HCHS/SOL Analysis Methods - Visit 2

August 2022

Version 3.1

Prepared by the

HCHS/SOL Coordinating Center

Collaborative Studies Coordinating Center
UNC Department of Biostatistics

Jianwen Cai
Daniela Sotres-Alvarez
Donglin Zeng
Beibo Zhao, Leo Li, Wenyi Xie
Franklyn Gonzalez II
Marston Youngblood
Former contributors: Pedro Baldoni, Nicole Butera

This document is CONFIDENTIAL and for EXCLUSIVE use by HCHS/SOL investigators and NHLBI-NIH. Its purpose is to illustrate methods not to report results. Please send questions, suggestions and comments to fgonzale@email.unc.edu

Table of Contents

Table of Contents

Table of Contents

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 data from 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, sex, and Hispanic/Latino background distributions from the 2010 US Census for the four study field centers based on participants' visit 1 age. HCHS/SOL Analyses Methods at Baseline describe differences between these and their proper use.
- The document is not intended for direct citation.
- The document uses variable called GENDER (at baseline), but it refers to SEX.
- 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

MINOR Update in Version 3.1 (August 2022)

- Chapter 5: Corrected HOUSEHOLD_ID variable name to be HH_ID.

MAIN Updates in Version 3.0 (June 2022)

- Used data from most updated baseline file PART_DERV_INV4 (Jan 2020; N=16,415). No update to V2 file.
- Chapter 1 is updated by adding four subsections that introduce the following chapters: 1.2) multilevel sampling weights; 1.4) design-based and model-based procedures; 1.5) marginal (Generalized Estimating Equation) and conditional (Multilevel Modelling) Approaches; 1.6) analytic dataset.
- Chapters 2, 3, and 4 are updated to adjust for center in all analysis.
- Chapters 2, 3, and 4 now include model-based procedures.

- Chapters 3 and 4 are updated to use INCIDENT_CKD_V1V2 instead of CKD2_V2 as the outcome variable to incorporate a GFR decreasing rate of 1+ ml/min/1.73 m^2 per year.
- NEW Chapter 5 illustrates how to analyze change in continuous outcomes with multilevel modelling, using BMI_V2V1 as an example.
- Chapter 6 (current) was Chapter 5 from the Version 2.0 update (July 2020).
- Chapter 6 is updated to include the model-based Kaplan–Meier estimator in section 6.2.

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), the study design specifications are accounted for 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 focuses on analysis methods for longitudinal data with two visits.** Specifically, it provides guidelines for analyzing changes in continuous measures using linear regression models, binary incident disease/conditions using logistic regression and Poisson regression models, changes in continuous measures using multilevel modelling to estimate conditional effects and variance components, and right censored incident event time data using the Kaplan–Meier estimator and Cox regression models. For these analyses, examples with SAS/SUDAAN/R/Stata/Mplus program codes and the corresponding results are presented using Body Mass Index (BMI) as a continuous outcome, incidence of Chronic Kidney Disease (CKD) as a discrete outcome, and right censored Diabetes incident event time as a survival outcome. We present both design-based complex survey procedures and model-based non-survey procedures, which are viable alternatives to the design-based counterparts. HCHS/SOL study design specifications are included in all the analyses presented. Sampling weights, both overall and multilevel, are adjusted for non-response to visits 1 and 2, trimmed, calibrated to the age, sex, and Hispanic/Latino background distributions from the 2010 US Census for the four study field centers based on participants' visit 1 age, and normalized.

1.1. Visit 2 Sampling Weights

The HCHS/SOL cohort at baseline was selected through a stratified multi-stage probability sampling design. Briefly, at the 1st stage, the Primary Sampling Units (PSUs) were the census block groups (BGs) and were selected with simple random sampling (SRS) at each field center, stratified by cross-classification of 2000 Census high/low socioeconomic status and high/low Hispanic/Latino concentration. At the 2nd stage, the Secondary Sampling Units (SSUs) were the households (HHs) and were selected with SRS in each of the sampled PSUs, stratified by having or not Hispanic/Latino surname from postal addresses purchased from Genesys. Households with Hispanic/Latino surname were oversampled. Lastly, at the 3rd stage, subjects (SUBs) were selected in each of the eligible sampled SSUs. Subjects aged 45-74 years were over-sampled. Therefore, in the final HCHS/SOL cohort, subjects are nested within household clusters, which are further nested within block group clusters with unequal probabilities of selection of BGs, HHs, and SUBs at their respective levels by this sampling design.

As in any complex survey design, and as was done for the HCHS/SOL baseline (visit 1), sampling weights account for non-response. 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 non-response for visit 2.

To identify baseline factors that are associated with the probability of attending visit 2, a classification tree approach (R package *rpart*) was used. The advantage of the classification tree approach is that it takes 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, Sex, 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, Sex, Strata, and Education to be associated with the probability of returning to visit 2.

Visit 1 non-response adjustment was stratified on field center, sex and 6-level age groups. Based on the classification tree results for visit 2 non-response adjustment and building on the strata formed by field center, sex 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), HCHS/SOL strata, and education. The smallest number of participants in strata formed by field center, sex and age groups is 90, hence we required the number of participants to be at least 90 to form a finer stratum to obtain a reliable non-response rate. The non-response rate for visit 2 is then calculated for each stratum. The sampling weights are calculated based on visit 1 non-response adjusted sampling weights and these non-response rates for visit 2. The sampling weights are then trimmed, calibrated to the age, sex, and Hispanic/Latino background distributions from the 2010 US Census for the four study field centers based on participants' visit 1 age, and normalized to the overall sample (WEIGHT_NORM_OVERALL_V2).

1.2. Visit 2 Multilevel Sampling Weights for Multilevel Modelling

Multilevel data have several levels of clustering. Multilevel modelling (MLM), presented in Chapter 5, requires specifying a sampling weight for each level of data to reflect unequal probabilities of selection at each clustering level and the individual level. In HCHS/SOL, depending on investigators' goals, data can be treated as having either three levels (BG, HH, SUB) with two-level nested clusters (BG and HH), or two levels (HH and SUB) with one-level cluster (HH). The latter is considered in this Guide because many survey studies do not release information on the PSUs. In addition, estimating the variance between PSUs is typically not of substantive interest to investigators and only treated as nuisance.

The steps to derive the multilevel sampling weights for visit 2 for each level of data are similar to the ones to derive V2 overall sampling weights. They are the product of a base weight for visit 1 and four adjustments. The multilevel base weights at each level are calculated as the inverse of the sampling probabilities for that level. Then, the four adjustments are done in the following order: 1) multiplicative adjustment for differential non-response. Basically, the response rate proportions are applied to the SUB and HH level base weights as multiplicative adjustment factors for their non-response adjusted weights, respectively; 2) cumulative trimming by 98th percentile to handle extreme weight values at each level of data. More specifically, extreme values in BG level base weights (if considered), in HH non-response adjusted weights, and in SUB non-response adjusted weights; 3) multiplicative adjustment to calibrate non-response adjusted and trimmed SUB level weights to the age, sex, and Hispanic/Latino background distributions from the 2010 US Census in the HCHS/SOL target area, for the four study field centers based on participants' visit 1 age; and 4) multiplicative adjustment to scale, i.e., normalize, weights at each level of data by effective cluster size to avoid bias of variance component estimation (Rabe-Hesketh and Skrondal 2006).

The details for calculating the multilevel sampling weights will be provided in a separate document. The levels, clusters, and corresponding multilevel sampling weights are summarized in the table below.

Levels	Clusters	Multilevel sampling weights needed
BG, HH, SUB	BG, HH	WEIGHT_3MLM_BG_V2 WEIGHT_3MLM_HH_V2 WEIGHT_MLM_SUB_V2
HH, SUB	HH	WEIGHT_2MLM_HH_V2 WEIGHT_MLM_SUB_V2

Note that the multilevel sampling weights are the inverse sampling probability at each level, with probability of selection conditional on higher levels. Therefore, **the SUB level weights WEIGHT_MLM_SUB_V2 are *conditional* weights, which are different from the visit 2 sampling weights, WEIGHT_NORM_OVERALL_V2, which were based on the participant's *unconditional* selection probability. The multilevel weights are specifically developed for MLM (Chapter 5) and should not be used in other settings.** Likewise, it is also inappropriate to substitute multilevel weights with the overall weights in MLM.

1.3. 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 for Visit 1 Data)				Visit 2 Target Population (N=11623 for Visit 2 Data)				
	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
Sex (%)										
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 for Visit 1 Data)				Visit 2 Target Population (N=11623 for Visit 2 Data)				
	N	Mean or %	Low 95%	Up 95%	N	Mean or %	Low 95%	Up 95%	Absolute Difference	Relative Difference
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 for Visit 1 Data)				Visit 2 Target Population (N=11623 for Visit 2 Data)				
	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).

1.4. Design-based Complex Survey Procedures and Model-Based Procedures

In all our analysis, we adopt the following perspective: observations are assumed to be sampled from a fixed finite population using a pre-specified sampling design, with the variation in the sample resulting from the randomness from sampling, instead of distributional assumption about the data-generating process (Sterba 2009). The values of variables of interest are treated as fixed in this finite population, and their inference considers the distribution of the estimator over repeated samples by using the same sampling design. For valid inference under this perspective, the sampling design (stratification, clustering and sampling weights) needs to be accounted for during the point and variance estimation of finite-population parameters. We call the analytic techniques that properly do so as design-based and refer to them with the general term **"complex survey procedures"**. However, for more complex models with longitudinal data or clustered data, such complex survey procedures either do not exist or have not been implemented in commercial software.

Simulation studies were conducted at the Coordinating Center to examine the prospect of using model-based non-survey procedures as alternatives to complex survey procedures for finite-population estimates. We use the general term “model-based procedures” to refer to these analytic techniques, as they stem from the different assumption that samples come from a hypothetical infinite population, and the observed values are regarded as realizations of the random variables that follow some distributions, i.e., model specifications. In this Guide, we only use the **model-based procedures** as tools to obtain finite-population estimates. Based on simulation results, which will be reported in a separate document, we present the model-based procedures that provide proper estimates and inference of the finite population parameters. These procedures use sampling weights to account for unequal probabilities of sampling, and robust variance estimation to account for intra-cluster correlation.

1.5. Marginal (Generalized Estimating Equation) and Conditional (Multilevel Modelling) Approaches

There are two modelling approaches that are commonly used to account for clustering in statistical analysis: population-averaged (marginal) and subject-specific (conditional).

Marginal methods describe linear relationships of a transformed mean response with the covariates without specifying the correlation structure for the responses within clusters. The coefficients (betas) of covariates have the interpretation of population-averaged effects; hence they are useful when one is interested in the covariate effects on the response but describing the amount of correlation of responses within clusters is not of particular interest. **Generalized estimating equation (GEE)** (Liang and Zeger 1986; Zeger, Liang, and Albert 1988) is an estimation method commonly used to estimate marginal effects, and it can be used in conjunction with intra-cluster robust variance estimation to account for the correlation within clusters. Various correlation matrix structures, termed working correlation matrix, can be used in GEE and a cluster-robust method is used for estimating the variance of the coefficient estimates. GEE provides asymptotically unbiased estimates and is robust against misspecification of the working correlation matrix. The cluster-robust variance estimator relaxes the assumption of independence of the subjects. In other words, subjects are still assumed to be independent across clusters, but not necessarily within clusters. Investigators can use this marginal modeling approach when interested in population-averaged effects and want a more robust method against model misspecification of the correlation structure. We refer to the software procedures that can provide estimation from GEE approach as model-based procedures. We will use the weight option in GEE procedures to accommodate sampling weights and refer to this approach as weighted GEE. In this Guide, we provide the sample code and finite-population estimates using model-based procedures of weighted GEE with intra-cluster robust variance estimation in different software, for linear regression (Chapter 2), logistic regression (Chapter 3), and Poisson regression (Chapter 4).

The conditional (mixed-effects) methods, on the other hand, model the transformed mean response conditional on the random effects for clusters, given the observed covariates. The random effects account for correlations within clusters and their distributions are usually specified, typically by the normal distribution. The coefficients (betas) of covariates have the interpretation of subject-specific effects because they are conditional on the subject’s random effects. As the result, they are useful when it is of interest to describe the correlation or variation of the responses within clusters. **Multilevel modelling (MLM)** is one approach for conditional modeling. The hierarchical nature of a multistage sampling design naturally corresponds to the hierarchical random-effects structure in the multilevel models. Random effects of the clustering levels are specified to address the cluster-specific variations, and survey sampling weights of the design are incorporated at each stage for valid inferences. Unlike GEE, MLM quantifies the covariates effects and variation of the effects through both fixed and random effects, allowing users to estimate within-cluster variability from the random-effects variance component. Like GEE, MLM can be used in conjunction with intra-robust cluster variance estimation. Investigators can use this method when interested in subject-specific effects as well as estimating correlation or variation within clusters. We refer to the software procedures that can provide estimation from the mixed effect approach as model-based procedures. We will use the weight option in the mixed effect procedures to accommodate sampling weights and refer to this approach as weighted MLM. In this Guide, we provide the sample program code and finite-population estimates from employing model-based procedures of weighted MLM with intra-cluster robust variance estimation in different software, for linear regression with either two or one random effect (Chapter 5).

The decision on which approach to use, marginal or conditional, depends on the research question of interest. Note that when the same set of covariates are included in the marginal model and fixed effect part of the conditional models, the coefficients of these covariates (betas) are identical for linear models, but not for non-linear models (Ritz and Spiegelman 2004).

1.6. Analytic Dataset

The following example code creates the analytic dataset SOL (N = 11,623) that will be used in examples throughout Chapters 2 to 5. It includes 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) for analysis with continuous outcomes; two binary flag variables KEEP_DATA (1 = those with non-missing BMI change between visit 1 and visit 2) and KEEP_DATA_CKD (1 = those without CKD at baseline), indicating the subpopulation of interest, for analysis with continuous and discrete measures, respectively.

```
data mylib.sol;
merge part_derv_inv4
part_derv_v2_inv3(keep=ID BMI_V2 YRS_BTWN_V1V2 CONSENT_V2
WEIGHT_NORM_OVERALL_V2 CKD2_V2 INCIDENT_CKD_V1V2 in=IN_V2)
        mlweights_v2_inv3; /* import multilevel weights */

by ID;

if IN_V2;

BMI_V2V1 = BMI_V2 - BMI;
RBMI_V2V1 = BMI_V2V1/YRS_BTWN_V1V2;

IF BKGRD1_C7<=.z THEN BKGRD1_C7=.;

KEEP_DATA = (BMI_V2V1 > .Z);

KEEP_DATA_CKD = (CKD2 = 0);

run;
```

Mplus data management: Indicator variables are created with desired reference levels and used in model fitting because Mplus cannot specify categorical variables in the models directly. Since variable names in Mplus cannot exceed 8 characters, they need to be renamed prior to input to avoid truncations. In addition, we use the value of 999 to represent missing values in covariates. These changes would be reflected in the results displayed. The following example code modifies the *SOL* dataset to *SOL_MPLUS* which will be used for examples with Mplus in Chapters 2, 3 and 4.

```
data mylib.sol_mplus;
set mylib.sol(keep = PSU_ID STRAT WEIGHT_NORM_OVERALL_V2
KEEP_DATA KEEP_DATA_CKD
RBMI_V2V1 BMI_V2V1 INCIDENT_CKD_V1V2
BKGRD1_C7 DIABETES2_INDICATOR
AGE BMI GENDERNUM CENTERNUM YRS_BTWN_V1V2 );

if GENDERNUM = 0 then gender_0 = 1; else gender_0 = 0;
if GENDERNUM = 1 then gender_1 = 1; else gender_1 = 0;
if CENTERNUM = 1 then center_1 = 1; else center_1 = 0;
if CENTERNUM = 2 then center_2 = 1; else center_2 = 0;
if CENTERNUM = 3 then center_3 = 1; else center_3 = 0;
if CENTERNUM = 4 then center_4 = 1; else center_4 = 0;

if BKGRD1_C7 = 0 then bkc7_0 = 1; else bkc7_0 = 0;
if BKGRD1_C7 = 1 then bkc7_1 = 1; else bkc7_1 = 0;
if BKGRD1_C7 = 2 then bkc7_2 = 1; else bkc7_2 = 0;
if BKGRD1_C7 = 3 then bkc7_3 = 1; else bkc7_3 = 0;
if BKGRD1_C7 = 4 then bkc7_4 = 1; else bkc7_4 = 0;
if BKGRD1_C7 = 5 then bkc7_5 = 1; else bkc7_5 = 0;
if BKGRD1_C7 = 6 then bkc7_6 = 1; else bkc7_6 = 0;


if DIABETES2_INDICATOR = 0 then d2_ind_0 = 1; else d2_ind_0 = 0;
if DIABETES2_INDICATOR = 1 then d2_ind_1 = 1; else d2_ind_1 = 0;

ly_v1v2 = log(YRS_BTWN_V1V2);


/* assign 999 as missing for Mplus */
if BMI = . then BMI = 999;
if BMI_V2V1 = . then BMI_V2V1 = 999;
      if RBMI_V2V1 = . then RBMI_V2V1 = 999;

if INCIDENT_CKD_V1V2 = . then INCIDENT_CKD_V1V2 = 999;
if BKGRD1_C7 = . then do;
bkc7_0 = 999; bkc7_1 = 999; bkc7_2 = 999; bkc7_3 = 999;
bkc7_4 = 999; bkc7_5 = 999; bkc7_6 = 999;
end;
if DIABETES2_INDICATOR = . then do;
d2_ind_0 = 999;
d2_ind_1 = 999;
end;

/* rename due to mplus character restriction */
rename
      YRS_BTWN_V1V2 = yrs_v1v2
      WEIGHT_NORM_OVERALL_V2 = weight_v2
KEEP_DATA_CKD = keep_ckd
      INCIDENT_CKD_V1V2 = ckd_v1v2
;

drop BKGRD1_C7 GENDERNUM CENTERNUM DIABETES2_INDICATOR;

run;
```

Case sensitivity: In R and Stata, variable names as well as commands are case-sensitive.

Disclaimer: The variable GENDER at baseline is an indication of biological sex, not self-identified gender.

2. Linear Regression Models for Change in Continuous Measures

The purpose of this chapter is to estimate change in continuous measures using linear regression models in SAS, SUDAAN, R, Stata, and Mplus. We present both design-based complex survey procedures and model-based non-survey procedures, when available in each software. See section 1.4 for a brief description of these procedures and their differences. 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 chapter, 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 sex (GENDER), field center (CENTERNUM), 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 visits 1 and 2 varies among participants, we will adjust for the time elapsed between visits (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.

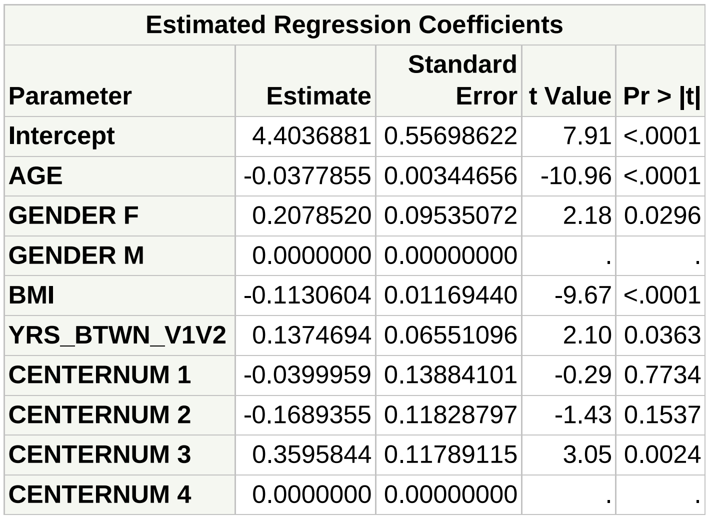
2.1.1. Complex Survey Procedures

In this section we fit a linear regression model for the difference in BMI between the first two visits (BMI_V2V1) using complex survey procedures in SAS, SUDAAN, R, Stata and Mplus. Note that the point estimates and robust standard error estimates are essentially identical among those from complex survey procedures in this section 2.1.1.

2.1.1.1. SAS

The procedure SURVEYREG is used to fit a linear regression 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 BKGRD1_C7, which will fit the model for each Hispanic/Latino background.

```
proc surveyreg data=worklib.sol; /* DEFAULT: order=formatted */
   strata STRAT; cluster PSU_ID; weight WEIGHT_NORM_OVERALL_V2;
   *domain BKGRD1_C7;
   class GENDER CENTERNUM;
   model BMI_V2V1 = AGE GENDER BMI YRS_BTWN_V1V2 CENTERNUM/ solution;
run;
```



Estimated Regression Coefficients				
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	4.4036881	0.55698622	7.91	<.0001
AGE	-0.0377855	0.00344656	-10.96	<.0001
GENDER F	0.2078520	0.09535072	2.18	0.0296
GENDER M	0.0000000	0.00000000	.	.
BMI	-0.1130604	0.01169440	-9.67	<.0001
YRS_BTWN_V1V2	0.1374694	0.06551096	2.10	0.0363
CENTERNUM 1	-0.0399959	0.13884101	-0.29	0.7734
CENTERNUM 2	-0.1689355	0.11828797	-1.43	0.1537
CENTERNUM 3	0.3595844	0.11789115	3.05	0.0024
CENTERNUM 4	0.0000000	0.00000000	.	.

This result indicates that after adjusting for sex, baseline BMI, center, and years elapsed between visits, a one-year increment in age at baseline is associated with a decrease of 0.0378 kg/m^2 in the change in BMI.

2.1.1.2. SUDAAN

In SUDAAN, PROC REGRESS is used to fit a linear regression model. 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 needs to specify an additional variable, for example, BKGRD1_C7=0, indicating the subpopulation “Dominican” 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 CENTERNUM;
   *subpopn BKGRD1_C7=0;
   model BMI_V2V1 = AGE GENDERNUM BMI YRS_BTWN_V1V2 CENTERNUM;
   reflevel GENDERNUM=1 CENTERNUM=4; /* reference: Male San Diego*/
   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                                                            P-value
  Variables and       Beta                   Lower 95%   Upper 95%             T-Test
  Effects             Coeff.        SE Beta  Limit Beta  Limit Beta  T-Test B=0  B=0
-------------------------------------------------------------------------------
Intercept                4.4037      0.5568      3.3102      5.4972     7.9084     0.0000
Age                     -0.0378      0.0034     -0.0446     -0.0310   -10.9670     0.0000
Gender (0=Female,
  1=Male)
  0                      0.2079      0.0953      0.0207      0.3950     2.1807     0.0296
  1                      0.0000      0.0000      0.0000      0.0000      .          .
BMI (kg/m2)             -0.1131      0.0117     -0.1360     -0.0901    -9.6700     0.0000
Elapsed time between
  visits 1 and 2
  (yrs)                  0.1375      0.0655      0.0089      0.2661     2.0991     0.0362
Participant's Field
  Center - numeric
  1                     -0.0400      0.1388     -0.3126      0.2326    -0.2881     0.7733
  2                     -0.1689      0.1183     -0.4013      0.0634    -1.4279     0.1538
  3                      0.3596      0.1179      0.1281      0.5910     3.0507     0.0024
  4                      0.0000      0.0000      0.0000      0.0000      .          .

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

2.1.1.3. R

In R, to fit linear regression models (or generalized linear models) 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.

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 BKGRD1_C7==0, indicating the subpopulation “Dominican” of interest. Finally, because we want to fit a linear regression model, we specify the Gaussian family with identity link through the ‘family’ argument. The *reflevel* function can be used to change the reference level of any “factor” (categorical) variables for all subsequent analyses.

```
sol$GENDER <- relevel(factor(sol$GENDER), ref='M')
sol$CENTER <- relevel(factor(sol$CENTER), ref='S')

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 + CENTER, design =
sol.design, #subset= BKGRD1_C7==0,# family=gaussian(link='identity'))

summary(model.diff)
Call:
svyglm(formula = BMI_V2V1 ~ AGE + GENDER + BMI + YRS_BTWN_V1V2 + CENTER, design = sol.design,
#subset= BKGRD1_C7==0,# 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)    4.403688   0.556835   7.908  1.2e-14 ***
AGE           -0.037786   0.003445 -10.967  < 2e-16 ***
GENDERF        0.207852   0.095314   2.181  0.02958 *
BMI           -0.113060   0.011692  -9.670  < 2e-16 ***
YRS_BTWN_V1V2  0.137469   0.065490   2.099  0.03621 *
CENTERB       -0.039996   0.138808  -0.288  0.77334
CENTERC       -0.168936   0.118309  -1.428  0.15382
CENTERM        0.359584   0.117871   3.051  0.00238 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 9.11461)
Number of Fisher Scoring iterations: 2
```

2.1.1.4. 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 is 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). If we are interested in making inference for a specific subpopulation, a domain variable BKGRD1_C7 is specified in the *subpop* option before the *regress* command (not shown in sample code), which will fit the model for each Hispanic/Latino background.

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.

```
fvset base last gendernum diabetes2_indicator bkgrd1_c7 centernum
svyset psu_id [pw=weight_norm_overall_v2], strata(strat)
svy: regress bmi_v2v1 age i.gendernum bmi yrs_btwn_v1v2 i.centernum
```

```
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(   7,    622)   =      58.92
                                              Prob > F          =     0.0000
                                              R-squared         =     0.0908
------------------------------------------------------------------------------
             |             Linearized
    bmi_v2v1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0377855   .0034455   -10.97   0.000    -.0445516   -.0310195
 0.gendernum |    .207852   .0953209     2.18   0.030     .0206656    .3950384
         bmi |  -.1130604   .0116908    -9.67   0.000    -.1360181   -.0901027
yrs_btwn_v~2 |   .1374693   .0654905     2.10   0.036     .0088625    .2660762
             |
   centernum |
          1  |  -.0399959   .1387977    -0.29   0.773    -.3125596    .2325678
          2  |  -.1689355    .118251    -1.43   0.154    -.4011509    .0632798
          3  |   .3595844   .1178543     3.05   0.002     .1281481    .5910207
             |
       _cons |   4.403688   .5568123     7.91   0.000     3.310249    5.497128
------------------------------------------------------------------------------
```

2.1.1.5. Mplus

In Mplus, the *ANALYSIS: TYPE = COMPLEX* statement is invoked to fit linear regression model using complex survey procedures. Design variables are specified through the statements *STRAT*, *CLUSTER*, and *WEIGHT*.

BMI_V2V1 is modelled through the *MODEL:* statement as the continuous outcome. If we are interested in making inference for a specific subpopulation, domain variable BKGRD1_C7 can be specified in the *SUBPOPULATION* statement, with ‘EQ 0’ indicating “Dominican” as the subpopulation of interest (not shown in sample code). Note that the statement *MISSING = ALL (999)* is invoked to indicate that the missing values in covariates are denoted with 999 in the input data.

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 “REGCOEFF.dat” in the example code) through the *savedata* statement and invoking the *format* statement.

```
! survey linear
DATA:
FILE IS sol_mplus.dat;
VARIABLE:
! variables in the same order of as created in the dataset;
NAMES = STRAT PSU_ID AGE BMI weight_v2 yrs_v1v2 ckd_v1v2
BMI_V2V1 RBMI_V2V1 KEEP_DATA keep_ckd gender_0 gender_1
center_1 center_2 center_3 center_4
bkc7_0 bkc7_1 bkc7_2 bkc7_3 bkc7_4 bkc7_5 bkc7_6 d2_ind_0 d2_ind_1 ly_v1v2;
! specify what variables we need to use in the analysis;
USEVARIABLES = STRAT PSU_ID AGE BMI weight_v2 yrs_v1v2 BMI_V2V1
gender_0 center_1 center_2 center_3;
! specify design features;
CLUSTER = PSU_ID;
STRAT = STRAT;
WEIGHT = weight_v2;
MISSING = ALL (999);
ANALYSIS:
! survey method used;
TYPE = COMPLEX;
ESTIMATOR=MLR;
!specify the model;
MODEL:
BMI_V2V1 on AGE gender_0 BMI yrs_v1v2 center_1 center_2 center_3;
SAVEDATA:
FORMAT IS f10.5;
RESULTS ARE Yourpath\regcoeff.dat;
```

```
SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                       11212

Number of dependent variables                                    1
Number of independent variables                                  7
Number of continuous latent variables                            0

MODEL RESULTS

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

 BMI_V2V1 ON
    AGE               -0.038      0.003    -10.967      0.000
    GENDER_0           0.208      0.095      2.180      0.029
    BMI               -0.113      0.012     -9.671      0.000
    YRS_V1V2           0.137      0.065      2.098      0.036
    CENTER_1          -0.040      0.139     -0.288      0.773
    CENTER_2          -0.169      0.118     -1.429      0.153
    CENTER_3           0.360      0.118      3.051      0.002

 Intercepts
    BMI_V2V1           4.404      0.557      7.910      0.000

 Residual Variances
    BMI_V2V1           9.189      0.388     23.677      0.000
```

2.1.2. Model-based Procedures

In this section we fit a linear regression model for the difference in BMI between visit 2 and visit 1 (BMI_V2V1) using the model-based procedure of weighted GEE with robust variance estimation that accounts for clustering within PSUs, instead of using complex survey procedures as done in previous section 2.1.1. See section 1.4 for a brief description of these procedures and their differences. We fit the model using SAS, R and Stata. Note that the point estimates are identical, and robust standard error estimates are the same up to the 2nd significant figure among those from model-based procedures in this section 2.1.2., and those from complex survey procedures in section 2.1.1.

2.1.2.1. SAS

In SAS, the procedure PROC GENMOD is used. The CLASS statement is used to specify categorical variables; the WEIGHT statement is used to specify the subject-level sampling weight; the MODEL statement is used to specify the analysis model; the DIST option is used to specify the marginal distribution of the outcome; and the cluster level, PSU_ID, and working correlation structure (independent, denoted with ind) are specified at SUBJECT and CORR options of the REPEATED statement to obtain robust variance estimation that accounts for clustering within PSUs.

```
proc genmod data=worklib.sol;
   class PSU_ID GENDER CENTERNUM; weight WEIGHT_NORM_OVERALL_V2;
   model BMI_V2V1 = AGE GENDER BMI YRS_BTWN_V1V2 CENTERNUM/ dist=normal;
   repeated subject=PSU_ID / corr=ind;
run;
```

The results from the model fitting are presented as follows:

Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

Parameter		**Estimate**	**Standard Error**	**95% Confidence Limits**		**Z**	**Pr > |Z|**
Intercept		4.4037	0.5569	3.3121	5.4953	7.91	<.0001
AGE		-0.0378	0.0035	-0.0446	-0.0310	-10.90	<.0001
GENDER	F	0.2079	0.0951	0.0214	0.3943	2.19	0.0289
GENDER	M	0.0000	0.0000	0.0000	0.0000	.	.
BMI		-0.1131	0.0117	-0.1359	-0.0902	-9.69	<.0001
YRS_BTWN_V1V2		0.1375	0.0657	0.0087	0.2662	2.09	0.0364
CENTERNUM	1	-0.0400	0.1384	-0.3113	0.2313	-0.29	0.7726
CENTERNUM	2	-0.1689	0.1210	-0.4062	0.0683	-1.40	0.1628
CENTERNUM	3	0.3596	0.1183	0.1278	0.5914	3.04	0.0024
CENTERNUM	4	0.0000	0.0000	0.0000	0.0000	.	.

This result indicates that, after adjusting for field center, sex, baseline BMI, and years elapsed between visits, a one-year increment in age at baseline is associated with a decrease of 0.0378 kg/m^2 in the change in BMI.

2.1.2.2. R

In R, the *geeglm* function from R package “*geepack*” is used. The “weights = WEIGHT_NORM_OVERALL_V2” statement indicates the subject-level sampling weight; the “data=sol”, “id=PSU_ID”, and “family=gaussian(link = "identity"))” statements indicate the working dataset, cluster level, and marginal distribution (and link function) of the outcome variable, respectively. The default of this function is to estimate robust variance.

```
model = geeglm(BMI_V2V1 ~ AGE + GENDER + BMI + YRS_BTWN_V1V2 + CENTER,
               weights = WEIGHT_NORM_OVERALL_V2,
               data=sol,
               id=PSU_ID,
               family=gaussian(link = "identity"))
summary(model)
```

```
Coefficients:
              Estimate Std.err   Wald Pr(>|W|)
(Intercept)     4.4037  0.5592  62.03  3.4e-15 ***
AGE            -0.0378  0.0033 130.74  < 2e-16 ***
GENDERF         0.2079  0.0912   5.20   0.0226 *
BMI            -0.1131  0.0118  92.59  < 2e-16 ***
YRS_BTWN_V1V2   0.1375  0.0668   4.23   0.0397 *
CENTERB        -0.0400  0.1323   0.09   0.7625
CENTERC        -0.1689  0.1134   2.22   0.1364
CENTERM         0.3596  0.1178   9.32   0.0023 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = independence
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)     9.19   0.376
Number of clusters:   11181  Maximum cluster size: 2
```

2.1.2.3. Stata

In Stata, the *meglm* command is used. The *pw* option is used to specify the subject-level sampling weight; the *vce* option is used to indicate the use of robust variance estimator; and *cluster psu_id* is used to specify the cluster level.

```
meglm bmi_v2v1 age ib1.gendernum ib4.centernum bmi yrs_btwn_v1v2 [pw=weight_norm_overall_v2],
vce(cluster psu_id)
```

```
Mixed-effects GLM                               Number of obs     =     11,212
Family:                Gaussian
```

```
Link:                 identity

                                                  Wald chi2(7)      =     414.26
Log pseudolikelihood = -28112.116                 Prob > chi2       =     0.0000
                               (Std. Err. adjusted for 648 clusters in psu_id)
------------------------------------------------------------------------------
             |               Robust
    bmi_v2v1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0377855   .0034677   -10.90   0.000    -.0445821    -.030989
 0.gendernum |    .207852   .0951906     2.18   0.029     .0212817    .3944222
             |
   centernum |
          1  |  -.0399959   .1385082    -0.29   0.773    -.3114671    .2314752
          2  |  -.1689355   .1211311    -1.39   0.163    -.4063482    .0684771
          3  |   .3595844   .1183513     3.04   0.002       .12762    .5915487
             |
         bmi |  -.1130604   .0116739    -9.68   0.000    -.1359408     -.09018
yrs_btwn_v~2 |   .1374693   .0657447     2.09   0.037     .0086121    .2663266
       _cons |   4.403688   .5573628     7.90   0.000     3.311277    5.496099
-------------+----------------------------------------------------------------
var(e.bmi_~1)|   9.188677   .3979405                      8.440913    10.00268
------------------------------------------------------------------------------
```

2.2. Linear Regression Model for the Rate of Change

In this section, the outcome of interest is the rate of change in BMI between the first two visits, denoted as RBMI_V2V1 and defined as the BMI change between visits 1 and 2, BMI_V2V1, divided by the time in years between the two visits (YRS_BTWN_V1V2). The rate of change, which is an annual 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. Complex Survey Procedures

In this section we fit a linear regression model for the rate of change in BMI between the first two visits (RBMI_V2V1) using complex survey procedures in SAS, SUDAAN, R, Stata and Mplus. Note that the point estimates and robust standard error estimates are essentially identical among those from complex survey procedures in this section 2.2.1.

2.2.1.1. SAS

The code provided below invokes the procedure SURVEYREG in SAS to fit a linear model for the rate of change in BMI between the first two visits (RBMI_V2V1). As before, statements and options specified are the same to the ones presented in section 2.1.1.1. 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 BKGRD1_C7;
   class GENDER CENTERNUM;
   model RBMI_V2V1 = AGE GENDER BMI CENTERNUM/ solution;
run;
```

Estimated Regression Coefficients				
Parameter	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	0.8688034	0.04823603	18.01	<.0001
AGE	-0.0062511	0.00056497	-11.06	<.0001
GENDER F	0.0335139	0.01556367	2.15	0.0317
GENDER M	0.0000000	0.00000000	.	.
BMI	-0.0187777	0.00185282	-10.13	<.0001
CENTERNUM 1	-0.0119286	0.02287691	-0.52	0.6023
CENTERNUM 2	-0.0270426	0.01911089	-1.42	0.1576
CENTERNUM 3	0.0551534	0.01878850	2.94	0.0035
CENTERNUM 4	0.0000000	0.00000000	.	.

The results indicate that, after adjusting for field center, sex, and baseline BMI, a one-year increment in age at baseline is associated with a decrease of 0.00625 kg/m^2 in the annual rate of change in BMI.

2.2.1.2. SUDAAN

The same model can be fitted in SUDDAN by invoking the procedure REGRESS. The same statements and options are used as in section 2.1.1.2.

```
proc regress data=worklib.sol filetype=sas design=wr notsorted;
   nest STRAT PSU_ID;
   weight WEIGHT_NORM_OVERALL_V2;   class GENDERNUM CENTERNUM;
   *subpopn BKGRD1_C7=0;
   model RBMI_V2V1 = AGE GENDERNUM BMI CENTERNUM;
   reflevel GENDERNUM=1 CENTERNUM=4; /* reference: Male San Diego*/
   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                                                                            P-value
  Variables and        Beta                     Lower 95%    Upper 95%                 T-Test
  Effects              Coeff.          SE Beta  Limit Beta   Limit Beta   T-Test B=0   B=0
-----------------------------------------------------------------------------------------------
Intercept                  0.8688       0.0482       0.7741       0.9635      18.0148      0.0000
Age                       -0.0063       0.0006      -0.0074      -0.0051     -11.0679      0.0000
Gender (0=Female,
  1=Male)
  0                        0.0335       0.0156       0.0030       0.0641       2.1541      0.0316
  1                        0.0000       0.0000       0.0000       0.0000        .            .
BMI (kg/m2)               -0.0188       0.0019      -0.0224      -0.0151     -10.1364      0.0000
Participant's Field
  Center - numeric
  1                       -0.0119       0.0229      -0.0568       0.0330      -0.5215      0.6022
  2                       -0.0270       0.0191      -0.0646       0.0105      -1.4148      0.1576
  3                        0.0552       0.0188       0.0183       0.0920       2.9359      0.0034
  4                        0.0000       0.0000       0.0000       0.0000        .            .
-----------------------------------------------------------------------------------------------
```

2.2.1.3. R

The R code provided below fits a linear model for the rate of change in BMI between the first two visits (RBMI_V2V1), and uses the survey design element ‘sol.design’ created in section 2.1.1.3.

```
> model.rdif = svyglm(RBMI_V2V1 ~ AGE + GENDER + BMI + CENTER, design = sol.design, #subset=
BKGRD1_C7==0,# family=gaussian(link='identity'))

> summary(model.rdif)

Call:
svyglm(formula = RBMI_V2V1 ~ AGE + GENDER + BMI + CENTER, design = sol.design, #subset=
BKGRD1_C7==0,# 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)  0.8688034  0.0482273  18.015  < 2e-16 ***
AGE         -0.0062511  0.0005648 -11.068  < 2e-16 ***
```

```
GENDERF      0.0335139  0.0155583   2.154  0.03162 *  
BMI         -0.0187777  0.0018525 -10.136  < 2e-16 ***
CENTERB     -0.0119286  0.0228733  -0.522  0.60220    
CENTERC     -0.0270426  0.0191141  -1.415  0.15763    
CENTERM      0.0551534  0.0187859   2.936  0.00345 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.2514612)

Number of Fisher Scoring iterations: 2
```

2.2.1.4. Stata

In Stata, the *regress* command is used to fit the linear regression model for the rate of change in BMI between the first two visits (RBMI_V2V1), 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. The same commands and options are used as in section 2.1.1.4.

```
fvset base last gendernum diabetes2_indicator bkgrd1_c7 centernum
svyset psu_id [pw=weight_norm_overall_v2], strata(strat)
svy: regress rbmi_v2v1 age i.gendernum bmi i.centernum
```

```
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(   6,     623)  =      68.81
                                              Prob > F          =     0.0000
                                              R-squared         =     0.0873

------------------------------------------------------------------------------
             |             Linearized
   rbmi_v2v1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0062511   .0005648   -11.07   0.000    -.0073602   -.0051419
 0.gendernum |   .0335139   .0155595     2.15   0.032     .0029589    .0640689
         bmi |  -.0187777   .0018523   -10.14   0.000    -.0224152   -.0151402
             |
   centernum |
          1  |  -.0119286   .0228708    -0.52   0.602     -.056841    .0329839
          2  |  -.0270426   .0191058    -1.42   0.157    -.0645616    .0104763
          3  |   .0551534   .0187835     2.94   0.003     .0182674    .0920395
             |
       _cons |   .8688034   .0482231    18.02   0.000     .7741053    .9635015
------------------------------------------------------------------------------
```

2.2.1.5. Mplus

The same syntax from section 2.1.1.5. is used. More decimal places can be viewed by saving the output as a text file (named as “REGCOEFF.dat” in the example code) through the *savedata* statement and invoking the *format* statement.

```
! survey linear
DATA:
FILE IS sol_mplus.dat;
VARIABLE:
! variables in the same order of as created in the dataset;
NAMES = STRAT PSU_ID AGE BMI weight_v2 yrs_v1v2 ckd_v1v2
BMI_V2V1 RBMI_V2V1 KEEP_DATA keep_ckd gender_0 gender_1
center_1 center_2 center_3 center_4
bkc7_0 bkc7_1 bkc7_2 bkc7_3 bkc7_4 bkc7_5 bkc7_6 d2_ind_0 d2_ind_1 ly_v1v2;
! specify what variables we need to use in the analysis;
USEVARIABLES = STRAT PSU_ID AGE BMI weight_v2 yrs_v1v2 RBMI_V2V1
gender_0 center_1 center_2 center_3;
! specify design features;
CLUSTER = PSU_ID;
STRAT = STRAT;
WEIGHT = weight_v2;
MISSING = ALL (999);
ANALYSIS:
! survey method used;
TYPE = COMPLEX;
ESTIMATOR=MLR;
!specify the model;
MODEL:
RBMI_V2V1 on AGE gender_0 BMI yrs_v1v2 center_1 center_2 center_3;
SAVEDATA:
FORMAT IS f10.5;
RESULTS ARE Yourpath\REGCOEFF.dat;
```

```
SUMMARY OF ANALYSIS

Number of groups                                                  1
Number of observations                                        11212

Number of dependent variables                                     1
Number of independent variables                                   7
Number of continuous latent variables                             0

MODEL RESULTS

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

 RBMI_V2V ON
    AGE               -0.006      0.001    -10.765      0.000
    GENDER_0           0.034      0.016      2.164      0.030
    BMI               -0.019      0.002    -10.116      0.000
    YRS_V1V2           0.004      0.009      0.391      0.696
    CENTER_1          -0.011      0.023     -0.499      0.618
    CENTER_2          -0.027      0.019     -1.384      0.166
    CENTER_3           0.056      0.019      2.941      0.003

 Intercepts
    RBMI_V2V1          0.844      0.082     10.247      0.000

 Residual Variances
    RBMI_V2V1          0.253      0.010     24.898      0.000
```

2.2.2. Model-based Procedures

In this section we fit a linear regression model for rate of change in BMI between the first two visits (RBMI_V2V1) using the model-based procedure of weighted GEE with robust variance estimation that accounts for clustering within PSUs, instead of using complex survey procedures as done in section 2.2.1. See section 1.4 for a brief description of these procedures and their differences. We fit the model using SAS, R and Stata. Note that the point estimates are identical, and robust standard error estimates are the same up to the 2nd significant figure among those from model-based procedures in this section 2.2.2., and those from complex survey procedures in section 2.2.1.

2.2.2.1.SAS

The code below invokes the procedure GENMOD in SAS to produce parameter estimates for the desired model. Statements and options specified are the same to the ones presented in section 2.1.2.1. for the model that fits the difference in BMI between visits 1 and 2. The only difference is the outcome and excluding years between visits from the covariates. Note the cluster level, PSU_ID, and working correlation structure (independent, denoted with ind) are specified at SUBJECT and CORR options of the REPEATED statement to obtain robust variance estimation that accounts for clustering within PSUs.

```
proc genmod data=sol;
class PSU_ID GENDER CENTERNUM;
weight WEIGHT_NORM_OVERALL_V2;
model RBMI_V2V1 = AGE GENDER BMI CENTERNUM / dist=normal;
repeated subject=PSU_ID / corr=ind;
run;
```

```
                   Analysis Of GEE Parameter Estimates
                   Empirical Standard Error Estimates
Parameter          Estimate  Standard  95% Confidence Limits        Z  Pr > |Z|
                                Error
Intercept            0.8688    0.0483     0.7742      0.9634   18.00   <.0001
AGE                 -0.0063    0.0006    -0.0074     -0.0051  -11.02   <.0001
GENDER        F      0.0335    0.0155     0.0031      0.0639    2.16   0.0309
```

Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
GENDER	M	0.0000	0.0000	0.0000	0.0000	.	.
BMI		-0.0188	0.0019	-0.0224	-0.0152	-10.15	<.0001
CENTERNUM	1	-0.0119	0.0228	-0.0567	0.0328	-0.52	0.6014
CENTERNUM	2	-0.0270	0.0196	-0.0655	0.0114	-1.38	0.1680
CENTERNUM	3	0.0552	0.0188	0.0182	0.0921	2.93	0.0034
CENTERNUM	4	0.0000	0.0000	0.0000	0.0000	.	.

The results indicate that, after adjusting for field center, sex, and baseline BMI, a one-year increment in age at baseline is associated with a decrease of 0.0063 kg/m^2 in the annual rate of change in BMI.

2.2.2.2.R

As in section 2.1.2.2, the *geeglm* function is used to model the rate of change for BMI from visit 1 to visit 2.

```
model = geeglm(RBMI_V2V1 ~ AGE + GENDER + BMI + CENTER,
               weights = WEIGHT_NORM_OVERALL_V2,
               data=sol,
               id=PSU_ID,
               family=gaussian(link = "identity"))
summary(model)
```

```
Coefficients:
                Estimate   Std.err   Wald Pr(>|W|)
(Intercept)     0.868803  0.047215 338.60   <2e-16 ***
AGE            -0.006251  0.000541 133.66   <2e-16 ***
GENDER.NUMTRUE  0.033514  0.014844   5.10    0.024 *
BMI            -0.018778  0.001872 100.59   <2e-16 ***
CENTERNUM1     -0.011929  0.021445   0.31    0.578
CENTERNUM2     -0.027043  0.018152   2.22    0.136
CENTERNUM3      0.055153  0.018609   8.78    0.003 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

2.2.2.3. Stata

As in section 2.1.2.3, the *meglm* command is used to fit linear regression model for the rate of change for BMI from visit 1 to visit 2.

```
meglm rbmi_v2v1 age ib1.gendernum ib4.centernum bmi [pw=weight_norm_overall_v2], vce(cluster
psu_id)
```

```
Mixed-effects GLM                               Number of obs     =     11,212
Family:                 Gaussian
Link:                   identity

                                                Wald chi2(6)      =     416.88
Log pseudolikelihood = -8148.6645               Prob > chi2       =     0.0000
                               (Std. Err. adjusted for 648 clusters in psu_id)
------------------------------------------------------------------------------
             |               Robust
   rbmi_v2v1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0062511   .0005676   -11.01   0.000    -.0073636   -.0051386
 0.gendernum |   .0335139    .015535     2.16   0.031     .0030658     .063962
             |
   centernum |
          1  |  -.0119286   .0228554    -0.52   0.602    -.0567244    .0328673
          2  |  -.0270426   .0196286    -1.38   0.168    -.0655139    .0114287
          3  |   .0551534   .0188559     2.92   0.003     .0181966    .0921103
             |
         bmi |  -.0187777   .0018521   -10.14   0.000    -.0224077   -.0151476
       _cons |   .8688034   .0483121    17.98   0.000     .7741134    .9634935
-------------+----------------------------------------------------------------
var(e.rbmi~1)|   .2535047   .0104777                      .2337785    .2748953
------------------------------------------------------------------------------
```

3.Logistic Regression for Visit 2 Binary Outcome

The purpose of this chapter is to estimate odds ratio of the incidence event using logistic regression in SAS, SUDAAN, R, Stata, and Mplus. We present both design-based complex survey procedures and model-based non-survey procedures, when available in each software. We use incidence of chronic kidney disease (CKD) at visit 2 as an example. The incidence of CKD is denoted by the binary variable, INCIDENT_CKD_V1V2, which is an indicator function for low eGFR at visit 2 with an annual eGFR decreasing rate of 1+ mL/min/1.73m^2, and/or high serum albumin-creatinine ratio at visit 2, among those without chronic kidney disease at baseline. To study CKD incidence, the population of interest is restricted to participants free of CKD at baseline. The flag variable KEEP_DATA_CKD is defined to select those without CKD at visit 1. Because the elapsed time between visit 1 and visit 2 (YRS_BTWN_V1V2) varies among participants, we will adjust for it. 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 Chapter 4 for details).

3.1. Complex Survey Procedures

In this section we fit a logistic regression model to estimate odds ratio of the CKD incidence event (INCIDENT_CKD_V1V2) using complex survey procedures in SAS, SUDAAN, R, Stata and Mplus. Note that the point estimates and robust standard error estimates are essentially identical among those from complex survey procedures in this section 3.1.

3.1.1. 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 of those without CKD at visit 1 is specified with KEEP_DATA_CKD 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. 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 CENTERNUM/ PARAM=REF;
   model INCIDENT_CKD_V1V2(EVENT=’1’) = AGE GENDER YRS_BTWN_V1V2 CENTERNUM/
link=logit;
run;
```

Type 3 Analysis of Effects				
Effect	**F Value**	**Num DF**	**Den DF**	**Pr > F**
AGE	60.38	1	630	<.0001
GENDER	0.12	1	630	0.7342
YRS_BTWN_V1V2	0.00	1	630	0.9689
CENTERNUM	2.16	3	628	0.0921

Analysis of Maximum Likelihood Estimates					
Parameter		**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept		-4.6852	0.6152	-7.62	<.0001
AGE		0.0375	0.00483	7.77	<.0001
GENDER	**F**	0.0440	0.1295	0.34	0.7342
YRS_BTWN_V1V2		0.00344	0.0882	0.04	0.9689
CENTERNUM	**1**	0.3579	0.1734	2.06	0.0395
CENTERNUM	**2**	0.1307	0.1743	0.75	0.4536
CENTERNUM	**3**	-0.0265	0.1825	-0.14	0.8848
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	1.048
GENDER F vs M	1.045	0.810	1.347
YRS_BTWN_V1V2	1.003	0.844	1.193
CENTERNUM 1 vs 4	1.430	1.017	2.011
CENTERNUM 2 vs 4	1.140	0.809	1.605
CENTERNUM 3 vs 4	0.974	0.680	1.394
NOTE: The degrees of freedom in computing the confidence limits is 630.			

The estimated odds ratio for incident CKD is exp(0.044) = 1.045 for females relative to males after adjusting for baseline age, center, and elapsed time between the two visits.

The associated 95% confidence interval is (0.810, 1.347). The interpretation is that, for females, the odds of developing CKD are 1.045 times as large as the odds for males developing CKD, given they are of the same age, Hispanic/Latino background, field center, and have same follow-up time.

3.1.2. 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 binary variable INCIDENT_CKD_V1V2. 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. The subpopulation of those without CKD at visit 1 is specified with KEEP_DATA_CKD=1 in the SUBPOPN statement. Note that, by default, SUDAAN outputs results using only two decimal places; in order to increase this number, one can use the statement SETENV and set the number of decimal places to be used through the option DECWIDTH.

```
Proc multilog data=worklib.sol filetype=sas design=wr notsorted;
   nest STRAT PSU_ID;
   weight WEIGHT_NORM_OVERALL_V2;
   class INCIDENT_CKD_V1V2 GENDERNUM CENTERNUM;
   subpopn KEEP_DATA_CKD=1;
   model INCIDENT_CKD_V1V2 = AGE GENDERNUM YRS_BTWN_V1V2 CENTERNUM;
   reflevel GENDERNUM=1 CENTERNUM=4; /* reference: Male San Diego*/
   setenv decwidth=4;
run;
```

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


Contrast               Degrees
                       of                      P-value
                       Freedom        Wald F   Wald F
-------------------------------------------------------
OVERALL MODEL            7.0000     303.0918     0.0000
MODEL MINUS
  INTERCEPT              6.0000      13.0931     0.0000
INTERCEPT                  .            .          .
AGE                      1.0000      60.4102     0.0000
GENDERNUM                1.0000       0.1155     0.7341
YRS_BTWN_V1V2            1.0000       0.0015     0.9690
CENTERNUM                3.0000       2.1640     0.0911
-------------------------------------------------------
--------------------------------------------------------------------------------------------------------
|                 |                   | Independent Variables and Effects                                |
| INCIDENT_CKD_V- |                   |----------------------------------------------------------------|
| 1V2 (log-odds)  |                   | Intercept  | Age         | Gender     | Gender     | Elapsed    |
|                 |                   |            |             | (0=Female, | (0=Female, | time       |
|                 |                  |            |            | 1=Male) =  | 1=Male) =  | between    |
|                 |                  |            |            | 0          | 1          | visits 1   |
--------------------------------------------------------------------------------------------------------
|                 |                  |            |            |            |            |            |
| 0 vs 1          | Beta Coeff.      |     4.6852 |    -0.0375 |    -0.0440 |     0.0000 |    -0.0034 |
|                 | SE Beta          |     0.6151 |     0.0048 |     0.1294 |     0.0000 |     0.0882 |
|                 | Lower 95% Limit  |            |            |            |            |            |
|                 |  Beta            |     3.4774 |    -0.0470 |    -0.2982 |     0.0000 |    -0.1766 |
|                 | Upper 95% Limit  |            |            |            |            |            |
|                 |  Beta            |     5.8930 |    -0.0280 |     0.2102 |     0.0000 |     0.1697 |
|                 | T-Test B=0       |     7.6174 |    -7.7724 |    -0.3398 |      .     |    -0.0389 |
|                 | P-value T-Test   |            |            |            |            |            |
|                 |  B=0             |     0.0000 |     0.0000 |     0.7341 |      .     |     0.9690 |
--------------------------------------------------------------------------------------------------------


-------------------------------------------------------------------------------------------
|                 |                  | Independent Variables and Effects                 |
| INCIDENT_CKD_V- |                  |---------------------------------------------------|
| 1V2 (log-odds)  |                  | Participa- | Participa- | Participa- | Participa- |
|                 |                  | nt's Field | nt's Field | nt's Field | nt's Field |
|                 |                  | Center -   | Center -   | Center -   | Center -   |
|                 |                  | numeric =  | numeric =  | numeric =  | numeric =  |
|                 |                  | 1          | 2          | 3          | 4          |
-------------------------------------------------------------------------------------------
|                 |                  |            |            |            |            |
| 0 vs 1          | Beta Coeff.      |    -0.3579 |    -0.1308 |     0.0265 |     0.0000 |
|                 | SE Beta          |     0.1734 |     0.1743 |     0.1825 |     0.0000 |
|                 | Lower 95% Limit  |            |            |            |            |
|                 |  Beta            |    -0.6984 |    -0.4730 |    -0.3319 |     0.0000 |
|                 | Upper 95% Limit  |            |            |            |            |
|                 |  Beta            |    -0.0175 |     0.2115 |     0.3848 |     0.0000 |
|                 | T-Test B=0       |    -2.0644 |    -0.7502 |     0.1450 |      .     |
|                 | P-value T-Test   |            |            |            |            |
|                 |  B=0             |     0.0394 |     0.4534 |     0.8848 |      .     |
-------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------
|                 |                  | Independent Variables and Effects                              |
| INCIDENT_CKD_V- |                  |----------------------------------------------------------------|
| 1V2 (log-odds)  |                  | Intercept  | Age        | Gender     | Gender     | Elapsed    |
|                 |                  |            |            | (0=Female, | (0=Female, | time       |
|                 |                  |            |            | 1=Male) =  | 1=Male) =  | between    |
|                 |                  |            |            | 0          | 1          | visits 1   |
|                 |                  |            |            |            |            | and 2      |
|                 |                  |            |            |            |            | (yrs)      |
--------------------------------------------------------------------------------------------------------
|                 |                  |            |            |            |            |            |
| 0 vs 1          | Odds Ratio       |   108.3274 |     0.9632 |     0.9570 |     1.0000 |     0.9966 |
|                 | Lower 95% Limit  |            |            |            |            |            |
|                 |  OR              |    32.3741 |     0.9541 |     0.7422 |     1.0000 |     0.8381 |
|                 | Upper 95% Limit  |            |            |            |            |            |
|                 |  OR              |   362.4764 |     0.9723 |     1.2339 |     1.0000 |     1.1850 |
--------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
|                 |                  | Independent Variables and Effects      |            |
| INCIDENT_CKD_V- |                  |----------------------------------------|-------------
| 1V2 (log-odds)  |                  | Participa- | Participa- | Participa- | Participa- |
|                 |                  | nt's Field | nt's Field | nt's Field | nt's Field |
|                 |                  | Center -   | Center -   | Center -   | Center -   |
|                 |                  | numeric =  | numeric =  | numeric =  | numeric =  |
|                 |                  | 1          | 2          | 3          | 4          |
------------------------------------------------------------------------------------------|
|                 |                  |            |            |            |            |
| 0 vs 1          | Odds Ratio       |     0.6991 |     0.8774 |     1.0268 |     1.0000 |
|                 | Lower 95% Limit |            |            |            |            |
|                 |  OR             |     0.4974 |     0.6231 |     0.7175 |     1.0000 |
|                 | Upper 95% Limit |            |            |            |            |
|                 |  OR             |     0.9827 |     1.2355 |     1.4694 |     1.0000 |
-------------------------------------------------------------------------------------------
```

The estimated odds ratio for incident CKD is 1/0.957 = 1.045 for females relative to males after adjusting for baseline age, center, and time between the two visits. The associated 95% confidence interval is (1/1.234 = 0.810, 1/0.742 = 1.347).

3.1.3. 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 rate of change models in section 2.1.1.3. and 2.2.1.3. 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. The subpopulation of those without CKD at visit 1 is specified with KEEP_DATA_CKD==1 with the ‘subset’ argument.

```
> model.bin = svyglm(INCIDENT_CKD_V1V2 ~ AGE + GENDER + YRS_BTWN_V1V2 + CENTER, design =
sol.design,subset=KEEP_DATA_CKD==1,family=quasibinomial(link='logit'))

> summary(model.bin)

Call:

svyglm(formula = INCIDENT_CKD_V1V2 ~ AGE + GENDER + YRS_BTWN_V1V2 + CENTER , design =
sol.design, subset = KEEP_DATA_CKD == 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)   -4.685158   0.615075  -7.617 9.65e-14 ***
AGE            0.037532   0.004829   7.772 3.18e-14 ***
GENDERF        0.043981   0.129435   0.340   0.7341
YRS_BTWN_V1V2  0.003429   0.088184   0.039   0.9690
CENTERB        0.357913   0.173384   2.064   0.0394 *
CENTERC        0.130757   0.174298   0.750   0.4534
CENTERM       -0.026458   0.182503  -0.145   0.8848
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.011708)

Number of Fisher Scoring iterations: 5
```

```
exp(cbind(Odds=coef(model.bin), confint(model.bin)))
                      Odds       2.5 %      97.5 %
(Intercept)   0.009231273 0.002758614 0.03089101
AGE           1.038244900 1.028446132 1.04813703
GENDERF       1.044962520 0.810421728 1.34738079
YRS_BTWN_V1V2 1.003435092 0.843880609 1.19315691
CENTERB       1.430340833 1.017578873 2.01053201
CENTERC       1.139690970 0.809349464 1.60486361
CENTERM       0.973888980 0.680550846 1.39366478
```

The estimated odds ratio for incident CKD is exp(0.044) = 1.045 for females relative to males after adjusting for baseline age, center, and time between the two visits. The associated 95% confidence interval is (0.810, 1.347).

3.1.4. Stata

Logistic regression can be fit using the *logit* command with the usual syntax. 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. Domain variable KEEP_DATA_CKD==1 indicating those without CKD at visit 1 is specified in the *subpop* option before the *logit* 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).

```
fvset base last gendernum diabetes2_indicator bkgrd1_c7 centernum
svyset psu_id [pw=weight_norm_overall_v2], strata(strat)
svy,  subpop(if keep_data_ckd==1):logit incident_ckd_v1v2 age i.gendernum yrs_btwn_v1v2 i.centernum
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(   6,     626)  =      12.99
                                              Prob > F          =     0.0000

------------------------------------------------------------------------------
             |             Linearized
incident_c~2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0375317   .0048287     7.77   0.000     .0280494     .047014
 0.gendernum |    .043981   .1294356     0.34   0.734    -.2101957    .2981578
yrs_btwn_v~2 |   .0034292    .088182     0.04   0.969    -.1697365     .176595
             |
   centernum |
          1  |   .3579128   .1733846     2.06   0.039      .017432    .6983935
```

```
           2  |    .1307571   .1742996     0.75   0.453    -.2115204    .4730346
           3  |    -.026458   .1825056    -0.14   0.885    -.3848498    .3319339
              |
        _cons |   -4.685158   .6150812    -7.62   0.000    -5.893012   -3.477305
------------------------------------------------------------------------------
```

```
.
.
.
. 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(   6,    626)   =      12.99
                                              Prob > F          =     0.0000

------------------------------------------------------------------------------
             |             Linearized
incident_c~2 | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.038245   .0050134     7.77   0.000     1.028446    1.048137
 0.gendernum |   1.044963   .1352554     0.34   0.734     .8104256    1.347374
yrs_btwn_v~2 |   1.003435   .0884849     0.04   0.969     .8438871    1.193148
             |
   centernum |
          1  |   1.430341   .2479991     2.06   0.039     1.017585     2.01052
          2  |   1.139691   .1986477     0.75   0.453     .8093528    1.604857
          3  |    .973889   .1777402    -0.14   0.885     .6805529    1.393661
             |
       _cons |   .0092313    .005678    -7.62   0.000     .0027587    .0308906
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

The estimated odds ratio for incident CKD is exp(0.044) = 1.045 for females relative to males after adjusting for baseline age, center, and elapsed time between the two visits. The associated 95% confidence interval is (0.810, 1.347).

3.1.5. Mplus

The *ANALYSIS: TYPE = COMPLEX* statement in Mplus and the specification of the outcome in the *CATEGORICAL* statement is invoked to fit logistic regression model using complex survey procedures. Design variables are specified through the statements *STRAT*, *CLUSTER*, and *WEIGHT*.

INCIDENT_CKD_V1V2 (renamed as ckd_v1v2 to shorten it) is modelled through the *MODEL* statement as the binary outcome which is specified through the *CATEGORICAL* statement. Since we are interested in making inference for those without CKD at visit 1, domain variable KEEP_DATA_CKD (renamed as keep_ckd) is specified in the *SUBPOPULATION* statement, with ‘EQ 1’ indicating subpopulation of interest.

More decimal places can only be viewed by saving the output as a text file (named as “REGCOEFF.dat” in the example code) through the *savedata* statement and invoking the *format* statement.

```
! survey logistic
DATA:
FILE IS sol_mplus.dat;
VARIABLE:
! variables in the same order of as created in the dataset;
NAMES = STRAT PSU_ID AGE BMI weight_v2 yrs_v1v2 ckd_v1v2
BMI_V2V1 RBMI_V2V1 KEEP_DATA keep_ckd gender_0 gender_1
center_1 center_2 center_3 center_4
bkc7_0 bkc7_1 bkc7_2 bkc7_3 bkc7_4 bkc7_5 bkc7_6 d2_ind_0 d2_ind_1 ly_v1v2;
! specify what variables we need to use in the analysis;
USEVARIABLES = STRAT PSU_ID AGE weight_v2 yrs_v1v2
ckd_v1v2 gender_0 center_1 center_2 center_3;
! specify design features;
SUBPOPULATION = keep_ckd EQ 1;
CLUSTER = PSU_ID;
STRAT = STRAT;
WEIGHT = weight_v2;
CATEGORICAL = ckd_v1v2;
MISSING = ALL (999);
ANALYSIS:
! survey method used;
TYPE = COMPLEX;
ESTIMATOR=MLR;
!specify the model;
MODEL:
ckd_v1v2 on AGE gender_0 yrs_v1v2 center_1 center_2 center_3;
SAVEDATA:
FORMAT IS f10.5;
RESULTS ARE Yourpath\REGCOEFF.dat;
```

```
SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                       10090

Number of dependent variables                                    1
Number of independent variables                                  6
Number of continuous latent variables                            0

MODEL RESULTS


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

 CKD_V1V2   ON
    AGE                0.038      0.005      7.773      0.000
    GENDER_0           0.044      0.129      0.340      0.734
    YRS_V1V2           0.003      0.088      0.039      0.969
    CENTER_1           0.358      0.173      2.064      0.039
    CENTER_2           0.131      0.174      0.750      0.453
    CENTER_3          -0.026      0.182     -0.145      0.885

 Thresholds
    CKD_V1V2$1         4.685      0.615      7.617      0.000
```

```
LOGISTIC REGRESSION ODDS RATIO RESULTS

                                              (Est. - 1) Two-Tailed
                    Estimate       S.E.     / S.E.    P-Value

 CKD_V1V2   ON
    AGE                1.038      0.005      7.629      0.000
    GENDER_0           1.045      0.135      0.332      0.740
    YRS_V1V2           1.003      0.088      0.039      0.969
    CENTER_1           1.430      0.248      1.735      0.083
    CENTER_2           1.140      0.199      0.703      0.482
    CENTER_3           0.974      0.178     -0.147      0.883
```

The estimated odds ratio for incident CKD is exp(0.044) = 1.045 for females relative to males after adjusting for baseline age, center, and elapsed time between the two visits. The associated 95% confidence interval is (exp(0.044-invnormal(0.975)*0.129)=0.810, exp(0.044+invnormal(0.975)*0.129)=1.347). Note that invnormal denotes inverse cumulative standard normal distribution, and invnormal(0.975) = 1.96.

3.2. Model-based Procedures

In this section we fit a logistic regression model to estimate odds ratio of the CKD incidence event (INCIDENT_CKD_V1V2) using the model-based procedure of weighted GEE with robust variance estimation that accounts for clustering within PSUs, instead of using complex survey procedures as done in previous section 3.1. See section 1.4 for a brief description of these procedures and their differences. We fit the model using SAS, R and Stata. Note that the point estimates are identical, and robust standard error estimates are the same up to the 2nd significant figure among those from model-based procedures in this section 3.2., and those from complex survey procedures in section 3.1.

3.2.1. SAS

In SAS, the model-based procedure PROC GENMOD is used. The CLASS statement is used to specify categorical variables; the WEIGHT statement is used to specify the subject-level sampling weight; the MODEL statement is used to specify the analysis model; the DIST option is used to specify the marginal distribution of the outcome; and the cluster level, PSU_ID, and working correlation structure (independent, denoted with ind) are specified at SUBJECT and CORR options of the REPEATED statement to obtain robust variance estimators that account for clustering within PSUs.. Odds ratios are obtained by using the ESTIMATE statement with the EXP option. When the logistic regression is used, note that: (1) the marginal distribution of the outcome is specified as binomial distribution and the logit link function is used; (2) the analysis is restricted to the subset of the data in which the subjects does not have CKD at visit 1, which can be specified at the WHERE statement; and (3) the DESCENDING option is necessary if the interest lies in modeling probability of INCIDENT_CKD_V1V2 = 1. Otherwise, probability of INCIDENT_CKD_V1V2 = 0 would be modeled.

```
proc genmod data=sol descending ;
where KEEP_DATA_CKD=1;
class PSU_ID GENDER CENTERNUM;
weight WEIGHT_NORM_OVERALL_V2;
model INCIDENT_CKD_V1V2 = AGE GENDER YRS_BTWN_V1V2 CENTERNUM / dist=binomial
link=logit;
repeated subject=PSU_ID / corr=ind ;
      estimate "Beta" GENDER 1 -1 / exp;
run;
```

Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-4.6852	0.6182	-5.8969	-3.4735	-7.58	<.0001
AGE		0.0375	0.0049	0.0280	0.0470	7.73	<.0001
GENDER	F	0.0440	0.1287	-0.2082	0.2962	0.34	0.7325
GENDER	M	0.0000	0.0000	0.0000	0.0000	.	.
YRS_BTWN_V1V2		0.0034	0.0882	-0.1695	0.1764	0.04	0.9690
CENTERNUM	1	0.3579	0.1713	0.0222	0.6936	2.09	0.0367
CENTERNUM	2	0.1308	0.1718	-0.2060	0.4675	0.76	0.4466
CENTERNUM	3	-0.0265	0.1820	-0.3832	0.3303	-0.15	0.8844
CENTERNUM	4	0.0000	0.0000	0.0000	0.0000	.	.

Contrast Estimate Results

Label	Mean Estimate	Mean Confidence Limits		L'Beta Estimate	Standard Error	Alpha	L'Beta Confidence Limits		Chi-Square	Pr > ChiSq
Beta	0.5110	0.4481	0.5735	0.0440	0.1287	0.05	-0.2082	0.2962	0.12	0.7325
Exp(Beta)				1.0450	0.1345	0.05	0.8120	1.3447		

The estimated odds ratio for incident CKD is 1.045 for females relative to males after adjusting for baseline age, center, and elapsed time between the two visits. The associated 95% confidence interval is (0.812, 1.347).

3.2.2. R

The *geeglm* function from R package "*geepack*" is used. The "weights = WEIGHT_NORM_OVERALL_V2" statement indicates the subject-level sampling weight; the "data=sol" and "id=PSU_ID" statements indicate the working dataset and cluster level respectively. The default of this function is to estimate robust variance, and "id=PSU_ID" identifies the cluster level (PSU) for the robust variance. When the logistic regression is used, note that: (1) the marginal distribution of the outcome is specified as binomial distribution and the logit link function is used; (2) the analysis is restricted to the subset of the participants free of CKD at baseline, which can be specified by "subset=(KEEP_DATA_CKD==1)".

```
model = geeglm(INCIDENT_CKD_V1V2 ~ AGE + GENDER + YRS_BTWN_V1V2 + CENTER,
               weights=WEIGHT_NORM_OVERALL_V2,
               data=sol,
               id=PSU_ID,
               family=binomial(link="logit"),
               subset=(KEEP_DATA_CKD==1))
summary(model)
```

```
Coefficients:
              Estimate  Std.err  Wald Pr(>|W|)
(Intercept)   -4.68516  0.59300 62.42  2.8e-15 ***
AGE            0.03753  0.00512 53.78  2.2e-13 ***
GENDERF        0.04398  0.13147  0.11    0.738
YRS_BTWN_V1V2  0.00343  0.08368  0.00    0.967
CENTERB        0.35791  0.17779  4.05    0.044 *
CENTERC        0.13076  0.17573  0.55    0.457
CENTERM       -0.02646  0.17627  0.02    0.881
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The estimated odds ratio for incident CKD is exp(0.0440) = 1.045 for females relative to males after adjusting for baseline age, center, and elapsed time between the two visits. The associated 95% confidence interval is (exp(0.044-invnormal(0.975)*0.131)=0.808, exp(0.044+invnormal(0.975)*0.131)=1.352).

3.2.3. Stata

In Stata, the *meglm* command is used. The *pw* option is used to specify the subject-level sampling weight; the *vce* option is used to indicate the use of robust variance estimator; and *cluster psu_id* is used to specify the cluster level (PSU) for the robust variance. When the logistic regression is used, note that: (1) the marginal distribution of the outcome is specified as binomial distribution by *family(bernoulli)* and the logit link function is specified by *link (logit)*; (2) the analysis is restricted to the subset of the participants free of CKD at baseline, and those ineligible subjects can be excluded from the analysis by using the *drop* command before the analysis is conducted. Odds ratios can be requested by using the option *or* (either with the original *meglm* command call, or by using the statement *meglm, or* after the logistic regression was fit).

```
drop if keep_data_ckd == 0
meglm incident_ckd_v1v2 age ib1.gendernum ib4.centernum yrs_btwn_v1v2
[pw=weight_norm_overall_v2], family(bernoulli) link(logit) vce(cluster psu_id)
meglm, or
```

```
Mixed-effects GLM                               Number of obs     =     10,090
Family:               Bernoulli
Link:                     logit

                                                Wald chi2(6)      =      77.87
Log pseudolikelihood = -2053.0528               Prob > chi2       =     0.0000
                               (Std. Err. adjusted for 650 clusters in psu_id)
------------------------------------------------------------------------------
             |               Robust
incident_c~2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0375317   .0048567     7.73   0.000     .0280128    .0470506
 0.gendernum |    .043981   .1287875     0.34   0.733    -.2084379    .2963999
             |
   centernum |
          1  |   .3579128   .1714227     2.09   0.037     .0219305    .6938951
          2  |   .1307571     .17193     0.76   0.447    -.2062195    .4677338
          3  |   -.026458   .1821393    -0.15   0.885    -.3834444    .3305285
             |
yrs_btwn_v~2 |   .0034292   .0882969     0.04   0.969    -.1696296     .176488
       _cons |  -4.685158   .6186993    -7.57   0.000    -5.897787


. meglm, or

Mixed-effects GLM                               Number of obs     =     10,090
Family:               Bernoulli
Link:                     logit

                                                Wald chi2(6)      =      77.87
Log pseudolikelihood = -2053.0528               Prob > chi2       =     0.0000
                               (Std. Err. adjusted for 650 clusters in psu_id)
------------------------------------------------------------------------------
             |               Robust
incident_c~2 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
         age |   1.038245   .0050424     7.73   0.000     1.028409    1.048175
 0.gendernum |   1.044963   .1345781     0.34   0.733     .8118515    1.345008
             |
   centernum |
          1  |   1.430341   .2451929     2.09   0.037     1.022173    2.001496
          2  |   1.139691   .1959471     0.76   0.447     .8136544    1.596372
          3  |    .973889   .1773834    -0.15   0.885       .68151    1.391703
             |
yrs_btwn_v~2 |   1.003435   .0886002     0.04   0.969     .8439774     1.19302
       _cons |   .0092313   .0057114    -7.57   0.000     .0027455    .0310384
------------------------------------------------------------------------------
Note: _cons estimates baseline odds (conditional on zero random effects).
```

The estimated odds ratio for incident CKD is 1.045 for females relative to males after adjusting for baseline age, center, and elapsed time between the two visits. The associated 95% confidence interval is (0.812,1.345).

4. Poisson Regression for Visit 2 Binary Outcome

The purpose of this chapter is to estimate incidence rate ratio and calculate adjusted incidence rate using Poisson regression in SAS, SUDAAN, R, Stata, and Mplus. Incidence rate ratios can be significantly different from odds ratios when the event of interest is not rare (incidence rate > 10%). Instead of logistic regression (Chapter 3), Poisson regression is recommended to provide estimation of covariate effect on the incidence rate. We present both design-based complex survey procedures and model-based non-survey procedures, given their availability in each software. We use incidence of chronic kidney disease (CKD) at visit 2 as an example. The incidence of CKD is denoted by the binary variable, INCIDENT_CKD_V1V2, which is an indicator function for low eGFR at visit 2 with an annual eGFR decreasing rate of 1+ mL/min/1.73m^2, and/or high serum albumin-creatinine ratio at visit 2, among those without chronic kidney disease at baseline. To study CKD incidence, the population of interest is restricted to participants free of CKD at baseline. The flag variable KEEP_DATA_CKD is defined to select those without CKD at visit 1. Because the elapsed time between visit 1 and visit 2 varies among participants, we will use time elapsed between visit 1 and visit 2 (YRS_BTWN_V1V2) as an offset.

4.1. Complex Survey Procedures

In this section we fit a Poisson regression model to estimate incidence rate ratio and calculate adjusted incidence rates of CKD (INCIDENT_CKD_V1V2) using complex survey procedures in SUDAAN, R, Stata and Mplus. Note that the point estimates and robust standard error estimates are essentially identical among those from complex survey procedures in this section 4.1.

4.1.1. 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, INCIDENT_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. The subpopulation of those without CKD at visit 1 is specified with KEEP_DATA_CKD=1 in the SUBPOPN statement. Note that an OFFSET option needs to be specified for YRS_BTWN_V1V2, the time elapsed between visit 1 and visit 2.

```
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 CENTERNUM;
   subpopn KEEP_DATA_CKD=1;
   model INCIDENT_CKD_V1V2 = AGE BKGRD1_C7 GENDERNUM DIABETES2_INDICATOR
CENTERNUM / OFFSET=YRS_BTWN_V1V2;
   reflevel GENDERNUM=1 BKGRD1_C7=3 DIABETES2_INDICATOR=1 CENTERNUM=4; /*
reference: Male Mexican San Diego*/
   setenv decwidth=4;
run;
```

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent

Link Function: Log
Response variable INCIDENT_CKD_V1V2: Incident Chronic Kidney Disease from V1 to
  V2 (using eGFR,ACR and >=1 avg decline in GGR per year)
Offset variable YRS_BTWN_V1V2: Elapsed time between visits 1 and 2 (yrs)
For Subpopulation: KEEP_DATA_CKD = 1
by: Contrast.

------------------------------------------------------------------------------------------
Independent                                                                        P-value
  Variables and        Beta                   Lower 95%    Upper 95%               T-Test
  Effects              Coeff.       SE Beta   Limit Beta   Limit Beta  T-Test B=0   B=0
------------------------------------------------------------------------------------------
Intercept                -5.0614     0.2809     -5.6129     -4.5099    -18.0213    0.0000
7-level re-
  classification of
  Hispanic/Latino
  Background
  0                      -0.7284     0.3386     -1.3933     -0.0634     -2.1511    0.0318
  1                      -0.4199     0.3065     -1.0218      0.1820     -1.3700    0.1712
  2                      -0.4875     0.3155     -1.1071      0.1321     -1.5449    0.1229
  3                       0.0000     0.0000      0.0000      0.0000           .         .
  4                      -0.2981     0.2595     -0.8078      0.2116     -1.1486    0.2511
  5                      -0.7573     0.3170     -1.3798     -0.1348     -2.3890    0.0172
  6                      -1.4669     0.4871     -2.4235     -0.5104     -3.0114    0.0027
Gender (0=Female,
  1=Male)
  0                       0.0122     0.1211     -0.2257      0.2500      0.1004    0.9200
  1                       0.0000     0.0000      0.0000      0.0000           .         .
Diabetes Indicator -
  ADA
  0                      -1.0916     0.1160     -1.3194     -0.8639     -9.4120    0.0000
  1                       0.0000     0.0000      0.0000      0.0000           .         .
Participant's Field
  Center - numeric
  1                       0.7549     0.2951      0.1755      1.3344      2.5584    0.0107
  2                       0.2385     0.1718     -0.0989      0.5758      1.3880    0.1656
  3                       0.5662     0.3151     -0.0525      1.1849      1.7970    0.0728
  4                       0.0000     0.0000      0.0000      0.0000           .         .
Age                       0.0240     0.0045      0.0152      0.0328      5.3467    0.0000
------------------------------------------------------------------------------------------
```

```
-----------------------------------------------------------
Independent            Incidence
  Variables and        Density      Lower 95%    Upper 95%
  Effects              Ratio        Limit IDR    Limit IDR
-----------------------------------------------------------
Intercept                  0.0063       0.0037       0.0110
7-level re-
  classification of
  Hispanic/Latino
  Background
  0                        0.4827       0.2483       0.9385
  1                        0.6571       0.3600       1.1996
  2                        0.6142       0.3305       1.1413
  3                        1.0000       1.0000       1.0000
  4                        0.7422       0.4458       1.2356
  5                        0.4689       0.2516       0.8739
  6                        0.2306       0.0886       0.6003
Gender (0=Female,
  1=Male)
  0                        1.0122       0.7980       1.2840
  1                        1.0000       1.0000       1.0000
Diabetes Indicator -
  ADA
  0                        0.3357       0.2673       0.4215
  1                        1.0000       1.0000       1.0000
Participant's Field
  Center - numeric
  1                        2.1275       1.1918       3.7977
  2                        1.2693       0.9058       1.7786
  3                        1.7615       0.9488       3.2703
  4                        1.0000       1.0000       1.0000
Age                        1.0243       1.0153       1.0334
-----------------------------------------------------------
```

The estimated incidence rate ratio (also called incidence density ratio, as shown in the example output) for CKD comparing those without diabetes at baseline and those with diabetes is 0.336, with the associated confidence interval (0.267, 0.421), after adjusting for age, field center, Hispanic/Latino background, and sex. The interpretation is that an individual without diabetes at baseline is expected to have 0.336 times the rate of developing CKD compared to those with diabetes at baseline, given they are of the same age, Hispanic/Latino background, sex, from the same field center, and have same follow-up time.

For individuals who are at population mean age (47 years), female, Dominican, without diabetes at baseline, and at Bronx field center, the adjusted incidence rate is exp(-5.0614+0.0240*47+0.0122-0.7284-1.0916+0.7549)*1000 = 6.83 per 1000 person-years, meaning that we expected 6.83 cases of CKD incidence in a year among 1000 individuals who are 47-year-old female Dominican with no diabetes at Bronx field center.

4.1.2. R

The Poisson regression model is fitted similarly as the logistic regression model; the only difference is the specification of the ‘quasipoisson’ family and the ‘log’ link. 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. The subpopulation of those without CKD at visit 1 is specified with KEEP_DATA_CKD==1 with the ‘subset’ argument. 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(INCIDENT_CKD_V1V2 ~ AGE +BKGRD1_C7+ GENDER+
DIABETES2_INDICATOR+ CENTER + offset(log(YRS_BTWN_V1V2)), design = sol.design,
subset=KEEP_DATA_CKD==1,family=quasipoisson(link='log'))
> summary(model.pois)

Call:
svyglm(formula = INCIDENT_CKD_V1V2 ~ AGE + BKGRD1_C7 + GENDER +
DIABETES2_INDICATOR + CENTER +
    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)          -5.06142    0.28086 -18.021  < 2e-16 ***
AGE                   0.02401    0.00449   5.347 1.26e-07 ***
BKGRD1_C70           -0.72838    0.33861  -2.151  0.03186 *
BKGRD1_C71           -0.41989    0.30650  -1.370  0.17120
BKGRD1_C72           -0.48747    0.31553  -1.545  0.12288
BKGRD1_C74           -0.29812    0.25954  -1.149  0.25115
BKGRD1_C75           -0.75729    0.31699  -2.389  0.01719 *
BKGRD1_C76           -1.46693    0.48712  -3.011  0.00271 **
GENDERF               0.01217    0.12112   0.100  0.92002
DIABETES2_INDICATOR0 -1.09164    0.11598  -9.412  < 2e-16 ***
CENTERB               0.75493    0.29508   2.558  0.01075 *
CENTERC               0.23845    0.17180   1.388  0.16564
CENTERM               0.56618    0.31506   1.797  0.07282 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 0.9687013)

Number of Fisher Scoring iterations: 6

> exp(cbind(IDR=coef(model.pois), confint(model.pois)))
                             IDR       2.5 %      97.5 %

(Intercept)          0.006336585 0.003650236 0.01099992
AGE                  1.024299430 1.015306542 1.03337197
```

```
BKGRD1_C70          0.482690745 0.248244568 0.93855168
BKGRD1_C71          0.657121151 0.359950994 1.19963055
BKGRD1_C72          0.614177026 0.330510738 1.14130458
BKGRD1_C74          0.742214264 0.445835063 1.23561841
BKGRD1_C75          0.468934517 0.251627302 0.87390986
BKGRD1_C76          0.230632277 0.088606765 0.60030684
GENDERF             1.012239579 0.797971117 1.28404267
DIABETES2_INDICATOR0 0.335664710 0.267292392 0.42152639
CENTERB             2.127461654 1.191786109 3.79773942
CENTERC             1.269284745 0.905808508 1.77861408
CENTERM             1.761526985 0.948809932 3.27038874
#End

> exp(sum(coef(model.pois)*c(1,47,c(1,0,0,0,0,0),1,1,c(1,0,0))))*1000
[1] 6.833462
```

The estimated incidence rate ratio for CKD comparing those without diabetes at baseline and those with diabetes is 0.336, with the associated confidence interval (0.267, 0.421), after adjusting for age, field center, Hispanic/Latino background, and sex. For individuals who are at population mean age (47 years), female, Dominican, without diabetes at baseline, and at Bronx field center, the adjusted incidence rate is 6.83 per 1000 person-years.

4.1.3. Stata

Poisson regression can be fit using the *poisson* command with the usual syntax. 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. Domain variable KEEP_DATA_CKD==1 indicating those without CKD at visit 1 is specified in the *subpop* option before the *poisson* command. The *offset* option should be used with the logarithm transformation of the time elapsed between visit 1 and visit 2, LOG_YRS_BTWN_V1V2, 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). To calculate adjusted incidence rate (per 1000 person-years) for individuals who are at population mean age (47 years), female, Dominican, without diabetes at baseline, and at Bronx field center, we replace YRS_BTWN_V1V2 with 1000, re-calculate LOG_YRS_BTWN_V1V2, and then set the covariates at desired levels with the *at* option in the *margins* command.

```
fvset base last gendernum diabetes2_indicator bkgrd1_c7 centernum
svyset psu_id [pw=weight_norm_overall_v2], strata(strat)
gen log_yrs_btwn_v1v2=ln(yrs_btwn_v1v2)

svy, subpop(if keep_data_ckd==1): poisson incident_ckd_v1v2 age ib(3).bkgrd1_c7 i.gendernum
i.diabetes2_indicator i.centernum, offset(log_yrs_btwn_v1v2)
```

```
poisson, irr

replace yrs_btwn_v1v2 = 1000
replace log_yrs_btwn_v1v2=ln(yrs_btwn_v1v2)
margins, at(age=47 bkgrd1_c7=0 gendernum=0 diabetes2_indicator=0 centernum=1)
```

```
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(  12,    620)   =      18.48
                                                Prob > F          =     0.0000

------------------------------------------------------------------------------
             |             Linearized
incident_c~2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0240089   .0044903     5.35   0.000     .0151911    .0328267
             |
   bkgrd1_c7 |
          0  |  -.7283791   .3386085    -2.15   0.032    -1.393315   -.0634433
          1  |  -.4198869    .306493    -1.37   0.171    -1.021756    .1819827
          2  |  -.4874721     .31553    -1.54   0.123    -1.107088     .132144
          4  |  -.2981173   .2595414    -1.15   0.251    -.8077868    .2115522
          5  |  -.7572921   .3169922    -2.39   0.017    -1.379779   -.1348049
          6  |  -1.466931   .4871189    -3.01   0.003    -2.423501   -.5103604
             |
 0.gendernum |   .0121653   .1211163     0.10   0.920    -.2256745    .2500051
0.diabetes~r |  -1.091643   .1159862    -9.41   0.000    -1.319408   -.8638768
             |
   centernum |
          1  |   .7549296   .2950837     2.56   0.011     .1754646    1.334395
          2  |   .2384536   .1718053     1.39   0.166    -.0989257    .5758328
          3  |    .566181   .3150683     1.80   0.073    -.0525282     1.18489
             |
       _cons |  -5.061415   .2808449   -18.02   0.000    -5.612919   -4.509912
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(  12,    620)   =      18.48
                                                Prob > F          =     0.0000

------------------------------------------------------------------------------
             |             Linearized
incident_c~2 |        IRR   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.024299   .0045994     5.35   0.000     1.015307    1.033371
             |
   bkgrd1_c7 |
          0  |   .4826907   .1634432    -2.15   0.032      .248251    .9385273
          1  |   .6571212    .201403    -1.37   0.171     .3599621    1.199593
          2  |    .614177   .1937913    -1.54   0.123       .33052    1.141273
          4  |   .7422143   .1926354    -1.15   0.251     .4458437    1.235594
          5  |   .4689345   .1486486    -2.39   0.017     .2516341    .8738864
          6  |   .2306323   .1123453    -3.01   0.003     .0886108    .6002792
             |
 0.gendernum |    1.01224   .1225987     0.10   0.920     .7979778    1.284032
0.diabetes~r |   .3356647   .0389325    -9.41   0.000     .2672934    .4215247
             |
   centernum |
          1  |   2.127462   .6277793     2.56   0.011       1.1918    3.797696
          2  |   1.269285   .2180698     1.39   0.166       .90581    1.778611
          3  |   1.761527   .5550012     1.80   0.073     .9488276    3.270328
             |
       _cons |   .0063366   .0017796   -18.02   0.000     .0036504    .0109994
log_yrs_btwn_v1v2|          1  (offset)
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.


Adjusted predictions

Number of strata   =        20                  Number of obs     =     11,551
Number of PSUs     =       651                  Population size   = 11,574.514
Model VCE    : Linearized                       Design df         =        631

Expression   : Predicted number of events, predict()
at           : age             =          47
               bkgrd1_c7       =           0
               gendernum       =           0
               diabetes2_indicator=           0
               centernum       =           1

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   6.833462   1.424082     4.80   0.000     4.036949    9.629976
------------------------------------------------------------------------------
```

The estimated incidence rate ratio for CKD comparing those without diabetes at baseline and those with diabetes is 0.336, with the associated confidence interval (0.267, 0.421), after adjusting for age, field center, Hispanic/Latino background, and sex. For individuals who are at population mean age (47 years), female, Dominican, without diabetes at baseline, and at Bronx field center, the adjusted incidence rate is 6.83 per 1000 person-years.

4.1.4. Mplus

The *ANALYSIS: TYPE = COMPLEX* statement in Mplus and the outcome specified through the *COUNT* statement is invoked to fit Poisson regression model using complex survey procedures. Design variables are specified through the statements *STRAT*, *CLUSTER*, and *WEIGHT*.

INCIDENT_CKD_V1V2 (renamed to ckd_v1v2) is modelled through the *MODEL:* statement as the count outcome, specified through the *COUNT* statement. Since we are interested in making inference for those without CKD at visit 1, domain variable KEEP_DATA_CKD (renamed to keep_ckd) is specified in the *SUBPOPULATION* statement, with ‘EQ 1’ indicating subpopulation of interest.

More decimal places can only be viewed by saving the output as a text file (named as “REGCOEFF.dat” in the example code) through the *savedata* statement and invoking the *format* statement.

```
! survey poisson
DATA:
FILE IS sol_mplus.dat;
VARIABLE:
! variables in the same order of as created in the dataset;
NAMES = STRAT PSU_ID AGE BMI weight_v2 yrs_v1v2 ckd_v1v2
BMI_V2V1 RBMI_V2V1 KEEP_DATA keep_ckd gender_0 gender_1
center_1 center_2 center_3 center_4
bkc7_0 bkc7_1 bkc7_2 bkc7_3 bkc7_4 bkc7_5 bkc7_6 d2_ind_0 d2_ind_1 ly_v1v2;
! specify what variables we need to use in the analysis;
USEVARIABLES = STRAT PSU_ID AGE weight_v2
ckd_v1v2 gender_0 center_1 center_2 center_3
bkc7_0 bkc7_1 bkc7_2 bkc7_4 bkc7_5 bkc7_6 d2_ind_0 ly_v1v2;
! specify design features;
SUBPOPULATION = keep_ckd EQ 1;
CLUSTER = PSU_ID;
STRAT = STRAT;
WEIGHT = weight_v2;
COUNT = ckd_v1v2;
MISSING = ALL (999);
ANALYSIS:
! survey method used;
TYPE = COMPLEX;
ESTIMATOR=MLR;
!specify the model;
MODEL:
ckd_v1v2 on AGE bkc7_0 bkc7_1 bkc7_2 bkc7_4 bkc7_5 bkc7_6
gender_0 d2_ind_0 center_1 center_2 center_3 ly_v1v2@1;
SAVEDATA:
FORMAT IS f10.5;
RESULTS ARE Yourpath\REGCOEFF.dat;
```

```
SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                       10071

Number of dependent variables                                    1
Number of independent variables                                 13
Number of continuous latent variables                            0

MODEL RESULTS

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

```
CKD_V1V2   ON
   AGE                0.024      0.004      5.347      0.000
   BKC7_0            -0.728      0.339     -2.151      0.031
   BKC7_1            -0.420      0.306     -1.370      0.171
   BKC7_2            -0.487      0.316     -1.545      0.122
   BKC7_4            -0.298      0.260     -1.149      0.251
   BKC7_5            -0.757      0.317     -2.389      0.017
   BKC7_6            -1.467      0.487     -3.011      0.003
   GENDER_0           0.012      0.121      0.100      0.920
   D2_IND_0          -1.092      0.116     -9.412      0.000
   CENTER_1           0.755      0.295      2.558      0.011
   CENTER_2           0.238      0.172      1.388      0.165
   CENTER_3           0.566      0.315      1.797      0.072
   LY_V1V2            1.000      0.000    999.000    999.000

Intercepts
   CKD_V1V2          -5.061      0.281    -18.022      0.000
```

The estimated incidence rate ratio for CKD comparing those without diabetes at baseline and those with diabetes is exp(-1.092) = 0.336. The associated 95% confidence interval is (exp(-1.092-invnormal(0.975)*0.116)=0.267, exp(-1.092 +invnormal(0.975)*0.116)=0.421), after adjusting for age, field center, Hispanic/Latino background, and sex. For individuals who are at population mean age (47 years), female, Dominican, without diabetes at baseline, and at Bronx field center, the adjusted incidence rate is exp(-5.061+0.024*47+0.012-0.728-1.092+0.755)*1000 = 6.83 per 1000 person-years. Note that invnormal denotes inverse cumulative standard normal distribution, and invnormal(0.975) = 1.96.

4.2. Model-based Procedures

In this section we fit a Poisson regression model to estimate incidence rate ratio and calculate adjusted incidence rates of CKD (INCIDENT_CKD_V1V2) using the model-based procedure of weighted GEE with robust variance estimation that accounts for clustering within PSUs, instead of using complex survey procedures as done in previous section 4.1. See section 1.4 for a brief description of these procedures and their differences. We fit the model using SAS, R and Stata. Note that the point estimates are identical, and robust standard error estimates are the same up to the 2nd significant figure among those from model-based procedures in this section 4.2., and those from complex survey procedures in section 4.1.

4.2.1. SAS

In SAS, the model-based procedure PROC GENMOD is used. The CLASS statement is used to specify the categorical variables; the WEIGHT statement is used to specify the subject-level sampling weight; the MODEL statement is used to specify the analysis model and the DIST option is used to specify the marginal distribution of the outcome; and finally, the cluster level, PSU_ID, and working correlation structure (independent, denoted with ind) are specified at SUBJECT and CORR options of the REPEATED statement to obtain robust variance estimation that accounts for clustering within PSUs. When the Poisson regression is used, note that: (1) the marginal distribution of the outcome is specified as Poisson distribution and the log link function is used; (2) the analysis is restricted to the subset of participants free of CKD at baseline, which can be specified at the WHERE statement; and (3) the DESCENDING option is necessary if the interest lies in modeling event rate of INCIDENT_CKD_V1V2 = 1. Otherwise, event rate of INCIDENT_CKD_V1V2 = 0 would be modeled instead. (4) An offset of time elapsed between baseline and V2 (logt, generated in a separate data step as the logarithm of YRS_BTWN_V1V2) is specified for modeling the rate of incidence as in contrast to modeling the incidence itself.

```
data sol;
set worklib.sol;
logt = log(YRS_BTWN_V1V2);
run;


proc genmod data=sol descending ;
where KEEP_DATA_CKD=1;
class PSU_ID GENDER BKGRD1_C7 (ref = '3') DIABETES2_INDICATOR  CENTERNUM;
weight WEIGHT_NORM_OVERALL_V2;
model INCIDENT_CKD_V1V2 = AGE BKGRD1_C7 GENDER DIABETES2_INDICATOR CENTERNUM
/ OFFSET=logt dist=poisson link=log;
repeated subject=PSU_ID / corr=ind ;
run;
```

Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-5.0614	0.2829	-5.6159	-4.5070	-17.89	<.0001
AGE		0.0240	0.0045	0.0151	0.0329	5.30	<.0001
BKGRD1_C7	0	-0.7284	0.3384	-1.3917	-0.0651	-2.15	0.0314
BKGRD1_C7	1	-0.4199	0.3110	-1.0295	0.1898	-1.35	0.1770
BKGRD1_C7	2	-0.4875	0.3207	-1.1160	0.1411	-1.52	0.1285
BKGRD1_C7	4	-0.2981	0.2589	-0.8056	0.2093	-1.15	0.2495

Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
BKGRD1_C7	5	-0.7573	0.3190	-1.3825	-0.1321	-2.37	0.0176
BKGRD1_C7	6	-1.4669	0.4869	-2.4212	-0.5126	-3.01	0.0026
BKGRD1_C7	3	0.0000	0.0000	0.0000	0.0000	.	.
GENDER	F	0.0122	0.1206	-0.2242	0.2486	0.10	0.9197
GENDER	M	0.0000	0.0000	0.0000	0.0000	.	.
DIABETES2_INDICATOR	0	-1.0916	0.1165	-1.3200	-0.8633	-9.37	<.0001
DIABETES2_INDICATOR	1	0.0000	0.0000	0.0000	0.0000	.	.
CENTERNUM	1	0.7549	0.2936	0.1794	1.3305	2.57	0.0101
CENTERNUM	2	0.2385	0.1699	-0.0945	0.5714	1.40	0.1604
CENTERNUM	3	0.5662	0.3181	-0.0573	1.1897	1.78	0.0751
CENTERNUM	4	0.0000	0.0000	0.0000	0.0000	.	.

Refer to section 4.1. for estimation and interpretation of incidence rate ratio and adjusted incidence rate.

4.2.2. R

In R, the *geeglm* function from R package "*geepack*" is used. The "weights = WEIGHT_NORM_OVERALL_V2" statement indicates the subject-level sampling weight; the "data=sol" and "id=PSU_ID" statements indicate the working dataset and cluster level respectively. The default of this function is to estimate robust variance, and "id=PSU_ID" identifies the cluster level (PSU) for the robust variance. When the Poisson regression is used, note that: (1) the marginal distribution of the outcome is specified as Poisson distribution and the log link function is used; (2) the analysis is restricted to the subset of the data in which the subjects does not have CKD at baseline, which can be specified by "subset=(KEEP_DATA_CKD==1)". (3) an offset item log(YRS_BTWN_V1V2) is necessary for modeling the rate of event.

```
sol$DIABETES2_INDICATOR <- relevel(factor(sol$DIABETES2_INDICATOR), ref='1')
sol$BKGRD1_C7 <- relevel(factor(sol$BKGRD1_C7), ref='3')

sol <- sol[(is.na(sol$BKGRD1_C7)==0),]
```

```
model = geeglm(INCIDENT_CKD_V1V2 ~ AGE + BKGRD1_C7 + GENDER +
DIABETES2_INDICATOR + CENTER + offset(log(YRS_BTWN_V1V2)),
               weights=WEIGHT_NORM_OVERALL_V2, data=sol, id=PSU_ID,
               family=poisson(link = "log"),
               subset=(KEEP_DATA_CKD==1))
summary(model)
```

```
Coefficients:
                     Estimate Std.err   Wald Pr(>|W|)
(Intercept)           -5.0614  0.2958 292.70  < 2e-16 ***
AGE                    0.0240  0.0048  25.04  5.6e-07 ***
BKGRD1_C70            -0.7284  0.3397   4.60   0.0320 *
BKGRD1_C71            -0.4199  0.3046   1.90   0.1681
BKGRD1_C72            -0.4875  0.3188   2.34   0.1262
BKGRD1_C74            -0.2981  0.2707   1.21   0.2708
BKGRD1_C75            -0.7573  0.2998   6.38   0.0115 *
BKGRD1_C76            -1.4669  0.4923   8.88   0.0029 **
GENDERF                0.0122  0.1223   0.01   0.9208
DIABETES2_INDICATOR0  -1.0916  0.1179  85.71  < 2e-16 ***
CENTERB                0.7549  0.3062   6.08   0.0137 *
CENTERC                0.2384  0.1755   1.85   0.1743
CENTERM                0.5662  0.3052   3.44   0.0635 .  ---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Refer to section 4.1. for estimation and interpretation of incidence rate ratio and adjusted incidence rate.

4.2.3.Stata

In State, the *meglm* command is used. The *pw* option is used to specify the subject-level sampling weight; the *vce* option is used to indicate the use of robust variance estimator; and *cluster psu_id* is used to specify the cluster level (PSU) for the robust variance. When the Poisson regression is used, note that: (1) the marginal distribution of the outcome is specified as Poisson distribution by *family(poisson)* and the log link function is specified by *link (log)*; (2) the analysis is restricted to the subset of the participants free of CKD at baseline, and those ineligible subjects can be excluded from the analysis by using the *drop* command before the analysis is conducted; and (3) The *offset* option is used with the logarithm transformation of the time elapsed between visit 1 and visit 2, LOG_YRS_BTWN_V1V2, for modeling the rate of event. Incidence-rate ratios can be requested by using the option *irr* (either with the original *meglm* command call, or by using the statement *meglm, irr* after the Poisson regression was fit). The calculation of adjusted incidence rate is the same as in section 4.1.3.

```
gen log_yrs_btwn_v1v2=ln(yrs_btwn_v1v2)
drop if keep_data_ckd == 0
```

```
meglm incident_ckd_v1v2 age ib3.bkgrd1_c7 ib1.gendernum ib4.centernum 0.diabetes2_indicator
[pw=weight_norm_overall_v2], offset(log_yrs_btwn_v1v2) family(poisson) link(log) vce(cluster psu_id)

meglm, irr

replace yrs_btwn_v1v2 = 1000
replace log_yrs_btwn_v1v2=ln(yrs_btwn_v1v2)
```

```
margins, at(age=47 bkgrd1_c7=0 gendernum=0 diabetes2_indicator=0 centernum=1)
```

```
Mixed-effects GLM                               Number of obs     =     10,071
Family:                 Poisson
Link:                       log

                                                Wald chi2(12)     =     228.44
Log pseudolikelihood = -2004.0589               Prob > chi2       =     0.0000
                               (Std. Err. adjusted for 650 clusters in psu_id)
------------------------------------------------------------------------------
             |               Robust
incident_c~2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0240089   .0045292     5.30   0.000     .0151318     .032886
             |
   bkgrd1_c7 |
          0  |  -.7283791   .3386963    -2.15   0.032    -1.392212   -.0645465
          1  |  -.4198869   .3112891    -1.35   0.177    -1.030002    .1902286
          2  |  -.4874721   .3209381    -1.52   0.129    -1.116499    .1415551
          4  |  -.2981173   .2591009    -1.15   0.250    -.8059457    .2097111
          5  |  -.7572921    .319221    -2.37   0.018    -1.382954   -.1316305
          6  |  -1.466931   .4872719    -3.01   0.003    -2.421966   -.5118953
             |
 0.gendernum |   .0121653   .1207138     0.10   0.920    -.2244295      .24876
             |
   centernum |
          1  |   .7549296   .2938718     2.57   0.010     .1789515    1.330908
          2  |   .2384536   .1699986     1.40   0.161    -.0947376    .5716448
          3  |    .566181   .3183734     1.78   0.075    -.0578193    1.190181
             |
0.diabetes~r |  -1.091643   .1166149    -9.36   0.000    -1.320204   -.8630815
       _cons |  -5.061415     .28311   -17.88   0.000    -5.616301    -4.50653
log_yrs_btwn_v1v2|          1  (offset)
------------------------------------------------------------------------------

Mixed-effects GLM                               Number of obs     =     10,071
Family:                 Poisson
Link:                       log

                                                Wald chi2(12)     =     228.44
Log pseudolikelihood = -2004.0589               Prob > chi2       =     0.0000
                               (Std. Err. adjusted for 650 clusters in psu_id)
------------------------------------------------------------------------------
             |               Robust
incident_c~2 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.024299   .0046393     5.30   0.000     1.015247    1.033433
             |
   bkgrd1_c7 |
          0  |   .4826907   .1634856    -2.15   0.032      .248525    .9374925
          1  |   .6571212   .2045547    -1.35   0.177     .3570061    1.209526
```

```
           2 |    .614177   .1971128    -1.52   0.129      .327424    1.152064
           4 |   .7422143   .1923084    -1.15   0.250     .4466653    1.233322
           5 |   .4689345   .1496937    -2.37   0.018     .2508365    .8766649
           6 |   .2306323   .1123806    -3.01   0.003      .088747    .5993586
             |
 0.gendernum |    1.01224   .1221913     0.10   0.920     .7989719    1.282434
             |
   centernum |
           1 |   2.127462   .6252009     2.57   0.010     1.195963    3.784477
           2 |   1.269285   .2157767     1.40   0.161     .9096115    1.771178
           3 |   1.761527   .5608233     1.78   0.075     .9438205    3.287677
             |
0.diabetes~r |   .3356647   .0391435    -9.36   0.000     .2670809    .4218601
       _cons |   .0063366    .001794   -17.88   0.000     .0036381    .0110367
log_yrs_btwn_v1v2|          1  (offset)
------------------------------------------------------------------------------


Adjusted predictions                            Number of obs     =     10,071
Model VCE    : Robust

Expression   : Predicted mean, predict()
at           : age             =          47
               bkgrd1_c7       =           0
               gendernum       =           0
               centernum       =           1
               diabetes2_indicator=           0

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   6.833462   1.422525     4.80   0.000     4.045365    9.621559
------------------------------------------------------------------------------
```

Refer to section 4.1. for estimation and interpretation of incidence rate ratio and adjusted incidence rate.

5. Multilevel Modelling (MLM) for Change in Continuous Outcomes

MLM can be used to analyze data coming from complex survey designs involving multistage sampling with unequal sampling probabilities and within-cluster dependence. See section 1.2 for a brief description of MLM. In this chapter, we present both design-based complex survey procedures and model-based non-survey procedures for analyzing change from visit 1 to visit 2 in continuous outcomes using MLM in SAS, R, and Stata, when available in each software. See section 1.4 for a brief description of these procedures and their differences. For illustration we use the BMI difference between visit 2 and visit 1. Specifically, we examine the effect of baseline age (AGE) on the change in BMI after adjusting for sex (GENDERNUM, Male as the reference level), field center (CENTERNUM, San Diego as the reference level) and baseline BMI (BMI). We model the difference in BMI between visit 2 and visit 1, denoted as BMI_V2V1 and defined as BMI_V2 - BMI. Because the time elapsed 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.

5.1.Cluster Levels and Random Effects

In this section, we provide sample code for fitting the following MLMs for change in continuous outcomes:

1. MLM with two random effects, one for BG and one for HH cluster
2. MLM with one random effect for HH cluster

MLM with two random effects for BG and HH clusters require sampling weights at all three levels (BG, HH, SUB) and are also referred to as 3-level MLM in the literature. MLM with one random effect for HH cluster, also referred to as 2-level MLM, require sampling weights at two levels (HH and SUB). In section 1.2., details on the cluster levels and multilevel sampling weights are explained.

The analysis dataset *sol* includes an indicator variable KEEP_DATA with = 1 identifying the subpopulation of interest – those with no missing covariates and outcome (11,212 participants (ID) from 7,386 households (HH_ID) nested under 652 block groups (PSU_ID); see section 1.6. for analytic file creation and derived variables. Design features such as PSU stratifications (STRAT) and SSU stratifications (LISTNUM) are also included.

Note: the default option when incorporating the study design for SAS, R, and Stata is sampling with replacement (WR).

5.2.Multilevel Modelling with two Random Effects (3-Level MLM)

5.2.1. Complex Survey Procedures

In this section, we illustrate how to fit a 3-level MLM for BMI change between visits 1 and 2 with two random effects using complex survey procedures. The two random effects are at the block group and household levels, respectively. The three sampling weights needed are at the block group, household, and subject levels. We present sample code and results in Stata, which to our knowledge is the only software available to run this type of analysis.

5.2.1.1. Stata

To specify HCHS/SOL study design levels, use the *svyset* command, in the order of highest to lowest level, separated by "||" sign. First the level-3 cluster, block group (PSU_ID) with its corresponding sampling weight (WEIGHT_3MLM_BG_V2), and stratification (STRAT). Then the level-2 cluster, household (HH_ID) with its sampling weight (WEIGHT_3MLM_HH_V2). Lastly the level-1 subject (ID) with its corresponding sampling weight (WEIGHT_MLM_SUB_V2).

The MLM is fit using the *meglm* command with the usual syntax. The prefix *svy* is invoked to account for the complex survey procedures specified with the *svyset* command. Domain variable KEEP_DATA is specified in the *subpop* option before the *meglm* command to indicate subpopulation of interest. After the model covariates, specify a random intercept at the block group level identified by PSU_ID with '|| psu_id:', then a random intercept at the household level identified by HH_ID with '|| hh_id:'. The order in which the levels are specified (from left to right) is important — *meglm* assumes that HH_ID is nested within PSU_ID.

By default, *meglm* sets the smallest numerical level of each of the class variables as the reference level. To identify a class variable and change its reference level in this procedure, invoke *ib* option.

```
svyset psu_id, weight (weight_3mlm_bg_v2) strata(strat) || hh_id, weight (weight_3mlm_hh_v2) || id,
weight (weight_mlm_sub_v2)

svy, subpop (keep_data): meglm bmi_v2v1 age ib1.gendernum ib4.centernum bmi yrs_btwn_v1v2 ||
psu_id: || hh_id:
```

```
Survey: Mixed-effects GLM

Number of strata   =        20                  Number of obs     =     11,623
Number of PSUs     =       652                  Population size   =  18,706.14
```

```
                                                Subpop. no. obs   =     11,212
                                                Subpop. size      = 18,001.216
                                                Design df         =        632
                                                F(   7,    626)   =      79.82
                                                Prob > F          =     0.0000

------------------------------------------------------------------------------
             |             Linearized
    bmi_v2v1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0406448   .0026101   -15.57   0.000    -.0457704   -.0355192
 0.gendernum |   .1720994   .0605897     2.84   0.005     .0531178     .291081
             |
   centernum |
          1  |   -.076517   .0974969    -0.78   0.433     -.267974    .1149401
          2  |  -.1827363   .0829531    -2.20   0.028    -.3456333   -.0198393
          3  |   .3921352   .0895609     4.38   0.000     .2162622    .5680082
             |
         bmi |  -.1043499   .0081497   -12.80   0.000    -.1203537   -.0883461
yrs_btwn_v~2 |   .1124002   .0516468     2.18   0.030     .0109802    .2138203
       _cons |      4.488   .4234283    10.60   0.000     3.656503    5.319496
-------------+----------------------------------------------------------------
psu_id       |
   var(_cons)|   .0127855   .0270503                      .0002006    .8148305
-------------+----------------------------------------------------------------
psu_id>      |
hh_id        |
   var(_cons)|   .6847898    .187979                      .3994376    1.173994
-------------+----------------------------------------------------------------
var(e.bmi_~1)|    7.08014    .283365                      6.544994    7.659042
------------------------------------------------------------------------------
```

Note that in MLM, fixed effects have an interpretation conditional on random effects. For example, this result indicates that after adjusting for sex, center, baseline BMI, and years elapsed between visits, and conditional on block group and household memberships, a one-year increment in age at baseline is associated with a decrease of 0.041 kg/m^2 in the change in BMI. This result indicates that older age at baseline is associated with less BMI change on average 6 years later.

With the variance estimates for the random effects, we can estimate the within-block group variability and the within-household variability. The total variance (7.778) is the sum of the variances of block group (0.013), household (0.685) and the error term (7.08). Almost all the variability in the changes in BMI is explained by differences between individuals. The conditional correlation between changes in BMI of individuals from different households in the same block group is 0.0017, which is calculated by dividing the block-group variance (0.013) by the total variance (7.778). The conditional correlation between changes in BMI of two different individuals from the same households is given as 0.685 / (0.013 + 0.685 + 7.08) = 0.088. That is, of the variability in changes in BMI that is not explained by age, sex, center, baseline BMI, and years elapsed between visits, 0.17% is due to unobserved block-group-specific attributes, and 8.8% is due to unobserved household-specific attributes. These results indicate that changes in BMI of individuals from the same household are more similar than those from different households but in the same block group.

5.2.2. Model-based Procedures

In this section, we illustrate how to fit a 3-level MLM with two random effects for BMI change from visit 1 to visit 2 using the model-based procedure of weighted analysis with robust variance estimation that accounts for clustering within PSUs, instead of using complex survey procedures as done in previous section 5.2.1. See section 1.4 for a brief description of these procedures and their differences. The two random effects are at the block group and household levels. Three sampling weights are used at the block group, household, and subject levels. Robust variance estimation is used to account for clustering at the block group level.

In SAS, the model-based GLIMMIX procedure (generalized linear mixed effect models) can be used to fit a MLM. However, due to the large sample size of HCHS/SOL, it is not feasible computationally to fit the MLM.

In R, the *mix* function from R package “WeMix” can be used for MLM, but to our knowledge it only allows random effects to be specified as crossed instead of nested – which is not appropriate for our data structure.

Hence, we only present code and results from Stata in the following. Note that for both fixed and random effects, the point estimates are identical, and robust standard error estimates are the same up to the 2nd significant figure between those from Stata model-based procedures in this section 5.2.2. and those from Stata complex survey procedures in section 5.2.1.

5.2.2.1. Stata

The MLM with two random effects is fit using the *meglm* command and the usual syntax, but without the *svy* component. The *drop* statement is used to carry out the exclusion criterion (KEEP_DATA=0) to select the subpopulation of interest. The subject level sampling weight (WEIGHT_MLM_SUB_V2) is specified via the *pw* option. A random intercept at the block group level identified by PSU_ID is specified with ‘|| psu_id:’, followed by the sampling weight (WEIGHT_3MLM_BG_V2) at this level. A random intercept at the household level identified by HH_ID is specified with ‘|| hh_id:’, followed by the sampling weight (WEIGHT_ 3MLM_HH_V2) at this level. Note that the sampling weights for levels with random effects are specified via the *pw* option. The order (from left to right) assumes that HH_ID is nested within PSU_ID. Clustering on the block group level can be accounted for by specifying the clustering variable PSU_ID in the *vce(cluster variable-name)* option, which requests the robust variance estimation that accounts for clustering.

```
drop if keep_data == 0

meglm bmi_v2v1 age ib1.gendernum ib4.centernum bmi yrs_btwn_v1v2 [pw=weight_mlm_sub_v2] ||
psu_id:, pw(weight_3mlm_bg_v2) || hh_id:, pw(weight_3mlm_hh_v2)  vce(cluster psu_id)
```

```
Mixed-effects GLM                              Number of obs     =     11,212
Family:                Gaussian
Link:                  identity
-------------------------------------------------------------
                |     No. of       Observations per Group
 Group Variable |     Groups    Minimum    Average    Maximum
----------------+--------------------------------------------
         psu_id |        648          1       17.3        156
          hh_id |      7,386          1        1.5         10
-------------------------------------------------------------

Integration method: mvaghermite                Integration pts.  =          7

                                               Wald chi2(7)      =     554.74
Log pseudolikelihood = -43979.792              Prob > chi2       =     0.0000
                               (Std. Err. adjusted for 648 clusters in psu_id)
------------------------------------------------------------------------------
             |               Robust
    bmi_v2v1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0406448   .0026651   -15.25   0.000    -.0458683   -.0354213
 0.gendernum |   .1720994   .0605646     2.84   0.004      .053395    .2908038
             |
   centernum |
          1  |   -.076517   .0964238    -0.79   0.427    -.2655041    .1124702
          2  |  -.1827363   .0852037    -2.14   0.032    -.3497326   -.0157401
          3  |   .3921352   .0895002     4.38   0.000     .2167179    .5675524
             |
         bmi |  -.1043499   .0082023   -12.72   0.000    -.1204261   -.0882737
yrs_btwn_v~2 |   .1124002   .0520614     2.16   0.031     .0103618    .2144387
       _cons |      4.488    .419428    10.70   0.000     3.665936    5.310063
-------------+----------------------------------------------------------------
psu_id       |
   var(_cons)|   .0127855   .0267705                      .0002111    .7744344
-------------+----------------------------------------------------------------
psu_id>      |
hh_id        |
   var(_cons)|   .6847898   .1895399                      .3980678    1.178033
-------------+----------------------------------------------------------------
var(e.bmi_~1)|    7.08014   .2954596                        6.5241    7.683572
------------------------------------------------------------------------------
```

5.3. Multilevel Modelling with one Random Effect (2-Level MLM)

5.3.1. Complex Survey Procedures

In this section, we illustrate how to fit a 2-level MLM for BMI change between visits 1 and 2 with one random effect using complex survey procedures. The random effect is at the household level. The two sampling weights are at the household and subject levels. We present code and results in R and Stata. Note that the results using R and Stata are slightly different because they maximize different objective functions: Stata maximizes a pseudo-likelihood while R maximizes a profile pairwise composite likelihood.

5.3.1.1.R

The *svy2lme* function from R package "*svylme*" is used to fit the MLM with one random effect. Design variables including level identifiers (HH_ID, ID), level weights (WEIGHT_2MLM_HH_V2, WEIGHT_MLM_ SUB_V2), and household stratification (LISTNUM) are specified through the *svydesign* function to generate a design object, which is then invoked in *svy2lme*. The order (from left to right) assumes that ID is nested within HH_ID.

The domain variable KEEP_DATA is specified through the *subset* function from the "survey" package, with '==1' indicating subpopulation of interest. A random intercept at the household level is specified through the "(1|HH_ID)" syntax. A nested level structure is specified in the "method =" option, as it is not the default setup.

Categorical variables (GENDERNUM, CENTERNUM) need to be set as factors with desired reference level for model fitting with *svy2lme*, which cannot specify class variables.

```
sol$GENDERNUM <- as.factor(sol$GENDERNUM)
sol <- within(sol, GENDERNUM <- relevel(GENDERNUM, ref = "1"))
sol$CENTERNUM <- as.factor(sol$CENTERNUM)
sol <- within(sol, CENTERNUM <- relevel(CENTERNUM, ref = "4"))

sol.design <-svydesign(id=~HH_ID+ID,
weights=~WEIGHT_2MLM_HH_V2+WEIGHT_MLM_SUB_V2,
                       strata=~LISTNUM, data=sol)
OneRE_Survey <- svy2lme(BMI_V2V1~ (1|HH_ID) + AGE + GENDERNUM + CENTERNUM +
BMI + YRS_BTWN_V1V2, design=sol.design, method = "nested")

coef(OneRE_Survey,random=TRUE)  # print point estimates for variance
components ($s2 contains variance for the residual; $varb contains the
variance-covariance matrix for the random effects)
OneRE_Survey  # print standard errors of random effects, and fixed effects
```

```
> coef(OneRE_Survey,random=TRUE)
$s2
[1] 6.847683

$varb
            (Intercept)
(Intercept)    0.613654

> OneRE_Survey
Linear mixed model fitted by pairwise likelihood
Formula: BMI_V2V1 ~ (1 | HH_ID) + AGE + GENDERNUM + CENTERNUM + 
    BMI + YRS_BTWN_V1V2
Random effects:
            Std.Dev.
(Intercept)   0.7834
Residual: 2.6168
 Fixed effects:
                   beta        SE      t        p
(Intercept)    4.110435  0.716590  5.736 9.69e-09
AGE           -0.041254  0.004286 -9.626  < 2e-16
GENDERNUM0     0.155617  0.097919  1.589  0.11201
CENTERNUM1     0.137586  0.155678  0.884  0.37681
CENTERNUM2    -0.008621  0.161492 -0.053  0.95743
CENTERNUM3     0.502723  0.143016  3.515  0.00044
BMI           -0.093823  0.014238 -6.590 4.41e-11
YRS_BTWN_V1V2  0.104158  0.071877  1.449  0.14731
```

5.3.1.2. Stata

Specify the design variables using the *svyset* command, in the order of highest to lowest level, separated by “||” sign. First the level-2 cluster, household identifier HH_ID, the household level sampling weight (WEIGHT_2MLM_HH_V2), and stratification

(LISTNUM). Then the level-1 unit subject with identifier ID, and the subject level sampling weight (WEIGHT_MLM_SUB_V2).

The MLM can then be fit using the *meglm* command with the usual syntax. The prefix *svy* is invoked to account for the complex survey procedures specified with the *svyset* command. Domain variable KEEP_DATA is specified in the *subpop* option before the *meglm* command to indicate subpopulation of interest. Invoke *ib* option to set reference levels for categorical covariates. After the model covariates, specify a random intercept at the household level identified by HH_ID with '|| hh_id:'.

```
svyset hh_id, weight(weight_2mlm_hh_v2) strata(listnum) || id, weight(weight_mlm_sub_v2)
svy, subpop(keep_data):meglm bmi_v2v1 age ib1.gendernum ib4.centernum bmi yrs_btwn_v1v2 || hh_id:
```

```
Survey: Mixed-effects GLM
Number of strata   =         2                Number of obs     =     11,623
Number of PSUs     =     7,576                Population size   = 54,356.644
                                              Subpop. no. obs   =     11,212
                                              Subpop. size      =  52,330.74
                                              Design df         =      7,574
                                              F(   7,   7568)   =      77.12
                                              Prob > F          =     0.0000

------------------------------------------------------------------------------
             |             Linearized
    bmi_v2v1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age | -.0391828   .0028034   -13.98   0.000    -.0446783   -.0336873
 0.gendernum |   .1741528   .0652273     2.67   0.008     .0462893    .3020163
             |
   centernum |
          1  |  -.089458    .098569    -0.91   0.364    -.2826805    .1037645
          2  | -.1416074   .0901346    -1.57   0.116    -.3182962    .0350813
          3  |  .4226248   .0902455     4.68   0.000     .2457185     .599531
             |
         bmi | -.1018099    .008348   -12.20   0.000    -.1181743   -.0854454
yrs_btwn_v~2 |  .1524935   .0468532     3.25   0.001     .0606483    .2443388
       _cons |  4.073591   .3956936    10.29   0.000     3.297922     4.84926
-------------+----------------------------------------------------------------
       hh_id |
   var(_cons)|  .6521151   .2052923                       .3518153    1.208743
-------------+----------------------------------------------------------------
var(e.bmi_~1)|  7.123565   .2989418                       6.561013    7.734352
------------------------------------------------------------------------------
```

This result indicates that after adjusting for sex, center, baseline BMI, and years elapsed between visits, and conditional on household memberships, a one-year increment in age at baseline is associated with a decrease of 0.039 kg/m^2 in the change in BMI. This result indicates that older age at the baseline is associated with less BMI change on average 6 years later.

With the variance estimates for the random effects, we can estimate the conditional correlation between changes in BMI of two different patients from the same households: 0.652 / (0.652 + 7.12) = 0.084. That is, of the variability in changes in BMI that is not explained by age, sex, center, baseline BMI, and elapsed time between visits, 8.4% is due to unobserved household-specific attributes.

5.3.2. Model-based Procedures

In this section, we illustrate how to fit a 2-level MLM with one random effect for BMI change from visit 1 to visit 2 using the model-based procedure of weighted regression with robust variance estimation that accounts for clustering within HHs, instead of using complex survey procedures as done in previous section 5.3.1. See section 1.4 for a brief description of these procedures and their differences. The random effect is at the household level. Two sampling weights are used at the household and subject levels. Robust variance estimation is used to account for clustering at the household level. We present sample code and results using SAS, R, Stata. Note that for both fixed and random effects, the point estimates are identical, and robust standard error estimates are the same up to the 2nd significant figure between those from model-based procedures in this section 5.3.2. and those from Stata complex survey procedures in section 5.3.1.

5.3.2.1. SAS

The model-based procedure GLIMMIX is used, while accounting for clustering on the household level.

WARNING: this procedure is computationally intensive and will take a considerable amount of time with HCHS/SOL data. SAS's default of 2G limit on the virtual memory that can be used by a SAS session may not be sufficient for complex models, resulting in error messages in the log indicating insufficient memory. The sample code includes a few options to alleviate this issue.

In the procedure statement, "noclprint" option is invoked to suppresses the display of the levels for variables specified in the *class* statement to save computational resources. To request the robust variance estimators, invoke the "empirical=classical" option, and the highest-level cluster is accounted for by default. It is recommended to invoke the "method = quad" estimation option, with at least 5 quadrature points specified through the "(qpoints=)" sub-option for reliable estimates. However, increasing the number of quadrature points will significantly increase the amount of computation.

In the *where* statement, KEEP_DATA is specified to select the subpopulation of interest. In the *class* statement, the cluster variable (HH_ID) is specified. Class variables, if any, in the model should be specified here as well. By default, GLIMMIX sets the last category 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 the 'ref = ' option in the *class* statement. In the *model* statement, the sampling weight at subject level (WEIGHT_MLM_SUB_V2) is specified by the "obsweight= " option, and the "solution" option requests the fixed-effects parameters be produced. The method for computing denominator degrees of freedom for the tests of fixed effects is changed from the default containment method to the between-within method through the “ddfm=bw” option to save computational resources.

A random intercept at household level is specified in the *random* statement with the “int” option, with the “subject= ” option identifying HH_ID as the clusters, followed by the “weight = ” option setting the sampling weight (WEIGHT_2MLM_HH_V2) at this level.

```
proc glimmix data = sol method=quad(qpoints=5) noclprint empirical=classical;
where KEEP_DATA = 1;
class HH_ID CENTERNUM GENDERNUM;
model BMI_V2V1 = AGE GENDERNUM CENTERNUM BMI YRS_BTWN_V1V2
/obsweight=WEIGHT_MLM_SUB_V2
 ddfm=bw solution;
random int/subject=HH_ID weight=WEIGHT_2MLM_HH_V2;
RUN;
```

Number of Observations Read	11212
Number of Observations Used	11212

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
Intercept	HH_ID	0.6521	0.2053
Residual		7.1236	0.2989

Solutions for Fixed Effects							
Effect	Participant's Field Center - numeric	Gender (0=Female, 1=Male)	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept			4.0736	0.3957	7382	10.30	<.0001
AGE			-0.03918	0.002803	44941	-13.98	<.0001
GENDERNUM		F	0.1742	0.06522	2003	2.67	0.0076
GENDERNUM		M	0	.	.	.	.
CENTERNUM	B		-0.08946	0.09856	7382	-0.91	0.3641
CENTERNUM	C		-0.1416	0.09013	7382	-1.57	0.1162
CENTERNUM	M		0.4226	0.09024	7382	4.68	<.0001
CENTERNUM	S		0	.	.	.	.
BMI			-0.1018	0.008349	44941	-12.19	<.0001
YRS_BTWN_V1V2			0.1525	0.04685	44941	3.25	0.0011

5.3.2.2. R

The *mix* function from R package “WeMix” is used to fit MLM with one random effect. KEEP_DATA is specified through the *subset* function with ‘==1’ to select the subpopulation of interest. A random intercept at the household level is specified through the “(1|HH_ID)” syntax. The “cWeights=TRUE” option requests the function to use conditional weights. The sampling weights (WEIGHT_ 2LEVEL_HH_V2, WEIGHT_MLM_SUB_V2) are specified in the “weights= ” option in the order from left to right corresponding with low to high level. The default of this function is to estimate robust variance, and “(1|HH_ID)” identifies the cluster level (HH) for the robust variance.

Indicator variables would need to be created for categorical variables in model fitting with *mix*, which cannot specify class variables.

Categorical variables (GENDERNUM, CENTERNUM) need to be set as factors with desired reference level for model fitting with *mix*, which cannot specify class variables.

```
sol$GENDERNUM <- as.factor(sol$GENDERNUM)
sol <- within(sol, GENDERNUM <- relevel(GENDERNUM, ref = "1"))
sol$CENTERNUM <- as.factor(sol$CENTERNUM)
sol <- within(sol, CENTERNUM <- relevel(CENTERNUM, ref = "4"))
```

```
OneRE_Weight <- mix(BMI_V2V1 ~ AGE + GENDERNUM + CENTERNUM + BMI +
YRS_BTWN_V1V2 +(1|HH_ID),
    data=subset(sol,KEEP_DATA == 1), weights=c("WEIGHT_MLM_SUB_V2",
"WEIGHT_2MLM_HH_V2"), cWeights=TRUE)

summary(OneRE_Weight)
```

```
Variance terms:
 Level        Group        Name Variance Std. Error Std.Dev.
     2     HH_ID (Intercept)      0.6521     0.2045   0.8075
     1     Residual               7.1236     0.2990   2.6690
Groups:
 Level        Group n size mean wgt sum wgt
     2        HH_ID   7386    5.087   37573
     1          Obs  11212    4.667   52331

Fixed Effects:
               Estimate Std. Error t value
(Intercept)    4.073591   0.395705  10.295
AGE           -0.039183   0.002799 -14.001
GENDERNUM0     0.174153   0.065023   2.678
CENTERNUM1    -0.089458   0.098667  -0.907
CENTERNUM2    -0.141607   0.090216  -1.570
CENTERNUM3     0.422625   0.090262   4.682
BMI           -0.101810   0.008351 -12.192
YRS_BTWN_V1V2  0.152494   0.046841   3.256

lnl= -127856.61
Intraclass Correlation= 0.08387
```

5.3.2.3. Stata

The MLM with one random effect is fit using the *meglm* command with the usual syntax, but without the *svy* component. KEEP_DATA is specified through the drop statement to select the subpopulation of interest. The subject level sampling weight (WEIGHT_MLM_SUB_V2) is specified via the *pw* option. A random intercept at the household level identified by HH_ID is specified with '|| hh_id:', followed by the sampling weight (WEIGHT_2MLM_HH_V2) at this level specified via the *pw* option. Clustering on the household level can be accounted for by specifying the clustering variable HH_ID in the *vce(cluster variable-name)* option, which requests the robust variance estimation that accounts for clustering.

```
drop if keep_data == 0
meglm bmi_v2v1 age ib1.gendernum ib4.centernum bmi yrs_btwn_v1v2 [pw=weight_mlm_sub_v2] || hh_id:,
pw(weight_2mlm_hh_v2) vce(cluster hh_id)
```

```
Mixed-effects GLM                               Number of obs     =     11,212
Family:                Gaussian
Link:                  identity
Group variable:           hh_id                 Number of groups  =      7,386

                                                Obs per group:
                                                              min =          1
                                                              avg =        1.5
                                                              max =         10

Integration method: mvaghermite                 Integration pts.  =          7

                                                Wald chi2(7)      =     539.91
Log pseudolikelihood = -127856.61               Prob > chi2       =     0.0000
                       (Std. Err. adjusted for 7,386 clusters in hh_id)
------------------------------------------------------------------------------
             |               Robust
    bmi_v2v1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0391828   .0028033   -13.98   0.000    -.0446771   -.0336884
 0.gendernum |   .1741528   .0652245     2.67   0.008     .0463152    .3019904
             |
   centernum |
          1  |   -.089458   .0985614    -0.91   0.364    -.2826347    .1037187
          2  |  -.1416074   .0901325    -1.57   0.116     -.318264    .0350491
          3  |   .4226248   .0902383     4.68   0.000     .2457609    .5994887
             |
         bmi |  -.1018099   .0083494   -12.19   0.000    -.1181744   -.0854453
yrs_btwn_v~2 |   .1524935   .0468515     3.25   0.001     .0606662    .2443208
       _cons |   4.073591    .395681    10.30   0.000     3.298071    4.849112
-------------+----------------------------------------------------------------
       hh_id |
   var(_cons)|   .6521151   .2053046                      .3518371    1.208668
-------------+----------------------------------------------------------------
var(e.bmi_~1)|   7.123565   .2989203                      6.561138    7.734205
------------------------------------------------------------------------------
```

6. 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; (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. Kaplan–Meier estimator can be used to estimate the survival function; 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 model-based procedures can be used as viable alternatives. 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, weighted analysis with robust variance estimation using model-based procedures estimated the finite sample parameters well.

In this chapter, we present sample code for obtaining Kaplan-Meier estimator and fitting Cox regression model. Since we cannot find any complex survey procedures that provide reasonable Kaplan-Meier estimates, we only provide sample code for the model-based procedures in SAS, R, and Stata. We note that since the model-based procedures we use for obtaining the Kaplan-Meier estimates do not have an option for obtaining the robust variance, we can only obtain the correct point estimates. Hence the sample code should only be used if one is interested in providing some descriptive statistics by plotting Kaplan-Meier curves without confidence intervals. In this chapter, we also provide sample code for fitting Cox regression model using both complex survey procedures in SAS, SUDAAN, R, Stata, and Mplus as well as model-based procedures in SAS, R, and Stata. For illustration we use diabetes incidence between visit 1 and visit 2 as the outcome of interest.

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

In this section, we present different definitions for diabetes incidence and introduces the variables needed for diabetes incidence analysis. 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 Chapter 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. The following is how this pair of variables are defined for a prevalent case at baseline.

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:

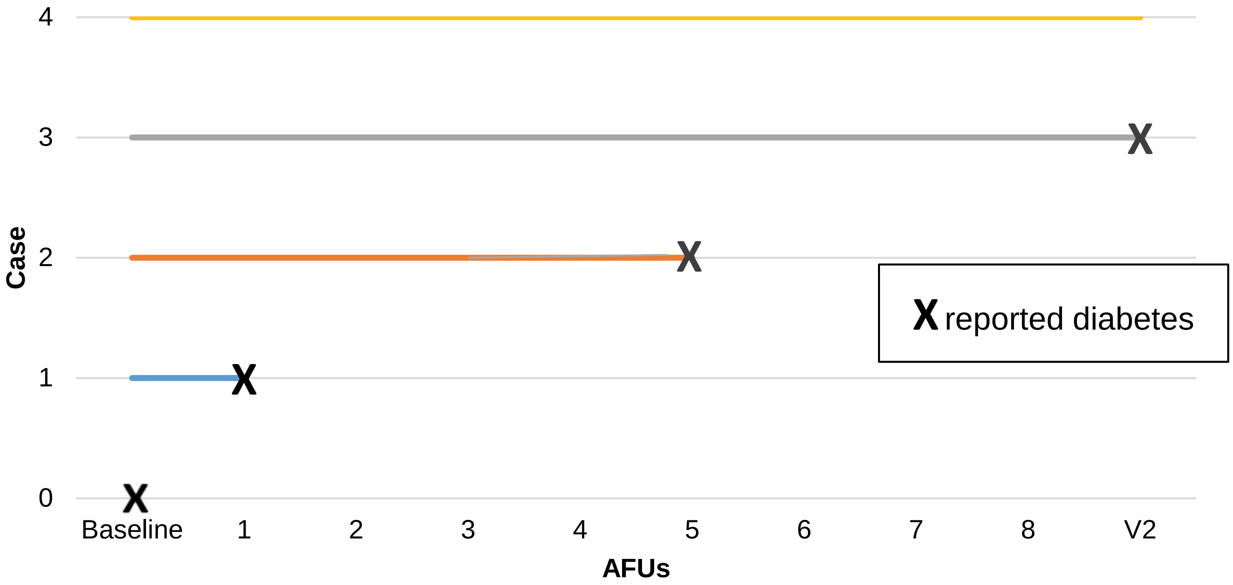
DIABETES5_TIME_V2 = AFU5 time - Visit 1 time, and
DIABETES5_INDICATOR_V2 = 1;

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.



<u>Definition 3 (DIABETES3)</u>: 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 is 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.

In this chapter, we use Definition 5 to illustrate in the examples. 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 obtain the Kaplan-Meier curve for time to incident diabetes based on Definition 5 and 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, sex, Hispanic/Latino background group, education, and income.

The following example code creates the analysis dataset that will be used throughout Chapter 6. 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;
```

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.

6.2.Kaplan–Meier Estimator

In this section, as we noted in introduction for chapter 6, since there are no complex survey procedures that provide reasonable Kaplan-Meier estimates of the survival function in the software, we present Kaplan–Meier estimator of survival function from model-based procedures in SAS, R, and Stata. Since the model-based procedures for obtaining Kaplan-Meier estimates cannot specify the use of robust variance, we can only obtain the correct point estimates. Hence the sample code should only be used if one is interested in providing some descriptive statistics by plotting Kaplan-Meier curves without confidence intervals. To illustrate that the sample results are consistent among software, we provide the 5th, 25th, 50th, 75th, and 95th quantiles of Kaplan-Meier estimates of the survival function excluding censored observations as the estimates only change at events. Note that, based on the quantiles, the Kaplan-Meier estimates are very close among software using model-based procedures in section 6.2.1.

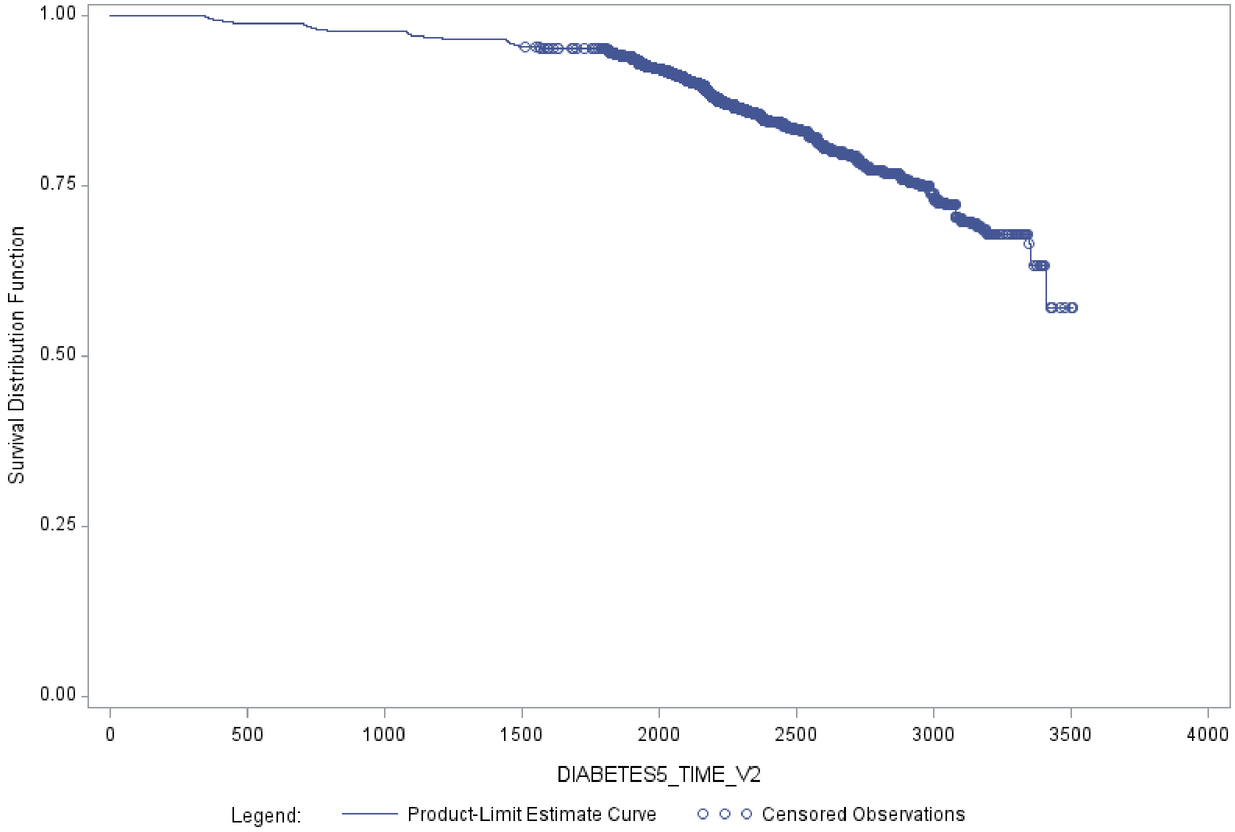
6.2.1. Model-based Procedures

6.2.1.1. SAS

The model-based procedure PROC LIFETEST with weight option is used to produce Kaplan–Meier estimator of survival function. KEEP_DATA_DIABETES5 is specified through the *where* statement to select the subpopulation of interest. Sampling weights are used in the *weight* 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. The PROC MEANS procedure produces the quantiles of Kaplan-

Meier estimates of the survival function, through specifying SURVIVAL in *var* statement, and excluding censored observations (_CENSOR_ = 0) with the *where* statement.

```
proc lifetest data=sol(where = (KEEP_DATA_DIABETES5 = 1)) notable
plots=(survival(atrisk test nocensor)) outsurv = sol_km_weighted;
weight WEIGHT_NORM_OVERALL_V2;
time DIABETES5_TIME_V2*DIABETES5_INDICATOR_V2(0);
run;

proc means data=sol_km_weighted StackODSOutput P5 P25 P50 P75 P95;
where _CENSOR_ = 0;
var SURVIVAL;
run;
```



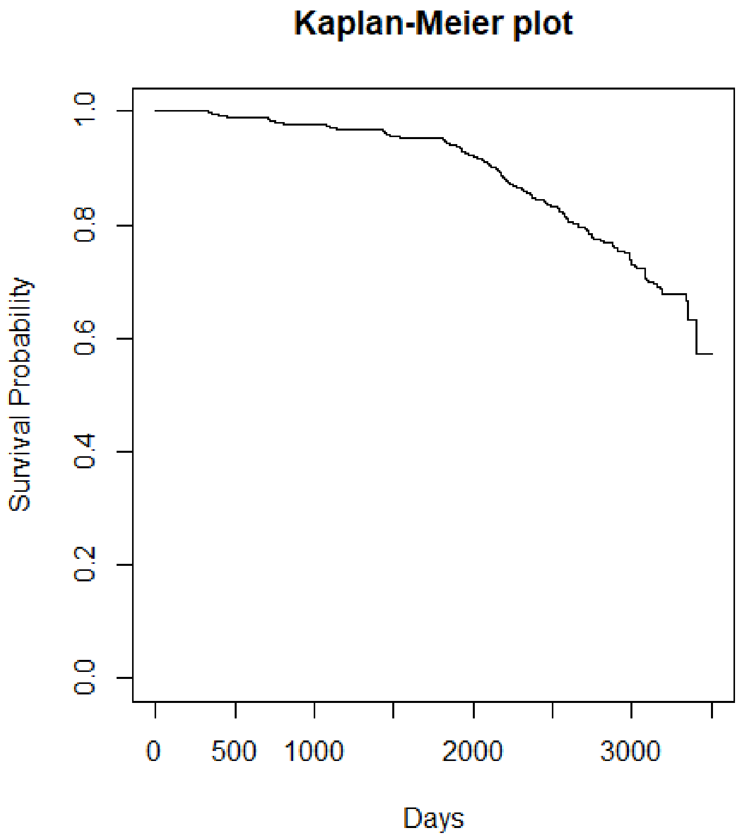
Job HC170106 run by beibo on 28APR22 at 19:15

Label	5th Pctl	25th Pctl	50th Pctl	75th Pctl	95th Pctl
Survival Distribution Function Estimate	0.770574	0.878468	0.935814	0.970288	0.993839

6.2.1.2. R

The *survfit* function from R package “survival” is used to obtain the Kaplan-Meier estimate. 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 “weights” option is set to be the sampling weights. The “subset” option is to select the subpopulation of interest, KEEP_DATA_DIABETES5 == 1. The *quantile* function produces the quantiles of Kaplan-Meier estimates of the survival function excluding censored observations (n.event !=0).

```
sol.km.weight <- survfit(Surv(DIABETES5_TIME_V2,DIABETES5_INDICATOR_V2==1) ~
1, se.fit = FALSE,
                         weights= WEIGHT_NORM_OVERALL_V2, subset =
(KEEP_DATA_DIABETES5 == 1), data=sol)

plot(sol.km.weight,
     main="Kaplan-Meier plot",
     xlab = "Days",
     ylab = "Survival Probability")


quantile(sol.km.weight$surv[which(sol.km.weight$n.event !=0)],
probs=c(0.05,0.25,0.50,0.75,0.95))
```

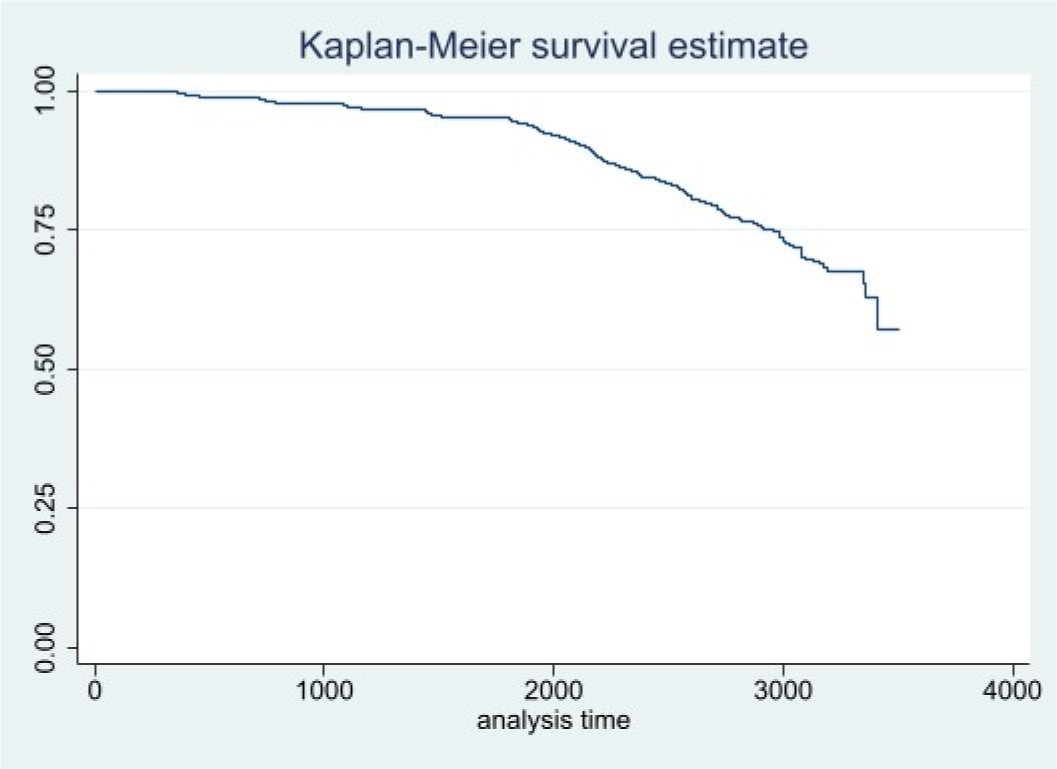


```
> quantile(sol.km.weight$surv[which(sol.km.weight$n.event !=0)],
probs=c(0.05,0.25,0.50,0.75,0.95))

       5%       25%       50%       75%       95%

0.7714357 0.8784953 0.9358137 0.9702007 0.9936946
```

6.2.1.3. Stata

The Kaplan-Meier estimate is obtained using the *sts graph* command. First, KEEP_DATA_DIABETES5 is specified through the *drop* statement to select the subpopulation of interest. We then specify DIABETES5_ INDICATOR_V2 as the event indicator, and DIABETES5_TIME_V2 as the observed event time in the *stset* command. Sampling weights are specified through the *pw* option in the *stset* command. After generating the estimates with *sts graph*, *sts list* command is used to save the results in *km*. Quantiles of Kaplan-Meier estimates of the survival function are produced with the *use km* and *summarize* commands, excluding censored observations (fail !=0).

```
drop if keep_data_diabetes5 ~= 1
stset diabetes5_time_v2 [pw=weight_norm_overall_v2], failure(diabetes5_indicator_v2)
sts graph
sts list, saving(km)
use km
summarize survivor if fail != 0,detail
```



```
summarize survivor if fail != 0,detail

                      Survivor Function
-------------------------------------------------------------
      Percentiles      Smallest
 1%      .692648       .5701267
 5%     .7679646       .6302448
10%      .813907        .652378       Obs                 863
25%      .873292       .6636853       Sum of Wgt.         863

50%     .9341545                      Mean            .914972
                        Largest       Std. Dev.      .0707202
75%     .9696292       .9996141
90%     .9891261       .9997623       Variance       .0050013
95%     .9932228       .9998225       Skewness      -1.242502
99%     .9982215        .999953       Kurtosis        4.53023
```

6.3.Cox Regression

This section illustrates how to fit a Cox regression model to estimate the hazard ratio of diabetes incidence using complex survey procedures and model-based procedures.

Different tie handling methods, such as Breslow or Efron methods provide very similar results. Our examples with SAS provide results for both tie handling 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.

6.3.1. Complex Survey Procedures

In this section, we use diabetes incidence based on Definition 5 as an example to illustrate the complex survey procedures for fitting Cox regression model. Specifically, we present examples and sample code using SAS, SUDAAN, R, Stata, and Mplus for such analysis. Note that the point estimates and robust standard error estimates are essentially identical among those from complex survey procedures in this section 6.3.1.

6.3.1.1. SAS

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 sex, 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
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

Model Information

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

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio
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

These results indicate that after adjusting for baseline age, center, sex, 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.

6.3.1.2. SUDAAN

The following code invokes the SUDAAN procedure SURVIVAL to fit Cox regression model. 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                                                                             P-value
  Variables and        Beta                      Lower 95%    Upper 95%                 T-Test
  Effects              Coeff.          SE Beta   Limit Beta   Limit Beta   T-Test B=0   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        Hazards      Lower 95%    Upper 95%
  Effects              Ratio        Limit        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
-----------------------------------------------------------
```

6.3.1.3. R

The *svycoxph* function from R package “survey” is used to fit Cox regression model. 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 <- dummy_cols(sol, select_columns = c("CENTERNUM", "GENDERNUM",
"BKGRD1_C7_NOMISS", "EDUCATION_C3_V1", "INCOME_C3_V1"))
```

```
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
```

6.3.1.4. Stata

Cox regression is fit with the *stcox* command and 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:

```
------------------------------------------------------------------------------
             |             Linearized
          _t | Haz. Ratio   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
------------------------------------------------------------------------------
```

6.3.1.5. Mplus

The *ANALYSIS: TYPE = COMPLEX* statement in Mplus is invoked to fit Cox regression model. 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 is used 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 “REGCOEFF.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 Yourpath\REGCOEFF.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
```

6.3.2. Model-based Procedures

In this 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 code 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. Note that the point estimates are identical, and robust standard error estimates are the same up to the 2nd significant figure among those from model-based procedures in this section 6.3.2., and those from complex survey procedures in section 6.3.1.

6.3.2.1. 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 section 6.3.1.1. The default reference levels and ties handling method are the same as the SURVEYPHREG procedure presented in section 6.3.1.1.

```
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
```

Model Information

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

```
                         Model Information

            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  Standard  StdErr  Chi-Square  Pr > ChiSq  Hazard
                              Estimate     Error   Ratio                           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
```

6.3.2.2. 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.

```
sol <- dummy_cols(sol, select_columns = c("CENTERNUM", "GENDERNUM",
"BKGRD1_C7_NOMISS", "EDUCATION_C3_V1", "INCOME_C3_V1"))

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
```

6.3.2.3. 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 variable-name)* option, which requests the robust variance estimation. 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 section 6.3.1.4.

```
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:

```
------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   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
------------------------------------------------------------------------------
```

REFERENCES

Liang, Kung Yee, and Scott L. Zeger. 1986. “Longitudinal Data Analysis Using Generalized Linear Models.” *Biometrika* 73(1): 13–22. https://academic.oup.com/biomet/article/73/1/13/246001 (December 2, 2021).

Rabe-Hesketh, Sophia, and Anders Skrondal. 2006. “Multilevel Modelling of Complex Survey Data.” *Journal of the Royal Statistical Society. Series A: Statistics in Society* 169(4): 805–27.

Ritz, John, and Donna Spiegelman. 2004. “Equivalence of Conditional and Marginal Regression Models for Clustered and Longitudinal Data.” : 309–23.

Schneiderman, Neil et al. 2014. “Prevalence of Diabetes among Hispanics/Latinos from Diverse Backgrounds: The Hispanic Community Health Study/Study of Latinos (HCHS/SOL).” *Diabetes care* 37(8): 2233–39. https://pubmed.ncbi.nlm.nih.gov/25061138/ (April 27, 2022).

Sterba, Sonya K. 2009. “Alternative Model-Based and Design-Based Frameworks for Inference from Samples to Populations: From Polarization to Integration.” *Multivariate Behavioral Research* 44(6): 711–40.

Zeger, Scott L., Kung-Yee Liang, and Paul S. Albert. 1988. “Models for Longitudinal Data: A Generalized Estimating Equation Approach.” *Biometrics* 44(4): 1049.

.