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable RBMI_V2V1: RBMI_V2V1
by: Independent Variables and Effects.
```

Independent Variables and Effects	Beta Coeff.	SE Beta	Lower 95% Limit Beta	Upper 95% Limit Beta	T-Test B=0	P-value T-Test B=0
Intercept	0.8736	0.0484	0.7786	0.9686	18.0609	0.0000
Age	-0.0059	0.0006	-0.0070	-0.0048	-10.5604	0.0000
Gender (0=Female, 1=Male)						
0	0.0323	0.0154	0.0020	0.0626	2.0954	0.0365
1	0.0000	0.0000	0.0000	0.0000	.	.
BMI (kg/m2)	-0.0191	0.0018	-0.0227	-0.0155	-10.4129	0.0000

2.2.3. R

As for the linear regression model that used the difference in BMI as the outcome, the model for the rate of change also considers the survey design element ‘sol.design’, which has already been created for the first model. The only part that is different from the previous code is the specification of the linear model; because we are modeling the rate of change, only age, gender, and baseline BMI are entered into the model. We also use the Gaussian and identity link for this class of models.

```
> model.rdif = svyglm(RBMI_V2V1 ~ AGE + GENDER + BMI, design =
sol.design,subset=KEEP_DATA==1,family=gaussian(link='identity'))

> summary(model.rdif)

Call:
svyglm(formula = RBMI_V2V1 ~ AGE + GENDER + BMI, design = sol.design, subset = KEEP_DATA == 1,
family = gaussian(link = "identity"))

Survey design:
svydesign(id = ~PSU_ID, strata = ~STRAT, weights = ~WEIGHT_NORM_OVERALL_V2, data = sol)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.873609  0.048370  18.061   <2e-16 ***
AGE         -0.005893  0.000558  -10.560   <2e-16 ***
GENDERF      0.032307  0.015418  2.095    0.0365
BMI         -0.019128  0.001837 -10.413   <2e-16 ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.2523977)

Number of Fisher Scoring iterations: 2
```

2.2.4. Stata

Again, the *regress* command is used to specify the linear regression model, and the *svy* prefix is used to indicate that the survey design specified using the *svyset* command (run earlier in the program) should be used.

```
. svy: regress rbmi_v2v1 age i.gendernum bmi
(running regress on estimation sample)

Survey: Linear regression
Number of strata   =       20                  Number of obs     =     11,212
Number of PSUs     =      648                  Population size   = 11,120.631
                                                Design df         =       628
                                                F(   3,    626)   =    123.76
                                                Prob > F          =    0.0000
                                                R-squared         =    0.0839
------------------------------------------------------------------------------
             |             Linearized
  rbmi_v2v1  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0058929   .000558   -10.56   0.000    -.0069887   -.0047971
  0.gendernum |   .0323073   .0154192    2.10   0.037     .0020277    .0625868
         bmi |  -.0191282   .0018368   -10.41   0.000    -.0227352   -.0155212
       _cons |   .8736091   .0483688    18.06   0.000     .7786248    .9685933
------------------------------------------------------------------------------
```

3. Logistic Regression for Visit 2 Binary Outcome

The purpose of this section is to estimate odds ratio of the incidence event using Logistic regression in SAS, SUDAAN and R. We use incidence of Chronic Kidney Disease (CKD) at visit 2 as an example. CKD at visit 2 is defined by eGFR less than 60 mL/min per 1.73 in binary variable CKD_V2. To study CKD incidence, the population of interest is those who did not have CKD at baseline visit. The flag variable KEEP_DATA_CKD is defined to select those without CKD at visit 1. Because the time length between visit 1 and visit 2 varies among participants, we will adjust for time elapsed between visit 1 and visit 2 (YRS_BTWN_V1V2). Note that odds ratios are different from incidence rate ratios when the event is not rare (incidence rate > 10%). If incidence rates are of interest, we recommend Poisson regression which provides direct estimates related to incidence rate (see Section 4 for details).

3.1. SAS

The code below invokes the SAS procedure SURVEYLOGISTIC; this procedure fits Logistic Regression models for either binary, nominal or ordinal variables while accounting for the survey design. Similar to REGRESS, study design variables are specified in the statements STRAT, CLUSTER and WEIGHT. The subpopulation is specified in the DOMAIN statement and the categorical covariates are included in the CLASS statement. Note that it is not necessary to include the outcome in the CLASS statement. Finally, since we are fitting models for the odds ratio of an outcome, we include the option LINK as logit. This option is important when the outcome of interest is either nominal or ordinal, and the user might want to choose one among several types of link functions available. Note that the default parameterization of SURVEYLOGISTIC is the effect coding; in order to change it to the reference cell parameterization, we use the option PARAM=REF.

```
proc surveylogistic data=worklib.sol; /* DEFAULT: order=formatted */
   strata STRAT; cluster PSU_ID; weight WEIGHT_NORM_OVERALL_V2;
   domain KEEP_DATA_CKD;
   class GENDER / PARAM=REF;
   model CKD2_V2(EVENT='1') = AGE GENDER YRS_BTWN_V1V2/ link=logit;
run;
```

Type 3 Analysis of Effects

Effect	F Value	Num DF	Den DF	Pr > F
AGE	61.86	1	630	<.0001
GENDER	0.39	1	630	0.5331
YRS_BTWN_V1V2	0.00	1	630	0.9672

Analysis of Maximum Likelihood Estimates					
Parameter		**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept		-4.5078	0.6049	-7.45	<.0001
AGE		0.0370	0.00470	7.87	<.0001
GENDER	**F**	0.0788	0.1264	0.62	0.5331
YRS_BTWN_V1V2		0.00354	0.0860	0.04	0.9672
NOTE: The degrees of freedom for the t tests is 630.					

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Confidence Limits**	
AGE	1.038	1.028	1.047
GENDER F vs M	1.082	0.844	1.387
YRS_BTWN_V1V2	1.004	0.848	1.188
NOTE: The degrees of freedom in computing the confidence limits is 630.			

The results indicate that the odds ratio for incident CKD is 1.082 for females relative to males after adjusting for baseline age and time between the two visits.

3.2. SUDAAN

The following code invokes the MULTILOG procedure and fits the equivalent model fitted by the SAS procedure SURVEYLOGISTIC. The study design variables are specified in the statements NEST (strata and primary sampling unit) and WEIGHT (weight_norm_overall_v2). The outcome of interest is the categorical variable CKD_V2, which assumes either 0 or 1; as such, this variable should be included in the CLASS statement along with any other categorical predictor that one might want to include in the statistical model. Note that, by default, SUDAAN outputs results using only two decimal places; in order to increase this number, one might want to use the statement SETENV and set the number of decimal places to be used through the option DECWIDTH. Note that results from both SAS and SUDAAN agree.

```
proc multilog data=worklib.sol filetype=sas design=wr notsorted;
   nest STRAT PSU_ID;
   weight WEIGHT_NORM_OVERALL_V2;
   class CKD2_V2_SUDAAN GENDERNUM;
   subpopn KEEP_DATA_CKD=1;
   model CKD2_V2_SUDAAN = AGE GENDERNUM YRS_BTWN_V1V2;
   reflevel GENDERNUM=1; /* reference: Male */
   setenv decwidth=4;
run;
```

Contrast	Degrees of Freedom	Wald F	P-value Wald F
AGE	1.0000	61.8739	0.0000
GENDERNUM	1.0000	0.3890	0.5331
YRS_BTWN_V1V2	1.0000	0.0017	0.9672

CKD2_V2_SUDAAN (log-odds)		Independent Variables and Effects				
		Intercept	Age	Gender (0=Female, 1=Male) = 0	Gender (0=Female, 1=Male) = 1	Elapsed time between visits 1 and 2 (yrs)
1 vs 2	Beta Coeff.	-4.5078	0.0370	0.0788	0.0000	0.0035
	SE Beta	0.6048	0.0047	0.1264	0.0000	0.0860
	Lower 95% Limit Beta	-5.6955	0.0278	-0.1693	0.0000	-0.1653
	Upper 95% Limit Beta	-3.3201	0.0462	0.3269	0.0000	0.1724
	T-Test B=0	-7.4533	7.8660	0.6237	.	0.0411
	P-value T-Test B=0	0.0000	0.0000	0.5331	.	0.9672

CKD2_V2_SUDAAN (log-odds)		Independent Variables and Effects				
		Intercept	Age	Gender (0=Female, 1=Male) = 0	Gender (0=Female, 1=Male) = 1	Elapsed time between visits 1 and 2 (yrs)
1 vs 2	Odds Ratio	0.0110	1.0377	1.0820	1.0000	1.0035
	Lower 95% Limit OR	0.0034	1.0282	0.8442	1.0000	0.8476
	Upper 95% Limit OR	0.0361	1.0473	1.3867	1.0000	1.1882

3.3. R

Fitting generalized linear models (when the outcome is not continuous and is not normally distributed) while taking into account the study design is straightforward and relatively similar to the regular linear model that we fitted for the difference and relative change models. The only difference between them is the specification of a new family, in this case the 'quasibinomial' family, and the 'logit' link function; the choice of the quasibinomial family is recommended by the package developers as it avoids some warnings from the package. It provides exactly the same point estimates and standard errors as the usual 'binomial' family.

```
> model.bin = svyglm(CKD_V2 ~ AGE + GENDER + YRS_BTWN_V1V2, design =
sol.design,subset=KEEP_DATA==1,family=quasibinomial(link='logit'))

> summary(model.bin)

Call:

svyglm(formula = CKD_V2 ~ AGE + GENDER + YRS_BTWN_V1V2, design = sol.design, subset =
KEEP_DATA == 1, family = quasibinomial(link = "logit"))

Survey design:
svydesign(id = ~PSU_ID, strata = ~STRAT, weights = ~WEIGHT_NORM_OVERALL_V2, data = sol)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4.507825   0.604828   -7.453 3.04e-14 ***
AGE           0.037003   0.004704  7.866 1.61e-14 ***
GENDERF        0.078804   0.126356   0.624 0.533 .
YRS_BTWN_V1V2 0.003535   0.086003  0.041 0.967 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 1.014773)

Number of Fisher Scoring iterations: 5


exp(cbind(Odds=coef(model.bin), confint(model.bin)))
                     Odds       2.5 %       97.5 %
(Intercept)      0.0110       0.0034  0.0361
AGE               1.0377       1.0282  1.0473
GENDERF           1.0819       0.8446  1.3805
YRS_BTWN_V1V2     1.0035       0.8469  1.1878
```

3.4. Stata

Logistic regression can be fit using the *logit* command with the usual syntax. Again, the prefix *svy* should be used with the *logit* command to ensure that the logistic regression accounts for the complex survey design specified using the *svyset* command. Odds ratios can be requested by using the option *or* (either with the original *logit* command call, or by using the statement *logit, or* after the logistic regression was fit).

```
. svy, subpop(if keep_data_ckd==1): logit ckd2_v2 age i.gendernum yrs_btwn_v1v2
(running logit on estimation sample)

Survey: Logistic regression

Number of strata   =       20                  Number of obs     =     11,593
Number of PSUs     =      651                  Population size   = 11,598.435
                                                Subpop. no. obs   =     10,090
                                                Subpop. size      = 10,277.241
                                                Design df         =       631
                                                F(   3,    629)   =     23.18
                                                Prob > F          =    0.0000
------------------------------------------------------------------------------
             |             Linearized
    ckd2_v2  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0370034   .0047041     7.87   0.000     .0277658    .046241
  0.gendernum|   .0788045   .1263553     0.62   0.533    -.1693234    .3269323
yrs_btwn_v1v2|   .0035349   .0860025     0.04   0.967    -.165351     .1724207
       _cons |  -4.507825   .6048276    -7.45   0.000    -5.695544   -3.320107
------------------------------------------------------------------------------


. logit, or

Survey: Logistic regression

Number of strata   =       20                  Number of obs     =     11,593
Number of PSUs     =      651                  Population size   = 11,598.435
                                                Subpop. no. obs   =     10,090
                                                Subpop. size      = 10,277.241
                                                Design df         =       631
                                                F(   3,    629)   =     23.18
                                                Prob > F          =    0.0000
------------------------------------------------------------------------------
             |             Linearized
    ckd2_v2  | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.037697   .0048814     7.87   0.000     1.028155    1.047327
  0.gendernum|   1.081993   .1367156     0.62   0.533     .8442359    1.386708
yrs_btwn_v1v2|   1.003541   .0863071     0.04   0.967     .8475962    1.188178
       _cons |   .0110224   .0066667    -7.45   0.000     .0033609     .036149
------------------------------------------------------------------------------
```

4. Poisson Regression with Robust Variance

When incidence rate is of interest, Poisson regression with robust variance can be used. Incidence rate ratio can be significantly different from odds ratios when the event of interest is not rare (incidence rate > 10%). The model can provide estimation of covariate effect on the incidence rate. Adjusted incidence rate can also be estimated. This section presents fitting Poisson Regression through generalized linear regression models. SUDAAN and R have procedures to fit this class of models while taking the study design into consideration; similar procedures are under development in SAS. We use incidence of Chronic Kidney Disease at visit 2 as an example. CKD at visit 2 is defined by eGFR less than 60 mL/min per 1.73 in binary variable CKD_V2. To study CKD incidence, the population of interest is those who did not have CKD at visit 1. The flag variable KEEP_DATA_CKD is defined to select those without CKD at visit 1.

4.1. SUDAAN

The following code invokes the SUDAAN procedure LOGLINK which uses the same set of statements and options as in the MULTILOG procedure. Note, however, that Poisson regression models assume, by default, that our response is a count variable; here, CKD_V2, can only assume two possible values (0 and 1). Thus, there is no need to specify our outcome of interest in the class statement when fitting this class of models in SUDAAN. Note that an OFFSET option needs to be specified for the time elapsed between visit 1 and visit 2. The option PREDMARG is specified for requesting the estimates of incidence rate.

```
proc loglink data=worklib.sol filetype=sas design=wr notsorted;
   nest STRAT PSU_ID;
   weight WEIGHT_NORM_OVERALL_V2;
   class BKGRD1_C7 GENDERNUM DIABETES2_INDICATOR;
   subpopn KEEP_DATA_CKD=1;
   model CKD2_V2 = AGE BKGRD1_C7 GENDERNUM DIABETES2_INDICATOR  /
OFFSET=YRS_BTWN_V1V2;
   reflevel GENDERNUM=1 BKGRD1_C7=6 DIABETES2_INDICATOR=1; /* reference: Male
*/
   TEST WALDCHI;
   PREDMARG / ALL;
   setenv decwidth=4;
run;
```

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Log
Response variable CKD2_V2: Chronic Kidney Disease using eGFR (creatinine only, no race) and albumin-creatinine ratio
  (NIDDK) at Visit 2
Offset variable YRS_BTWN_V1V2: Elapsed time between visits 1 and 2 (yrs)
For Subpopulation: KEEP_DATA_CKD = 1
by: Contrast.

Contrast	Degrees of Freedom	Wald ChiSq	P-value Wald ChiSq
OVERALL MODEL	10.0000	6727.2247	0.0000
MODEL MINUS INTERCEPT	9.0000	190.4564	0.0000
INTERCEPT	.	.	.
BKGRD1_C7	6.0000	9.6073	0.1422
GENDERNUM	1.0000	0.1464	0.7020
DIABETES2_INDICATOR	1.0000	85.0051	0.0000
AGE	1.0000	30.2283	0.0000

Independent Variables and Effects	Beta Coeff.	SE Beta	Lower 95% Limit Beta	Upper 95% Limit Beta	T-Test B=0	P-value T-Test B=0
Intercept	-5.5278	0.4290	-6.3702	-4.6855	-12.8866	0.0000
7-level re-classification of Hispanic/Latino Background						
0	0.4795	0.4037	-0.3132	1.2723	1.1878	0.2353
1	0.6015	0.4022	-0.1882	1.3912	1.4956	0.1352
2	0.5619	0.3677	-0.1603	1.2840	1.5279	0.1270
3	0.5602	0.3855	-0.1967	1.3171	1.4534	0.1466
4	0.8170	0.3786	0.0736	1.5604	2.1582	0.0313
5	0.2784	0.4053	-0.5174	1.0743	0.6870	0.4923
6	0.0000	0.0000	0.0000	0.0000	.	.
Gender (0=Female, 1=Male)						
0	0.0451	0.1179	-0.1864	0.2767	0.3826	0.7022
1	0.0000	0.0000	0.0000	0.0000	.	.
Diabetes Indicator - ADA						
0	-1.0393	0.1127	-1.2606	-0.8179	-9.2198	0.0000
1	0.0000	0.0000	0.0000	0.0000	.	.
Age	0.0239	0.0043	0.0154	0.0324	5.4980	0.0000

Predicted Marginal #1	Predicted Marginal	SE	Lower 95% Limit	Upper 95% Limit	T:Marg=0	P-value
Intercept	0.0092	0.0005	0.0082	0.0104	16.9489	0.0000
7-level re-classification of Hispanic/Latino Background						
0	0.0084	0.0016	0.0058	0.0121	5.3330	0.0000
1	0.0094	0.0015	0.0070	0.0128	6.4113	0.0000
2	0.0091	0.0012	0.0070	0.0118	7.4883	0.0000
3	0.0091	0.0009	0.0075	0.0110	10.2384	0.0000
4	0.0117	0.0012	0.0096	0.0143	9.7124	0.0000
5	0.0068	0.0015	0.0044	0.0106	4.4744	0.0000
6	0.0052	0.0019	0.0025	0.0107	2.7170	0.0068
Gender (0=Female, 1=Male)						
0	0.0094	0.0007	0.0081	0.0110	12.9499	0.0000
1	0.0090	0.0008	0.0075	0.0108	11.1004	0.0000
Diabetes Indicator - ADA						
0	0.0071	0.0005	0.0062	0.0082	13.6338	0.0000
1	0.0201	0.0019	0.0167	0.0242	10.5368	0.0000

Independent Variables and Effects	Incidence Density Ratio	Lower 95% Limit IDR	Upper 95% Limit IDR
Intercept	0.0040	0.0017	0.0092
7-level re-classification of Hispanic/Latino Background			
0	1.6153	0.7311	3.5692
1	1.8248	0.8284	4.0197
2	1.7540	0.8519	3.6110
3	1.7511	0.8214	3.7328
4	2.2637	1.0764	4.7608
5	1.3210	0.5960	2.9279
6	1.0000	1.0000	1.0000
Gender (0=Female, 1=Male)			
0	1.0461	0.8299	1.3187
1	1.0000	1.0000	1.0000
Diabetes Indicator - ADA			
0	0.3537	0.2835	0.4414
1	1.0000	1.0000	1.0000
Age	1.0242	1.0155	1.0330

The results indicate that the incidence rate for CKD is higher for those without diabetes at baseline than those who had diabetes at baseline after adjusting for age, Hispanic/Latino background, and gender. The table also provides the adjusted incidence rate ratios.

4.2. R

The Poisson regression model is fitted similarly as the logistic regression model; the only exception is the specification of the 'quasipoisson' family and the 'log' link. Again, the choice of the quasipoisson family avoids warnings from the package and produce exactly the same point estimates and standard errors as the regular 'Poisson' family. Note that the argument in the offset option for R is the logarithm transformation of the time elapsed between visit 1 and visit 2, which is different from the specification in SUDAAN in which the original variable for time elapsed between visit 1 and visit 2 is used.

```
#Start
> model.pois = svyglm(CKD2_V2 ~ AGE +BKGRD1_C7+ GENDER+ DIABETES2_INDICATOR+
offset(log(YRS_BTWN_V1V2)), design = sol.design,
+                     subset=KEEP_DATA_CKD==1,family=quasipoisson(link='log'))
> summary(model.pois)


Call:
svyglm(formula = CKD2_V2 ~ AGE + BKGRD1_C7 + GENDER + DIABETES2_INDICATOR +
    offset(log(YRS_BTWN_V1V2)), design = sol.design, subset = KEEP_DATA_CKD ==
    1, family = quasipoisson(link = "log"))


Survey design:
svydesign(id = ~PSU_ID, strata = ~STRAT, weights = ~WEIGHT_NORM_OVERALL_V2,
    data = sol)
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -5.527828   0.428959 -12.887  < 2e-16 ***
AGE                   0.023903   0.004348   5.498 5.61e-08 ***
BKGRD1_C70            0.479550   0.403715   1.188   0.2353
BKGRD1_C71            0.601483   0.402159   1.496   0.1353
BKGRD1_C72            0.561873   0.367733   1.528   0.1270
BKGRD1_C73            0.560219   0.385457   1.453   0.1466
BKGRD1_C74            0.817020   0.378567   2.158   0.0313 *
BKGRD1_C75            0.278427   0.405286   0.687   0.4923
GENDERF               0.045112   0.117911   0.383   0.7022
DIABETES2_INDICATOR0 -1.039266   0.112721  -9.220  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


(Dispersion parameter for quasipoisson family taken to be 0.9779706)
```

```
Number of Fisher Scoring iterations: 6


> exp(cbind(IDR=coef(model.pois), confint(model.pois)))
                              IDR       2.5 %      97.5 %
(Intercept)          0.003974613 0.001714604 0.009213524
AGE                  1.024190880 1.015500796 1.032955330
BKGRD1_C70           1.615346773 0.732188390 3.563762047
BKGRD1_C71           1.824823408 0.829664203 4.013648484
BKGRD1_C72           1.753954028 0.853107073 3.606059344
BKGRD1_C73           1.751055873 0.822617683 3.727365376
BKGRD1_C74           2.263742746 1.077929677 4.754049667
BKGRD1_C75           1.321049999 0.596951810 2.923474004
GENDERF              1.046144925 0.830282444 1.318128803
DIABETES2_INDICATOR0 0.353714382 0.283598934 0.441164790
#End
```

4.3. Stata

Poisson regression can be fit using the *poisson* command with the usual syntax. Again, the prefix *svy* should be used with the *poisson* command to ensure that the Poisson regression accounts for the complex survey design specified using the *svyset* command. The *offset* option should be used with the logarithm transformation of the time elapsed between visit 1 and visit 2, similar to R. Incidence-rate ratios can be requested by using the option *irr* (either with the original *poisson* command call, or by using the statement *poisson, irr* after the Poisson regression was fit).

```
. gen log_yrs_btwn_v1v2=ln(yrs_btwn_v1v2)
(4,792 missing values generated)

. svy, subpop(if keep_data_ckd==1): poisson ckd2_v2 age i.bkgrd1_c7 i.gendernum i.diabetes2_
> indicator, offset(log_yrs_btwn_v1v2)
(running poisson on estimation sample)

Survey: Poisson regression

Number of strata   =        20            Number of obs     =     11,574
Number of PSUs     =       651            Population size   = 11,574.514
                                                Subpop. no. obs   =     10,071
                                                Subpop. size      =  10,253.32
                                                Design df         =       631
                                                F(   9,    623)   =     20.89
                                                Prob > F          =    0.0000

------------------------------------------------------------------------------
```

```
                |               Linearized
        ckd2_v2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+------------------------------------------------------------------------------
            age |   .0239029   .0043474     5.50   0.000     .0153658    .0324401
                |
        bkgrd1_c7 |
              0 |   .4795496   .4037087     1.19   0.235    -.3132256    1.272325
              1 |   .6014832   .4021566     1.50   0.135    -.188244     1.39121
              2 |   .5618727    .367725     1.53   0.127    -.1602402    1.283986
              3 |    .560219   .3854509     1.45   0.147    -.1967028    1.317141
              4 |   .8170195    .378561     2.16   0.031     .0736276    1.560411
              5 |   .2784269   .4052823     0.69   0.492    -.5174385    1.074292
    0.gendernum |   .0451119   .1179122     0.38   0.702    -.1864359    .2766597
0.diabetes2_indicator|  -1.039266   .1127226    -9.22   0.000    -1.260622   -.8179087
          _cons |  -5.527828   .4289623   -12.89   0.000    -6.370194   -4.685462
log_yrs_btwn_v1v2 |          1  (offset)
----------------+------------------------------------------------------------------------------
```

```
. poisson, irr

Survey: Poisson regression

Number of strata   =       20                  Number of obs     =     11,574
Number of PSUs     =      651                  Population size   = 11,574.514
                                                Subpop. no. obs   =     10,071
                                                Subpop. size      =  10,253.32
                                                Design df         =        631
                                                F(   9,    623)   =      20.89
                                                Prob > F          =     0.0000

----------------+------------------------------------------------------------------------------
                |               Linearized
        ckd2_v2 |        IRR   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+------------------------------------------------------------------------------
            age |   1.024191   .0044526     5.50   0.000     1.015484    1.032972
                |
        bkgrd1_c7 |
              0 |   1.615347   .6521296     1.19   0.235     .731085    3.569141
              1 |   1.824823   .7338648     1.50   0.135     .8284125    4.019713
              2 |   1.753954   .6449727     1.53   0.127     .8519392    3.611003
              3 |   1.751056   .6749461     1.45   0.147     .8214347    3.732733
              4 |   2.263743   .8569648     2.16   0.031     1.076406    4.760779
              5 |    1.32105   .5353982     0.69   0.492     .5960454     2.92792
                |
    0.gendernum |   1.046145   .1233532     0.38   0.702     .8299118    1.318718
0.diabetes2_indicator |   .3537144   .0398716    -9.22   0.000     .2834775    .4413537
          _cons |   .0039746    .001705   -12.89   0.000     .0017118    .0092285
log_yrs_btwn_v1v2 |          1  (offset)
----------------+------------------------------------------------------------------------------
```

5. Survival Analysis for Right Censored Incident Event Time Data

When time to incident event is of interest, survival analysis methods can be used. In HCHS/SOL, data are collected through clinic visits and annual follow up calls. There are in general two ways in HCHS/SOL to define an incident event depending on the data that have been collected: (a) incident event that is determined by data collected at clinic visits only; and (b) incident event that is determined jointly by data from the clinic visits and the annual follow up calls. For both ways of definition, the incident event time for a participant can be right censored if the participant did not have the event of interest at the last contact with the participant either at clinic visit or through annual follow up calls. Such right censored data can be analyzed using survival analysis methods. Cox regression models can be used to study the association of covariate effects on the hazard of the incident event. Given that stratified multi-stage sampling was used in HCHS/SOL, analyses need to account for the design features of the study such as stratification, clustering, and unequal sampling proportions. Data collected from complex survey designs can be analyzed using complex survey procedures. However, software that has complex survey procedures is limited and it is of interest to examine whether other analysis approach that uses non-survey based procedures can be used for such analysis. Simulation studies were conducted at the Coordinating Center to examine the performance of various methods. Based on simulation results, which will be reported in a separate document, hazard ratios in the Cox regression model can be estimated in two ways for HCHS/SOL:
1) using complex survey procedures;
2) using weighted regression analysis and accounting for clustering.

This section illustrates how to fit a Cox regression model to estimate the hazard ratio of diabetes incidence using these two approaches. We organize this Section in the following way: Section 5.1 presents different definitions for diabetes incidence and introduces the variables needed for diabetes incidence analysis. Section 5.2 provides examples and sample program code (SAS, SUDAAN, R, STATA, and Mplus) using complex survey procedures. Section 5.3 provides examples and sample program code (SAS, R, and STATA) using weighted regression analysis taking the study design and clustering into consideration.

5.1. Diabetes Definitions and the Outcome Variables for Right Censored Incident Event Time Data

To study diabetes incidence, the population of interest consists of those who did not have diabetes at baseline visit. Based on the information that have been collected during the HCHS/SOL baseline visit, four definitions for diabetes have been derived and numbered

as definitions 2 to 5 in the order of creation; see their definitions in the baseline Derived Variable Dictionary. Briefly,

(a) Definition 2 (DIABETES2): based on ADA lab criteria plus **scanned medication**
(b) Definition 3 (DIABETES3): based on ADA lab criteria plus **self-reported diagnosis**
(c) Definition 4 (DIABETES4): based on ADA lab criteria plus **self-reported medication use**
(d) Definition 5 (DIEBETES5): based on ADA lab criteria, **self-reported medication use**, and **self-reported diagnosis**

Ideally, we would like to use the same algorithm as the one that was used at the baseline to define incidence. However, there are some complications that prevent us from using the same algorithm directly. We will discuss each definition for the incidence analysis related to the baseline definition in the following order: DIABETES2, DIABETES4, DIABETES5, and DIABETES3.

Definition 2 (DIABETES2): This definition was used in the HCHS/SOL diabetes prevalence paper (Schneiderman et al, 2014). However, scanned medication is not currently available at Visit 2, therefore for diabetes incidence, it is not feasible to use an equivalent definition.

Definition 4 (DIABETES4): This definition is an approximation to DIABETES2 by replacing scanned medication with self-reported medication use. Baseline self-reported medication use is based on the question MUEA33c "Were any of the medications you took during the last four weeks for high blood sugar or diabetes?" from the Medication Use form. The same question was administered at clinic visit 2 under MUE26c. Note that this question does not track back medication use history, it only asks for medication use information in the past four weeks. The main purpose for including this information in the diabetes definition is to account for the medication's influence on the lab measures. In other words, DIABETES4 is an objective classification based on ADA lab criteria accounting for the medication influence on the lab measurement at the respective visit.

For incident diabetes analysis using DIABETES4_V2 (i.e. diabetes definition 4 using V2 data), use survey procedure for Poisson regression model with time between visits as offset (see Section 4). In order to have a relatively pure group with no diabetes at baseline visit for the incidence analysis, we recommend excluding individuals with diabetes based on DIABETES4 and self-reported being diagnosed at baseline. Note that when DIABETES4_V2 is used for incident analysis, we do NOT recommend excluding individuals with self-reported diagnosis at Visit 2 because we would not want to treat the self-reported diagnosis information collected at Visit 2 differently from those at the Annual Follow-Up calls. More information on self-reported diagnosis is provided below for DIABETES5 and DIABETES3.

Definition 5 (DIABETES5): This definition includes self-reported diagnosis in addition to ADA lab criteria and self-reported medication use. Both at baseline and at clinic visit 2, self-reported diagnosis was asked in the Medical History Form (MHE). However, the question refers to a different time period. **At baseline, the question is: "MHE16. Has a doctor ever said** that you have diabetes (high sugar in blood or urine)?". In contrast, **at clinic visit 2 the question is: "MHE14. Since our last telephone interview** with you, has a doctor or health professional told you that you had diabetes or high sugar in the blood?". Therefore, to capture the self-reported diagnosis at visit 2, we need to also include data from all previous annual follow-up calls when the same question was asked under OPE7 of the Out-Patient Self-Reported Conditions Form.

We treat the incident diabetes data based on ADA lab criteria, self-reported medication use, and self-reported diagnosis as right censored data. Specifically, we define a pair of variables DIABETES5_TIME_V2 and DIABETES5_INDICATOR_V2 to capture the diabetes incidence information, where DIABETES5_TIME_V2 records the time, in days, when diabetes was first reported (baseline, annual follow-up or visit 2) or the time when the participant was last contacted if s/he did not develop diabetes. DIABETES5_INDICATOR_V2 is an indicator variable (1 or 0) of whether or not the participant has diabetes based on either ADA lab criteria, self-reported medication use, or self-reported diabetes status at the recorded time in DIABETES5_TIME_V2. For details on the derivation of variables DIABETES5_TIME_V2 and DIABETES5_INDICATOR_V2, see the Dictionary for Derived Variables for Visit 2.

Case 0) Prevalent case. If a participant reported having diabetes at baseline based on DIABETES5, then:

DIABETES5_TIME_V2 = 0, and
DIABETES5_INDICATOR_V2 = 1;

Below we provide four examples for participants who did not have diabetes at baseline based on DIABETES5:

Case 1) If a participant reported having diabetes at AFU1, then:

DIABETES5_TIME_V2 = AFU1 time - Visit 1 time, and
DIABETES5_INDICATOR_V2 = 1;

Case 2) If a participant did not report having diabetes at AFU1 through AFU4, but reported having diabetes at AFU5, then:

DIABETES5_TIME_V2 = AFU5 time - Visit 1 time, and
(DIABETES5_INDICATOR_V2 = 1;

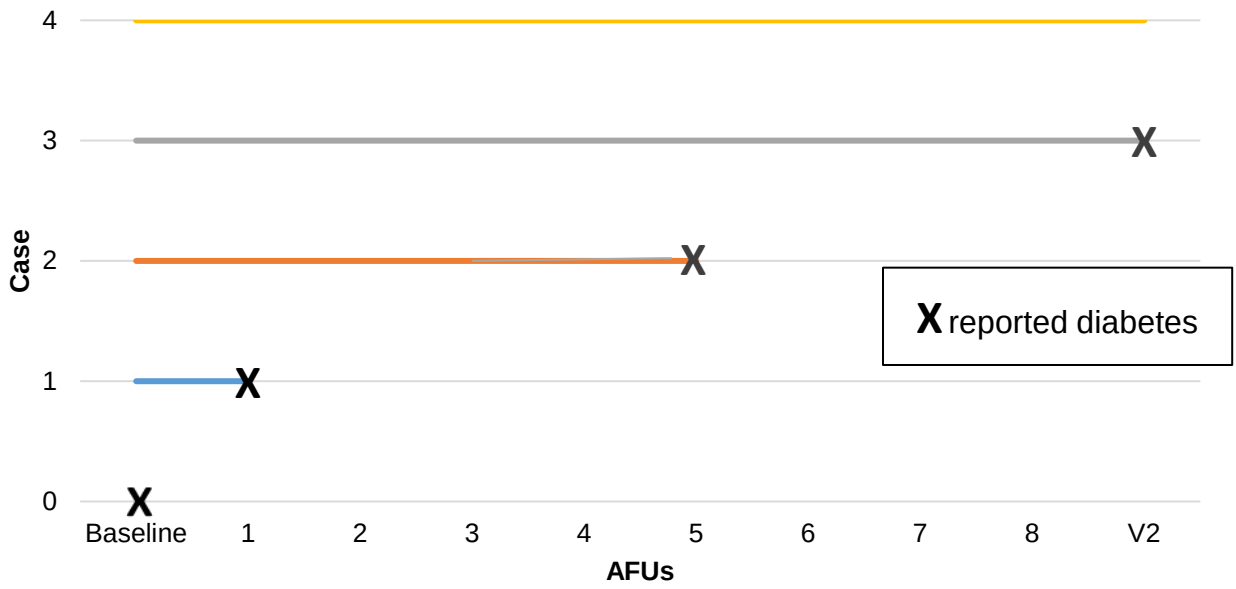
Case 3) If a participant did not report having diabetes at any of the AFUs before Visit 2, but reported having diabetes at Visit 2, then:

DIABETES5_TIME_V2 = Visit 2 time - Visit 1 time, and
DIABETES5_INDICATOR_V2=1;

Case 4) If a participant did not report having diabetes at any of the AFUs before Visit 2, and did not have diabetes based on Visit 2 lab values, did not report diabetes medication use at Visit 2, and did not report having diabetes since the last AFU before Visit 2, then:

DIABETES5_TIME_V2 = Visit 2 time - Visit 1 time, and
DIABETES5_INDICATOR_V2 = 0.

The figure below illustrates these five cases, with lines tracking the recorded follow-up time from baseline, through AFUs, to Visit 2, and crosses (X) marking time points of reported diabetes.



Definition 3 (DIABETES3): This definition is similar to DIABETES5 except that self-reported medication use is not included in the definition. Because self-reported diagnosis is included in the definition, the incidence data structure is similar to that based on Definition 5. Specifically, we treat the incident diabetes data based on ADA lab criteria and self-reported diagnosis as right censored data. We define a pair of variables DIABETES3_TIME_V2 and DIABETES3_INDICATOR_V2 that are similar to DIABETES5_TIME_V2 and DIABETES5_INDICATOR_V2 to capture the diabetes incidence information. For details on the DIABETES3_TIME_V2 and DIABETES3_INDICATOR_V2 derived variables, see the Dictionary for Derived Variables for Visit 2.

5.2. Complex Survey Procedures for Cox Regression Model

In this section, we use Definition 5 to illustrate how to fit Cox Regression Model using complex survey procedures using SAS, SUDAAN, R, STATA, and Mplus. Specifically, the potentially right censored outcome of interest is contained in the pair of variables DIABETES5_TIME_V2 and DIABETES5_INDICATOR_V2. In the examples provided, we examine the effect of baseline CES-D 10, a 10-item CES-D summary score assessing depressive symptoms, on diabetes incidence after adjusting for baseline age, center, gender, Hispanic/Latino background group, education, and income.

An indicator variable KEEP_DATA_DIABETES5 with = 1 identifying the subpopulation of interest – those without diabetes at baseline and having no missing covariates – is created for the incident diabetes analysis. This subpopulation contains 8938 participants with 7478 right-censored times and 1460 event times. Here are the unweighted descriptive statistics of the time variable DIABETES5_TIME_V2, in days, by the event indicator DIABETES5_INDICATOR_V2, within this subpopulation:

Analysis Variable : DIABETES5_TIME_V2 (Recorded Time in Days)

DIABETES5_INDICATOR_V2	N	Mean	Std Dev	Minimum	Maximum
0	7478	2200.775	287.079	1513.000	3506.000
1	1460	1690.325	679.022	300.000	3408.000

Note that the 7478 right-censored times range from 1,513 to 3,506 days (i.e., 4.1 to 9.6 years), with a mean of 2,201 days (i.e. 6 years); the 1460 event times range from 3000 to 3,408 days (i.e., 0.8 to 9.3 years), with a mean of 1690 days (i.e., 4.6 years).

There are a total of 855 distinct failure times for the 1460 events. More specifically, there are 531 distinct failure times at which only one event happened and another 324 distinct failure times at which 2 or more events happened with the number of tied events ranging from 2 to 8 with a median of 2. Different tie handling methods, such as Breslow or Efron methods provide very similar results. Our examples with SAS provide results for both methods for comparison to illustrate this point. Examples with other software provide results for only one method based on respective availability.

Note: the default option when incorporating the study design for SAS, R, STATA, and Mplus is sampling with replacement (WR), while for SUDAAN, the option `design= “wr”’ needs to be specified explicitly.

5.2.1. SAS

The following example code creates the analysis dataset that will be used throughout Section 5. Note the creation of the two derived variables COV_MISS (indicator for missing covariates) and KEEP_DATA_DIABETES5 (indicator for subpopulation of interest).

```
data sol;
      merge inv.part_derv_inv4(keep=ID STRAT PSU_ID DIABETES5 CENTERNUM
GENDERNUM AGE CESD10 BKGRD1_C7 INCOME_C3 EDUCATION_C3 rename =(INCOME_C3 =
INCOME_C3_V1 EDUCATION_C3 = EDUCATION_C3_V1 AGE = AGE_V1 CESD10 = CESD10_V1
DIABETES5 = DIABETES5_V1))
      inv_v2.PART_DERV_V2_inv3(keep=ID WEIGHT_NORM_OVERALL_V2 DIABETES5_V2
DIABETES5_TIME_V2 DIABETES5_INDICATOR_V2 in = inv2);

      by ID;
      if inv2;

      if not missing(BKGRD1_C7) then BKGRD1_C7_NOMISS = BKGRD1_C7; else
BKGRD1_C7_NOMISS = 6; label BKGRD1_C7_NOMISS = 'Missing collapsed with
mixed/other';
      if nmiss(CESD10_V1, EDUCATION_C3_V1, INCOME_C3_V1) > 0 then COV_MISS =
1; else COV_MISS = 0; label COV_MISS = 'Indicator of missing covariates';
      KEEP_DATA_DIABETES5 = (COV_MISS=0 and DIABETES5_V1 in (1, 2)); label
KEEP_DATA_DIABETES5 = "Subpopulation of interest - those without diabetes at
baseline and having no missing covariates";
run;
```

The procedure SURVEYPHREG is used to produce Cox regression estimates while accounting for the study design of the HCHS/SOL. Design variables are specified through the statements *strata*, *cluster*, and *weight*. If we are interested in making inference on a particular subpopulation, we need to use the domain statement, for example, domain KEEP_DATA_DIABETES5, which indicates the subpopulation of interest - those without diabetes at baseline and having no missing covariates. In the *model* statement, we use DIABETES5_INDICATOR_V2 as the event indicator (with '0' specified as the censoring value), and DIABETES5_TIME_V2 as the observed event time.

By default, SURVEYPHREG will set the last category of each of the class variables as the reference level. For example, for baseline gender, GENDERNUM=1 (Male) will be the reference level. In order to change the reference level of a class variable in this procedure, invoke the 'ref = ' option in the *class* statement. For example, for baseline Hispanic/Latino background group, BKGRD1_C7_NOMISS=3 (Mexicans) will be the reference level, set through 'ref = 3'.

By default, SURVEYPHREG will use the Breslow method to handle ties, we can invoke the 'ties = ' option to use the Efron method instead.

```
proc surveyphreg data= sol; /* DEFAULT: order=formatted */
   strata STRAT; cluster PSU_ID; weight WEIGHT_NORM_OVERALL_V2;
   domain KEEP_DATA_DIABETES5;
   class CENTERNUM GENDERNUM BKGRD1_C7_NOMISS(ref = '3') EDUCATION_C3_V1
      INCOME_C3_V1; /* ref: San Diego, Male, Mexicans */
   model DIABETES5_TIME_V2*DIABETES5_INDICATOR_V2(0)= CESD10_V1 AGE_V1
      CENTERNUM GENDERNUM BKGRD1_C7_NOMISS EDUCATION_C3_V1 INCOME_C3_V1 /
      ties = efron; /* DEFAULT: ties = breslow */
run;
```

Efron tie handling results:

Model Information

Data Set	WORK.SOL
Dependent Variable	DIABETES5_TIME_V2
Censoring Variable	DIABETES5_INDICATOR_V2
Censoring Value(s)	0
Weight Variable	WEIGHT_NORM_OVERALL_V2
Stratum Variable	STRAT
Cluster Variable	PSU_ID
Ties Handling	EFRON

Domain Analysis for domain KEEP_DATA_DIABETES5=1

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
8938	1460	7478	83.67

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
CESD10_V1	631	0.015722	0.007164	2.19	0.0286	1.016
AGE_V1	631	0.042941	0.002865	14.99	<.0001	1.044
CENTERNUM B	631	-0.291961	0.245316	-1.19	0.2344	0.747
CENTERNUM C	631	-0.110925	0.143999	-0.77	0.4414	0.895
CENTERNUM M	631	-0.441756	0.236058	-1.87	0.0618	0.643
CENTERNUM S	631	0	.	.	.	1.000
GENDERNUM F	631	0.023183	0.081379	0.28	0.7758	1.023
GENDERNUM M	631	0	.	.	.	1.000

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
BKGRD1_C7_NOMISS 0	631	-0.134836	0.251130	-0.54	0.5915	0.874
BKGRD1_C7_NOMISS 1	631	-0.345717	0.207434	-1.67	0.0961	0.708
BKGRD1_C7_NOMISS 2	631	0.059193	0.223767	0.26	0.7915	1.061
BKGRD1_C7_NOMISS 4	631	0.100840	0.227257	0.44	0.6574	1.106
BKGRD1_C7_NOMISS 5	631	-0.450973	0.243999	-1.85	0.0650	0.637
BKGRD1_C7_NOMISS 6	631	0.240836	0.247358	0.97	0.3306	1.272
BKGRD1_C7_NOMISS 3	631	0	.	.	.	1.000
EDUCATION_C3_V1 1	631	0.119106	0.112252	1.06	0.2891	1.126
EDUCATION_C3_V1 2	631	0.101482	0.102893	0.99	0.3244	1.107
EDUCATION_C3_V1 3	631	0	.	.	.	1.000
INCOME_C3_V1 1	631	0.184609	0.182626	1.01	0.3125	1.203
INCOME_C3_V1 2	631	0.029187	0.191390	0.15	0.8788	1.030
INCOME_C3_V1 3	631	0	.	.	.	1.000

Breslow tie handling results:

Model Information

Data Set	WORK.SOL
Dependent Variable	DIABETES5_TIME_V2
Censoring Variable	DIABETES5_INDICATOR_V2
Censoring Value(s)	0
Weight Variable	WEIGHT_NORM_OVERALL_V2
Stratum Variable	STRAT
Cluster Variable	PSU_ID
Ties Handling	BRESLOW

Domain Analysis for domain KEEP_DATA_DIABETES5=1

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
8938	1460	7478	83.67

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
CESD10_V1	631	0.015721	0.007158	2.20	0.0284	1.016
AGE_V1	631	0.042938	0.002863	15.00	<.0001	1.044
CENTERNUM B	631	-0.291828	0.245029	-1.19	0.2341	0.747
CENTERNUM C	631	-0.110773	0.143866	-0.77	0.4416	0.895
CENTERNUM M	631	-0.441616	0.235886	-1.87	0.0616	0.643
CENTERNUM S	631	0	.	.	.	1.000
GENDERNUM F	631	0.023244	0.081336	0.29	0.7751	1.024
GENDERNUM M	631	0	.	.	.	1.000
BKGRD1_C7_NOMISS 0	631	-0.134814	0.250936	-0.54	0.5913	0.874
BKGRD1_C7_NOMISS 1	631	-0.345685	0.207337	-1.67	0.0960	0.708
BKGRD1_C7_NOMISS 2	631	0.059126	0.223631	0.26	0.7916	1.061
BKGRD1_C7_NOMISS 4	631	0.100762	0.226949	0.44	0.6572	1.106
BKGRD1_C7_NOMISS 5	631	-0.450975	0.243903	-1.85	0.0649	0.637
BKGRD1_C7_NOMISS 6	631	0.240829	0.247252	0.97	0.3304	1.272
BKGRD1_C7_NOMISS 3	631	0	.	.	.	1.000
EDUCATION_C3_V1 1	631	0.119123	0.112208	1.06	0.2888	1.127
EDUCATION_C3_V1 2	631	0.101450	0.102836	0.99	0.3243	1.107
EDUCATION_C3_V1 3	631	0	.	.	.	1.000
INCOME_C3_V1 1	631	0.184749	0.182506	1.01	0.3118	1.203
INCOME_C3_V1 2	631	0.029379	0.191260	0.15	0.8780	1.030
INCOME_C3_V1 3	631	0	.	.	.	1.000

Note that Breslow and Efron methods provide very similar results. These results indicate that after adjusting for baseline age, center, gender, Hispanic/Latino background, education, and income, a one-point increment in baseline CES-D 10 score is significantly associated with a 1.6% increase in the hazard of diabetes incidence. In other words, the higher the baseline CES-D 10 score, the more likely an individual to develop diabetes between Visit 1 and Visit 2.

5.2.2. SUDAAN

The following code invokes the SUDAAN procedure SURVIVAL to fit Cox regression model using complex survey procedures. Design variables are specified through the statements *nest* and *weight*, and domain variable KEEP_DATA_DIABETES5 is specified through the *subpopn* statement, with '=1' indicating subpopulation of interest. The event indicator DIABETES5_INDICATOR_V2 is specified through the *event* statement, and the observed event time DIABETES5_TIME_V2 is modelled through the *model* statement.

By default, SURVIVAL will set the last category of each of the class variables as the reference level and SURVIVAL will use the Efron method to handle ties. Other tie handling methods are not supported.

```
proc survival data=sol filetype=sas design=wr notsorted;
   nest STRAT PSU_ID;
   weight WEIGHT_NORM_OVERALL_V2;
   class CENTERNUM GENDERNUM BKGRD1_C7_NOMISS EDUCATION_C3_V1 INCOME_C3_V1;
   subpopn KEEP_DATA_DIABETES5 = 1;
   event DIABETES5_INDICATOR_V2;
   model DIABETES5_TIME_V2 = CESD10_V1 AGE_V1 CENTERNUM GENDERNUM
      BKGRD1_C7_NOMISS EDUCATION_C3_V1 INCOME_C3_V1;
   reflevel BKGRD1_C7_NOMISS = 3; /* ref: San Diego, Male, Mexicans */
   setenv decwidth=6; /* display results with 6 decimals */
run;
```

```
DESIGN SUMMARY: Variances will be computed using the Taylor Linearization Method, Assuming a

With Replacement (WR) Design

    Sample Weight: WEIGHT_NORM_OVERALL_V2

    Stratification Variables(s): STRAT

    Primary Sampling Unit: PSU_ID

Summary of Event Values
by: DIABETES5_INDICATOR_V2.

-----------------------------------------------------
DIABETES5_INDICATOR-
  _V2                       Frequency      Weighted Sum
-----------------------------------------------------
Censored                     7478.000          8369.077
Non-Censored                 1460.000          1224.146

Variance Estimation Method: Taylor Series (WR)
Dependent Variable: DIABETES5_TIME_V2
Censoring Variable: DIABETES5_INDICATOR_V2
Ties Handling: EFRON
For Subpopulation: KEEP_DATA_DIABETES5 = 1
by: Independent Variables and Effects.
```

Independent Variables and Effects	Beta Coeff.	SE Beta	Lower 95% Limit Beta	Upper 95% Limit Beta	T-Test B=0	P-value T-Test B=0
CESD10_V1	0.015722	0.007164	0.001654	0.029790	2.194583	0.028557
AGE_V1	0.042941	0.002865	0.037314	0.048567	14.987460	0.000000
CENTERNUM						
B	-0.291961	0.245313	-0.773689	0.189766	-1.190156	0.234432
C	-0.110925	0.144000	-0.393700	0.171851	-0.770313	0.441402
M	-0.441756	0.236058	-0.905310	0.021797	-1.871387	0.061753
S	0.000000	0.000000	0.000000	0.000000	.	.
GENDERNUM						
F	0.023183	0.081379	-0.136623	0.182989	0.284876	0.775832
M	0.000000	0.000000	0.000000	0.000000	.	.
BKGRD1_C7_NOMISS						
0	-0.134836	0.251127	-0.627979	0.358308	-0.536923	0.591510
1	-0.345717	0.207432	-0.753057	0.061623	-1.666651	0.096079
2	0.059193	0.223768	-0.380226	0.498611	0.264527	0.791460
3	0.000000	0.000000	0.000000	0.000000	.	.
4	0.100840	0.227257	-0.345431	0.547110	0.443725	0.657393
5	-0.450973	0.244001	-0.930124	0.028177	-1.848245	0.065034
6	0.240836	0.247358	-0.244907	0.726578	0.973633	0.330611
EDUCATION_C3_V1						
1	0.119106	0.112254	-0.101331	0.339543	1.061034	0.289080
2	0.101482	0.102894	-0.100573	0.303538	0.986283	0.324372
3	0.000000	0.000000	0.000000	0.000000	.	.
INCOME_C3_V1						
1	0.184609	0.182625	-0.174017	0.543235	1.010860	0.312470
2	0.029187	0.191390	-0.346651	0.405024	0.152498	0.878843
3	0.000000	0.000000	0.000000	0.000000	.	.

Independent Variables and Effects	Hazards Ratio	Lower 95% Limit	Upper 95% Limit
CESD10_V1	1.015846	1.001655	1.030238
AGE_V1	1.043876	1.038019	1.049766
CENTERNUM			
B	0.746798	0.461308	1.208967
C	0.895006	0.674556	1.187501
M	0.642906	0.404417	1.022036
S	1.000000	1.000000	1.000000
GENDERNUM			
F	1.023454	0.872299	1.200801
M	1.000000	1.000000	1.000000
BKGRD1_C7_NOMISS			
0	0.873859	0.533669	1.430906
1	0.707713	0.470925	1.063561
2	1.060980	0.683707	1.646433
3	1.000000	1.000000	1.000000
4	1.106099	0.707915	1.728252
5	0.637008	0.394505	1.028578
6	1.272312	0.782778	2.067992
EDUCATION_C3_V1			
1	1.126489	0.903634	1.404305

```
  2                     1.106811     0.904319     1.354643
  3                     1.000000     1.000000     1.000000
INCOME_C3_V1
  1                     1.202748     0.840282     1.721567
  2                     1.029617     0.707052     1.499338
  3                     1.000000     1.000000     1.000000
------------------------------------------------------------
```

Note that these results are identical to those from other software with Efron tie handling methods.

5.2.3. R

The *svycoxph* function from R package "survey" is used to fit Cox regression model using complex survey procedures. Design variables are first specified through the *svydesign* function to generate a design object, which is then invoked in *svycoxph*. We use DIABETES5_INDICATOR_V2 as the event indicator (with '== 1' specified as the event value), and DIABETES5_TIME_V2 as the observed event time. Domain variable KEEP_DATA_DIABETES5 is specified in the 'subset' option, with '==1' indicating subpopulation of interest.

Indicator variables are created with desired reference levels, and used in model fitting with *svycoxph*, which cannot specify class variables. By default, *svycoxph* will use the Efron method to handle ties. Other tie handling methods are not supported.

```
sol.design<-svydesign(id=~PSU_ID, strata=~STRAT,
weights=~WEIGHT_NORM_OVERALL_V2, data=sol)

svycoxph(Surv(DIABETES5_TIME_V2,DIABETES5_INDICATOR_V2==1)~CESD10_V1 +AGE_V1
+ CENTERNUM_1 + CENTERNUM_2 + CENTERNUM_3 + GENDERNUM_0+ BKGRD1_C7_NOMISS_0
+BKGRD1_C7_NOMISS_1+BKGRD1_C7_NOMISS_2+BKGRD1_C7_NOMISS_4+BKGRD1_C7_NOMISS_5+
BKGRD1_C7_NOMISS_6+ EDUCATION_C3_V1_1+EDUCATION_C3_V1_2+INCOME_C3_V1_1+
INCOME_C3_V1_2, subset = (KEEP_DATA_DIABETES5 == 1), design=sol.design, data
= sol)  # ref: San Diego, Male, Mexicans
```

```
                       coef exp(coef)  se(coef)      z      p
CESD10_V1           0.015722  1.015846  0.007164  2.195 0.0282
AGE_V1              0.042941  1.043876  0.002865 14.987 <2e-16
CENTERNUM_1        -0.291961  0.746798  0.245316 -1.190 0.2340
CENTERNUM_2        -0.110925  0.895006  0.143999 -0.770 0.4411
CENTERNUM_3        -0.441756  0.642906  0.236058 -1.871 0.0613
GENDERNUM_0         0.023183  1.023454  0.081379  0.285 0.7757
BKGRD1_C7_NOMISS_0 -0.134836  0.873859  0.251130 -0.537 0.5913
BKGRD1_C7_NOMISS_1 -0.345717  0.707713  0.207434 -1.667 0.0956
BKGRD1_C7_NOMISS_2  0.059193  1.060980  0.223767  0.265 0.7914
BKGRD1_C7_NOMISS_4  0.100840  1.106099  0.227257  0.444 0.6572
BKGRD1_C7_NOMISS_5 -0.450973  0.637008  0.243999 -1.848 0.0646
```

```
BKGRD1_C7_NOMISS_6  0.240836  1.272312  0.247358  0.974 0.3302
EDUCATION_C3_V1_1   0.119106  1.126489  0.112252  1.061 0.2887
EDUCATION_C3_V1_2   0.101482  1.106811  0.102893  0.986 0.3240
INCOME_C3_V1_1      0.184609  1.202748  0.182626  1.011 0.3121
INCOME_C3_V1_2      0.029187  1.029617  0.191390  0.152 0.8788

Likelihood ratio test=  on 16 df, p=
n= 8938, number of events= 1460
```

Note that these results are identical to those from other software with Efron tie handling methods.

5.2.4. Stata

Cox regression can be fit using the *stcox* command with the usual syntax. First, we specify DIABETES5_INDICATOR_V2 as the event indicator, and DIABETES5_TIME_V2 as the observed event time in the *stset* command. The prefix *svy* is then used with the *stcox* command to ensure that the Cox regression accounts for the complex survey procedures specified using the *svyset* command. Domain variable KEEP_DATA_DIABETES5 is specified in the 'subpop' option before the *stcox* command.

By default, *stcox* will set the smallest numerical level of each of the class variables as the reference level. In order to change the reference level of a class variable in this procedure, invoke 'ib' option.

By default, *stcox* will output estimated hazard ratios, but 'nohr' option can be invoked to output coefficient estimates instead. Breslow method is the default ties handling method, and Efron method is not supported with weights, specified in the 'pw' option.

```
svyset psu_id [pw=weight_norm_overall_v2], strata(strat)

stset diabetes5_time_v2, failure(diabetes5_indicator_v2)

svy, subpop(keep_data_diabetes5): stcox cesd10_v1 age_v1 ib4.centernum ib1.gendernum ib3.bkgrd1_c7_nomiss
ib3.education_c3_v1 ib3.income_c3_v1, nohr
* ref: San Diego, Male, Mexicans
```

```
      pweight: weight_norm_overall_v2
          VCE: linearized
  Single unit: missing
     Strata 1: strat
         SU 1: psu_id
        FPC 1: <zero>

     failure event:  diabetes5_indicator_v2 != 0 & diabetes5_indicator_v2 < .
obs. time interval:  (0, diabetes5_time_v2]
 exit on or before:  failure

------------------------------------------------------------------------------
```

```
     11,623  total observations
          5  event time missing (diabetes5_time_v2>=.)          PROBABLE ERROR
      2,541  observations end on or before enter()
------------------------------------------------------------------------------
      9,077  observations remaining, representing
      1,488  failures in single-record/single-failure data
   19223924  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =     3,506

Survey: Cox regression

Number of strata   =        20                  Number of obs     =     11,618
Number of PSUs     =       652                  Population size   = 11,619.054
                                                Subpop. no. obs   =      8,938
                                                Subpop. size      = 9,593.2228
                                                Design df         =        632
                                                F(  16,    617)   =      25.64
                                                Prob > F          =     0.0000
```

Specifying ‘nohr’ option for coefficient estimates:

```
------------------------------------------------------------------------------
             |             Linearized
          _t |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   cesd10_v1 |   .0157209   .0071584     2.20   0.028     .0016636    .0297781
      age_v1 |   .0429378   .0028631    15.00   0.000     .0373154    .0485602
             |
   centernum |
           1 |  -.2918285   .2450266    -1.19   0.234    -.7729932    .1893362
           2 |  -.1107734   .1438666    -0.77   0.442    -.3932878    .1717411
           3 |  -.4416158    .235886    -1.87   0.062    -.9048309    .0215993
             |
 0.gendernum |   .0232439   .0813361     0.29   0.775    -.1364779    .1829657
             |
bkgrd1_c7_~s |
           0 |  -.1348145   .2509326    -0.54   0.591    -.6275769     .357948
           1 |  -.3456848    .207336    -1.67   0.096    -.7528357     .061466
           2 |   .0591257   .2236318     0.26   0.792    -.3800256    .4982769
           4 |   .1007617   .2269496     0.44   0.657    -.3449047    .5464282
           5 |  -.4509749   .2439043    -1.85   0.065    -.9299358    .0279859
           6 |   .2408288   .2472516     0.97   0.330    -.2447052    .7263628
             |
education_~1 |
           1 |   .1191234   .1122097     1.06   0.289    -.1012255    .3394723
           2 |   .1014503   .1028365     0.99   0.324    -.1004924    .3033929
             |
income_c3_v1 |
           1 |   .1847493   .1825061     1.01   0.312    -.1736424     .543141
           2 |   .0293787   .1912598     0.15   0.878     -.346203    .4049603
------------------------------------------------------------------------------
```

Default option for hazard ratios:

_t	Haz. Ratio	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
cesd10_v1	1.015845	.0072719	2.20	0.028	1.001665	1.030226
age_v1	1.043873	.0029888	15.00	0.000	1.03802	1.049759
centernum						
1	.7468966	.1830095	-1.19	0.234	.4616293	1.208447
2	.8951416	.128781	-0.77	0.442	.6748345	1.18737
3	.6429966	.1516739	-1.87	0.062	.4046103	1.021834
0.gendernum	1.023516	.0832488	0.29	0.775	.8724256	1.200773
bkgrd1_c7_~s						
0	.873878	.2192844	-0.54	0.591	.5338839	1.430391
1	.7077355	.1467391	-1.67	0.096	.471029	1.063394
2	1.060909	.2372529	0.26	0.792	.6838439	1.645883
4	1.106013	.2510092	0.44	0.657	.7082879	1.727073
5	.6370068	.1553687	-1.85	0.065	.394579	1.028381
6	1.272303	.314579	0.97	0.330	.7829353	2.067547
education_~1						
1	1.126509	.1264052	1.06	0.289	.9037292	1.404206
2	1.106775	.1138169	0.99	0.324	.904392	1.354447
income_c3_v1						
1	1.202917	.2195396	1.01	0.312	.8405974	1.721405
2	1.029814	.1969621	0.15	0.878	.7073689	1.499243

Note that these results are identical to those from other software with Breslow tie handling methods.

5.2.5. Mplus

The *ANALYSIS: TYPE = COMPLEX* statement in Mplus is invoked to fit Cox regression model using complex survey procedures. Design variables are specified through the statements *STRAT*, *CLUSTER*, and *WEIGHT*. Indicator variables are created with desired reference levels and used in model fitting because Mplus cannot specify class variables directly. Since variable names in Mplus cannot exceed 8 characters, they need to be renamed prior to input to avoid truncations.

Domain variable KEEP_DATA_DIABETES5 (renamed to KEEP_DATA) is specified in the *SUBPOPULATION* statement, with ‘EQ 1’ indicating subpopulation of interest. DIABETES5_INDICATOR_V2 (renamed to dm5_ind) as the event indicator is specified through the *TIMECENSORED* statement, with ‘(1 = NOT 0 = RIGHT)’ indicating censoring value. DIABETES5_TIME_V2 (renamed to dm5_time) is modelled through the *MODEL:* statement as the observed event time.

Mplus documentation does not specify which method uses to handle ties. By comparing Mplus output with other software output, we observe that *ANALYSIS: TYPE = COMPLEX* uses the Breslow method. Other tie handling methods are not supported.

By default, *ANALYSIS: TYPE = COMPLEX* will output coefficient estimates with 3 decimal places. More decimal places can only be viewed by saving the output as a text file (named as "estimates.dat" in the example code) through the *savedata* statement, and invoking the *format* statement. Hazard ratio estimates are not supported.

```
DATA:
FILE IS sol.csv;

! variables in the same order of as created in the dataset;
VARIABLE:
NAMES = dm5_time dm5_ind weight PSU_ID STRAT keep_data CESD10_V1 AGE_V1 center_1
center_2 center_3 gender_0 bkgrd_0 bkgrd_1 bkgrd_2 bkgrd_4 bkgrd_5 bkgrd_6 edu_1 edu_2 income_1 income_2;

! specify what variables we need to use in the analysis;
USEVARIABLES = dm5_time dm5_ind weight PSU_ID STRAT keep_data CESD10_V1 AGE_V1 center_1
center_2 center_3 gender_0 bkgrd_1 bkgrd_2 bkgrd_3 bkgrd_4 bkgrd_5 bkgrd_6 edu_1 edu_2 income_1 income_2;

! specify design features;
SUBPOPULATION = keep_data EQ 1;
CLUSTER = PSU_ID;
STRAT = STRAT;
WEIGHT = weight;
SURVIVAL = dm5_time;

! event indicator;
TIMECENSORED = dm5_ind (1 = NOT 0 = RIGHT);

! survey method used;
ANALYSIS:
TYPE = COMPLEX;

!specify the model;
MODEL:
dm5_time on CESD10_V1 AGE_V1 center_1 center_2 center_3 gender_0 bkgrd_1 bkgrd_2 bkgrd_3 bkgrd_4
bkgrd_5 bkgrd_6 edu_1 edu_2 income_1 income_2;

! save the output as a text file to view more decimal places in estimates;
savedata:
format is f10.5;
results are estimates.dat;
```

```
SUMMARY OF ANALYSIS

Number of groups                                 1
Number of observations                        8938

Number of dependent variables                    1
Number of independent variables                 16
Number of continuous latent variables            0
```

```
MODEL RESULTS

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

DM5_TIME    ON
    CESD10_V1          0.016      0.007      2.196      0.028
    AGE_V1             0.043      0.003     14.997      0.000
    CENTER_1          -0.292      0.245     -1.191      0.234
    CENTER_2          -0.111      0.144     -0.770      0.441
    CENTER_3          -0.442      0.236     -1.872      0.061
    GENDER_1          -0.023      0.081     -0.286      0.775
    BKGRD_0           -0.135      0.251     -0.537      0.591
    BKGRD_1           -0.346      0.207     -1.667      0.095
    BKGRD_2            0.059      0.224      0.264      0.791
    BKGRD_4            0.101      0.227      0.444      0.657
    BKGRD_5           -0.451      0.244     -1.849      0.064
    BKGRD_6            0.241      0.247      0.974      0.330
    EDU_1              0.119      0.112      1.062      0.288
    EDU_2              0.101      0.103      0.986      0.324
    INCOME_1           0.185      0.182      1.012      0.311
    INCOME_2           0.029      0.191      0.153      0.878
```

Estimates with more decimal places in estimates.dat:

```
   Estimate                S.E.

0.15720530E-01         0.71584377E-02
0.42937477E-01         0.28630898E-02
-0.29183347E+00        0.24502836E+00
-0.11077819E+00        0.14386542E+00
-0.44162171E+00        0.23588471E+00
-0.23246485E-01        0.81336166E-01
-0.13481485E+00        0.25093602E+00
-0.34568652E+00        0.20733715E+00
0.59126878E-01         0.22363073E+00
0.10076388E+00         0.22694838E+00
-0.45097420E+00        0.24390239E+00
0.24082629E+00         0.24725163E+00
0.11911927E+00         0.11220726E+00
0.10144444E+00         0.10283533E+00
0.18471550E+00         0.18249588E+00
0.29342200E-01         0.19124868E+00
```

Note that these results are almost identical to those from other software with Breslow tie handling methods.

5.3. Weighted Regression Analysis Approach for Cox Regression Model

In this sub-section, we use diabetes incidence based on Definition 5 as an example to illustrate the weighted regression approach that also account for clustering at the PSU level for fitting Cox regression model. Specifically, we present examples and sample codes using SAS, R, and STATA for such analysis. The weighted approach uses Visit 2 sampling weights (WEIGHT_NORM_OVERALL_V2) as weights and account for clustering on the PSU_ID level in the data.

5.3.1. SAS

The procedure PHREG is used to produce estimates for Cox regression model using the weighted regression analysis approach while accounting for clustering on the PSU level. KEEP_DATA_DIABETES5 is specified through the *where* statement to select the subpopulation of interest. The clustering variable PSU_ID is specified through the *id* statement, and the "covs(aggregate)" option is specified to request the corresponding robust sandwich estimate for output and testing. Sampling weights are used in the *weight* statement. The *class* statement and the *model* statement are the same as the ones presented in 5.2.1. The default reference levels and ties handling method are the same as the SURVEYPHREG procedure.

```
proc phreg data = sol covs(aggregate); /* DEFAULT: order=formatted */
   where KEEP_DATA_DIABETES5 = 1;
   id PSU_ID;
   weight WEIGHT_NORM_OVERALL_V2;
   class CENTERNUM GENDERNUM BKGRD1_C7_NOMISS(ref = '3') EDUCATION_C3_V1
INCOME_C3_V1; /* ref: San Diego, Male, Mexicans */
   model DIABETES5_TIME_V2*DIABETES5_INDICATOR_V2(0)= CESD10_V1 AGE_V1
      CENTERNUM GENDERNUM BKGRD1_C7_NOMISS EDUCATION_C3_V1 INCOME_C3_V1 /
      ties = efron; /* DEFAULT: ties = breslow */
run;
```

Efron tie handling results:

Model Information

Data Set	WORK.SOL
Dependent Variable	DIABETES5_TIME_V2
Censoring Variable	DIABETES5_INDICATOR_V2
Censoring Value(s)	0
Weight Variable	WEIGHT_NORM_OVERALL_V2
Ties Handling	EFRON

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
8938	1460	7478	83.67

Parameter		DF	Parameter Estimate	Standard Error	StdErr Ratio	Chi-Square	Pr > ChiSq	Hazard Ratio
CESD10_V1		1	0.01572	0.00713	1.468	4.8659	0.0274	1.016
AGE_V1		1	0.04294	0.00288	1.417	222.0823	<.0001	1.044
CENTERNUM	**B**	1	-0.29196	0.24557	2.063	1.4135	0.2345	0.747
CENTERNUM	**C**	1	-0.11092	0.14421	1.525	0.5916	0.4418	0.895
CENTERNUM	**M**	1	-0.44176	0.23716	1.514	3.4695	0.0625	0.643
GENDERNUM	**F**	1	0.02318	0.08096	1.378	0.0820	0.7746	1.023
BKGRD1_C7_NOMISS	**0**	1	-0.13484	0.25167	1.680	0.2871	0.5921	0.874
BKGRD1_C7_NOMISS	**1**	1	-0.34571	0.20690	1.200	2.7919	0.0947	0.708
BKGRD1_C7_NOMISS	**2**	1	0.05919	0.22434	1.372	0.0696	0.7919	1.061
BKGRD1_C7_NOMISS	**4**	1	0.10084	0.23044	2.021	0.1915	0.6617	1.106
BKGRD1_C7_NOMISS	**5**	1	-0.45097	0.24466	1.281	3.3977	0.0653	0.637
BKGRD1_C7_NOMISS	**6**	1	0.24084	0.24701	1.634	0.9507	0.3296	1.272
EDUCATION_C3_V1	**1**	1	0.11911	0.11207	1.554	1.1294	0.2879	1.126
EDUCATION_C3_V1	**2**	1	0.10148	0.10285	1.386	0.9735	0.3238	1.107
INCOME_C3_V1	**1**	1	0.18453	0.18124	1.308	1.0367	0.3086	1.203
INCOME_C3_V1	**2**	1	0.02911	0.19077	1.300	0.0233	0.8787	1.030

Breslow tie handling results:

Model Information

Data Set	WORK.SOL
Dependent Variable	DIABETES5_TIME_V2
Censoring Variable	DIABETES5_INDICATOR_V2
Censoring Value(s)	0
Weight Variable	WEIGHT_NORM_OVERALL_V2
Ties Handling	BRESLOW

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
8938	1460	7478	83.67

Parameter		DF	Parameter Estimate	Standard Error	StdErr Ratio	Chi-Square	Pr > ChiSq	Hazard Ratio
CESD10_V1		1	0.01572	0.00712	1.467	4.8727	0.0273	1.016
AGE_V1		1	0.04294	0.00288	1.416	222.3618	<.0001	1.044
CENTERNUM	**B**	1	-0.29183	0.24529	2.061	1.4155	0.2341	0.747
CENTERNUM	**C**	1	-0.11077	0.14408	1.524	0.5911	0.4420	0.895
CENTERNUM	**M**	1	-0.44161	0.23699	1.512	3.4724	0.0624	0.643
GENDERNUM	**F**	1	0.02324	0.08092	1.378	0.0825	0.7739	1.024
BKGRD1_C7_NOMISS	**0**	1	-0.13482	0.25147	1.678	0.2874	0.5919	0.874
BKGRD1_C7_NOMISS	**1**	1	-0.34568	0.20680	1.199	2.7940	0.0946	0.708
BKGRD1_C7_NOMISS	**2**	1	0.05912	0.22420	1.371	0.0695	0.7920	1.061
BKGRD1_C7_NOMISS	**4**	1	0.10076	0.23013	2.019	0.1917	0.6615	1.106
BKGRD1_C7_NOMISS	**5**	1	-0.45098	0.24456	1.280	3.4004	0.0652	0.637
BKGRD1_C7_NOMISS	**6**	1	0.24083	0.24690	1.633	0.9514	0.3294	1.272
EDUCATION_C3_V1	**1**	1	0.11912	0.11203	1.553	1.1306	0.2876	1.127
EDUCATION_C3_V1	**2**	1	0.10145	0.10279	1.385	0.9740	0.3237	1.107
INCOME_C3_V1	**1**	1	0.18467	0.18112	1.307	1.0396	0.3079	1.203
INCOME_C3_V1	**2**	1	0.02930	0.19064	1.299	0.0236	0.8778	1.030

Note that Breslow and Efron methods provide very similar results. The results are similar to those from Cox regression with complex survey procedures.

5.3.2. R

The *coxph* function from R package “survival” is used to fit Cox regression model using weighted regression analysis approach while accounting for clustering on the PSU level. We use DIABETES5_INDICATOR_V2 as the event indicator (with ‘== 1’ specified as the event value), and DIABETES5_TIME_V2 as the observed event time. The clustering variable PSU_ID is specified by adding a “cluster(PSU_ID)” term in the model, which requests the corresponding robust sandwich estimate for output and testing. The “weights” option is set to be the sampling weights. The “subset” option is to select the subpopulation of interest, KEEP_DATA_DIABETES5 == 1.

Indicator variables are created with desired reference levels and used in model fitting with *coxph*, which cannot specify class variables. By default, *coxph* will use the Efron method to handle ties. The Breslow method can be invoked through the “ties” option.

```
coxph(Surv(DIABETES5_TIME_V2,DIABETES5_INDICATOR_V2==1) ~ CESD10_V1 +AGE_V1 +
CENTERNUM_1 + CENTERNUM_2 + CENTERNUM_3 + GENDERNUM_0+ BKGRD1_C7_NOMISS_0
+BKGRD1_C7_NOMISS_1+BKGRD1_C7_NOMISS_2+BKGRD1_C7_NOMISS_4+BKGRD1_C7_NOMISS_5+
BKGRD1_C7_NOMISS_6+EDUCATION_C3_V1_1+EDUCATION_C3_V1_2+INCOME_C3_V1_1+
INCOME_C3_V1_2 + cluster(PSU_ID), weights = WEIGHT_NORM_OVERALL_V2, subset =
(KEEP_DATA_DIABETES5 == 1), ties = c("breslow"), data = sol)
```

```
                       coef exp(coef)  se(coef) robust se      z      p
CESD10_V1           0.015721  1.015845  0.004856  0.007122  2.207 0.0273
AGE_V1              0.042938  1.043873  0.002033  0.002879 14.912 <2e-16
CENTERNUM_1        -0.291828  0.746897  0.119005  0.245287 -1.190 0.2341
CENTERNUM_2        -0.110773  0.895142  0.094556  0.144077 -0.769 0.4420
CENTERNUM_3        -0.441616  0.642997  0.156694  0.236991 -1.863 0.0624
GENDERNUM_0         0.023244  1.023516  0.058734  0.080918  0.287 0.7739
BKGRD1_C7_NOMISS_0 -0.134814  0.873878  0.149838  0.251472 -0.536 0.5919
BKGRD1_C7_NOMISS_1 -0.345685  0.707736  0.172451  0.206805 -1.672 0.0946
BKGRD1_C7_NOMISS_2  0.059126  1.060909  0.163508  0.224201  0.264 0.7920
BKGRD1_C7_NOMISS_4  0.100762  1.106013  0.114007  0.230134  0.438 0.6615
BKGRD1_C7_NOMISS_5 -0.450975  0.637007  0.191046  0.244559 -1.844 0.0652
BKGRD1_C7_NOMISS_6  0.240829  1.272303  0.151181  0.246900  0.975 0.3294
EDUCATION_C3_V1_1   0.119123  1.126509  0.072118  0.112030  1.063 0.2876
EDUCATION_C3_V1_2   0.101450  1.106775  0.074208  0.102794  0.987 0.3237
INCOME_C3_V1_1      0.184749  1.202917  0.138594  0.181128  1.020 0.3077
INCOME_C3_V1_2      0.029379  1.029814  0.146771  0.190650  0.154 0.8775

Likelihood ratio test=553.5  on 16 df, p=< 2.2e-16
n= 8938, number of events= 1460
```

Note that these results are almost identical to those from other software with Breslow tie handling methods, with differences due to rounding.

5.3.3. Stata

Estimates for the Cox regression model using weighted regression analysis approach can be obtained using the *stcox* command without the *svy* prefix, and clustering on the PSU level can be accounted for by specifying the clustering variable PSU_ID in the "vce(cluster)" option. KEEP_DATA_DIABETES5 is specified through the *drop* statement to select the subpopulation of interest. Sampling weights are specified through the "[pw =]" option in the *stset* command. Other statements and options specified are the same to the ones presented in 5.2.4.

```
drop if keep_data_diabetes5 ~= 1

stset diabetes5_time_v2 [pw=weight_norm_overall_v2], failure(diabetes5_indicator_v2)

stcox cesd10_v1 age_v1 ib4.centernum ib1.gendernum_v2 ib6.bkgrd1_c7_nomiss ib3.education_c3_v1
ib3.income_c3_v1, nohr vce(cluster psu_id)
```

```
     failure event:  diabetes5_indicator_v2 != 0 & diabetes5_indicator_v2 < .
obs. time interval:  (0, diabetes5_time_v2]
 exit on or before:  failure
            weight:  [pweight=weight_norm_overall_v2]

------------------------------------------------------------------------------
      8,938  total observations
          0  exclusions
------------------------------------------------------------------------------
      8,938  observations remaining, representing
      1,460  failures in single-record/single-failure data
   18925267  total analysis time at risk and under observation
                                             at risk from t =         0
                                  earliest observed entry t =         0
                                       last observed exit t =     3,506

         failure _d:  diabetes5_indicator_v2
   analysis time _t:  diabetes5_time_v2
            weight:  [pweight=weight_norm_overall_v2]

Cox regression -- Breslow method for ties

No. of subjects      =        9,593             Number of obs    =       8,938
No. of failures      =        1,224
Time at risk         =  20839006.87
                                               Wald chi2(16)    =      415.41
Log pseudolikelihood =   -10248.005             Prob > chi2      =      0.0000

                              (Std. Err. adjusted for 646 clusters in psu_id)
```

Specifying ‘nohr’ option for coefficient estimates:

```
------------------------------------------------------------------------------
             |               Robust
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   cesd10_v1 |   .0157209   .0071273     2.21   0.027     .0017516    .0296902
      age_v1 |   .0429378   .0028817    14.90   0.000     .0372898    .0485858
             |
   centernum |
           1 |  -.2918285   .2454766    -1.19   0.235    -.7729538    .1892968
           2 |  -.1107734   .1441883    -0.77   0.442    -.3933772    .1718305
           3 |  -.4416158   .2371742    -1.86   0.063    -.9064687     .023237
             |
 0.gendernum |   .0232439   .0809805     0.29   0.774     -.135475    .1819628
             |
bkgrd1_c7_~s |
           0 |  -.1348145   .2516669    -0.54   0.592    -.6280725    .3584436
           1 |  -.3456848   .2069654    -1.67   0.095    -.7513296      .05996
           2 |   .0591257   .2243743     0.26   0.792    -.3806399    .4988913
           4 |   .1007617   .2303126     0.44   0.662    -.3506427    .5521662
           5 |  -.4509749   .2447488    -1.84   0.065    -.9306738    .0287239
           6 |   .2408288   .2470915     0.97   0.330    -.2434617    .7251193
             |
education_~1 |
           1 |   .1191234   .1121169     1.06   0.288    -.1006218    .3388685
           2 |   .1014503   .1028738     0.99   0.324    -.1001786    .3030791
             |
income_c3_v1 |
           1 |   .1847493   .1812686     1.02   0.308    -.1705306    .5400292
           2 |   .0293787   .1907982     0.15   0.878     -.344579    .4033363
------------------------------------------------------------------------------
```

Default option for hazard ratios:

_t	Haz. Ratio	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
cesd10_v1	1.015845	.0072403	2.21	0.027	1.001753	1.030135
age_v1	1.043873	.0030081	14.90	0.000	1.037994	1.049785
centernum						
1	.7468966	.1833456	-1.19	0.235	.4616475	1.2084
2	.8951416	.1290689	-0.77	0.442	.6747742	1.187476
3	.6429966	.1525022	-1.86	0.063	.4039482	1.023509
0.gendernum	1.023516	.0828849	0.29	0.774	.873301	1.19957
bkgrd1_c7_~s						
0	.873878	.2199262	-0.54	0.592	.5336194	1.4311
1	.7077355	.1464768	-1.67	0.095	.4717389	1.061794
2	1.060909	.2380406	0.26	0.792	.6834239	1.646894
4	1.106013	.2547288	0.44	0.662	.7042353	1.737012
5	.6370068	.1559067	-1.84	0.065	.3942879	1.02914
6	1.272303	.3143754	0.97	0.330	.7839095	2.064977
education_~1						
1	1.126509	.1263007	1.06	0.288	.904275	1.403359
2	1.106775	.1138581	0.99	0.324	.9046758	1.354022
income_c3_v1						
1	1.202917	.218051	1.02	0.308	.8432173	1.716057
2	1.029814	.1964868	0.15	0.878	.7085186	1.49681

Note that these results are almost identical to those from other software with Breslow tie handling methods, with differences due to rounding.

REFERENCES

Therneau, T., Atkinson, B., Ripley, B., & Ripley, M. B. (2018). Package 'rpart'. Available online: https://cran.r-project.org/web/packages/rpart/rpart.pdf (accessed on 20 March 2018).

LaVange LM, & Kalsbeek W, et. Al. (2010) Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. Annals of Epidemiology; 20(8): 642-649.

Schneiderman N, Llabre M, Cowie CC, Barnhart J, Carnethon M, Gallo LC, Giachello AL, Heiss G, Kaplan RC, LaVange LM, Teng Y, Villa-Caballero L, Avilés-Santa ML. Prevalence of diabetes among Hispanics/Latinos from diverse backgrounds: the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). Diabetes Care. 2014 Aug;37(8):2233-9.