HCHS/SOL Analysis Methods - Visit 3

November 2024

Version 1.0

Prepared by the

HCHS/SOL Coordinating Center

Collaborative Studies Coordinating Center
UNC Department of Biostatistics

Jianwen Cai
Beibo Zhao
Daniela Sotres-Alvarez
Wenyi Xie
Franklyn Gonzalez

Table of Contents

Foreword

Note to Users

- This document is for illustration purposes for longitudinal data analysis based on data from the first three HCHS/SOL clinic visits (Baseline/Visit 1, Visit 2, Visit 3).
- Because the HCHS/SOL cohort was selected through a stratified multi-stage area probability sample design (Lavange et al., 2010), the study design specifications are accounted for in all the analysis presented.
- For cross-sectional analysis based on Visit 3 data only, please refer to *HCHS/SOL Analysis Methods at Baseline* and use Visit 3 sampling weights.
- For longitudinal analysis using only two visits, for example, Visit 1 and Visit 3 or Visit 2 and Visit 3, please refer to *HCHS/SOL Analysis Methods – Visit 2* and use Visit 3 sampling weights.
- The document is not intended for direct citation.
- Statistical program outputs used in the examples throughout this document have been modified and/or formatted for presentation and clarity.

Additional Documentations

- HCHS/SOL Analysis Methods at Baseline
  https://sites.cscc.unc.edu/hchs/node/405
- HCHS/SOL Analysis Methods - Visit 2
  https://sites.cscc.unc.edu/hchs/node/6113
- SAS (Version 9.4)
  https://support.sas.com/documentation/onlinedoc/stat/
- STATA (Version 18)
  https://www.stata.com/features/documentation/
- R (Version 4.4.1)
  https://www.r-project.org/l

List of Abbreviations

BG	Block Group
CC	Coordinating Center
CT	Classification Tree
FCS	Fully Conditional Specification
GEE	Generalized Estimating Equation
HH	Household
IPW	Inverse Probability Weighting
MAR	Missing at Random
MCAR	Missing Completely at Random
MICE	Multiple Imputation by Chained Equations
MI	Multiple Imputation
MNAR	Missing Not at Random
PSU	Primary Sampling Unit
SRS	Simple Random Sampling
SSU	Secondary Sampling Unit
SUB	Subject

1. Introduction

In the HCHS/SOL, data are collected longitudinally, with participants invited to in-person clinic visits to obtain measurements of interest such as anthropometry and biospecimens. This document contains two general parts. The first part (Chapter 2) describes the calculation of Visit 3 sampling weights. For how to conduct **cross-sectional analysis** for HCHS/SOL data involving Visit 3 data only, please refer to *HCHS/SOL Analysis Methods at Baseline* and use Visit 3 sampling weights. The second part (Chapters 3 and 4) provides guidelines on **longitudinal analysis** with repeated measures for HCHS/SOL data involving more than two clinic visits, focusing on modeling a continuous outcome over time. For how to conduct **longitudinal analysis** for HCHS/SOL data involving only two clinic visits, for example, Visit 1 and Visit 3 data only or Visit 2 and Visit 3 data only, focusing on modelling the difference, rate of change, incident event odds ratio, or incidence rate, please refer to *HCHS/SOL Analysis Methods - Visit 2* and use Visit 3 sampling weights.

Because the HCHS/SOL cohort was selected through a stratified multi-stage area probability sample design (Lavange et al., 2010), the study design specifications are accounted for in all the presented analysis. Sample codes and results using readily available software (e.g., SAS, Stata, R) are provided.

1.1. Inferential Framework

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. However, complex survey procedures either do not exist or have not been implemented in commercial software to fit some models using longitudinal data. Simulation studies were conducted at the **Coordinating Center (CC)** to examine the prospect of using non-survey model-based procedures as alternatives for finite-population estimates. The simulation results, which will be communicated in a separate document, show that the non-survey model-based procedures can provide reasonable estimation and inference as long as the sampling weights and correlation in the repeated measures are accounted for in the analysis. In this document, we present the use of the model-based procedures as tools to obtain finite-population estimates.

1.2. Modelling Approaches

Two statistical modelling approaches are commonly adopted to analyze longitudinal data with repeated measures, the **marginal approach** modeling the population-averaged longitudinal trend and the **conditional approach** modeling the subject-specific longitudinal trend. The marginal approach describes the linear relationship 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. The conditional approach incorporates random effects to capture between-subject heterogeneity in response trend. The random effects are usually assumed to follow some parametric distribution. The coefficients of the covariates in the model (betas) represent subject-specific effects, quantifying how changes in covariates within a person affect individual responses conditioning on the random effects. By explicitly modeling the within-cluster correlation structure through random effects, this approach provides insights into how the responses within a person are correlated. The interpretation of covariate effects is specific to each subject rather than averaged across the population. The choice between the conditional and marginal approaches depends on whether or not the correlation of the responses within clusters is of interest. When the response variable is continuous and the link function is identity function, the beta coefficients in the marginal model are the same as the fixed effects in the conditional model.

Generalized Estimating Equation (GEE) is a marginal approach for longitudinal analysis with repeated measures (Liang & Zeger, 1986). GEE estimates the relationship of a mean response with the covariates through a quasi-likelihood function and accounts for the non-independence of units within clusters (e.g., repeated observations within participants) through the specification of a working correlation structure. GEE can provide asymptotically unbiased coefficient estimates, which are interpreted as population-averaged effects. The variance of the coefficients can be estimated using a cluster-robust variance estimator (also known as the sandwich estimator), which is robust against misspecification of the working correlation structure. Investigators can use this marginal approach when their primary interest lies in understanding the effects of change in covariates within a person/cluster on the response, rather than quantifying the correlation between responses within clusters. In **Section 4.1**, we present the use of the marginal approach (GEE) for longitudinal analysis.

2. Cross-Sectional Analysis at Visit 3

In this chapter, we describe the calculation of Visit 3 sampling weights. We also present estimates for baseline characteristics based on Visit 1 sample using Visit 1 sampling weights and based on Visit 3 sample using Visit 3 sampling weights. We expect the estimates to be similar because both are estimating the same population parameters.

For how to conduct **cross-sectional analysis** for HCHS/SOL data involving Visit 3 data only, please refer to *HCHS/SOL Analysis Methods at Baseline* and use Visit 3 sampling weights.

2.1. Visit 3 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 over-sampled. Lastly, at the 3rd stage, **Subjects (SUBs)**, i.e., study participants, were selected in each of the eligible sampled SSUs. Participants aged 45-74 years were over-sampled. Therefore, participants were nested within household clusters, which were further nested within block group clusters with unequal probabilities of selection of BGs, HHs, and SUBs at their respective levels by this sampling design. The product of the reciprocals of the probabilities of being selected at each stage was used to calculate the base sampling weight for each participant in the cohort, which remains the same through all subsequent visits. These base weights were then adjusted for differential non-response at both the household and subject-level at baseline, forming the Visit 1 non-response adjusted sampling weights. Non-response adjustment factors were defined as the reciprocal of an estimate of the probability that a sample household agrees to be screened and to participate in the study, and the probability that a person selected into the sample agrees to participate and completes the clinic exam.

Visit 3 data collection initially began in January 2020. Due to the COVID-19 pandemic, it was paused in March 2020. To navigate the challenges posed by the pandemic, the HCHS/SOL Steering Committee decided to split Visit 3 visit into two parts: phone interview and in-person exam. The phone interviews were initiated in May 2020 and the in-person exam was resumed during the first quarter of 2021. Consequently, for Visit 3, there are two definitions of participation: (1) In-person participation only (including home visits) (N=9,090, i.e., excluding those who had phone interviews only); and (2) All participation (including phone-only interviews) (N=9,864). Of the 7,179 participants who started with phone interviews during the COVID pandemic, 6,405 (89%) later completed an in-person visit, while 774 (11%) had phone interviews only. The variables PARTICIPANT_EXAMONLY_V3 and PARTICIPANT_ALL_V3

are the indicator variables for Visit 3 participation based on the “Exam Only” definition and the “All” definition, respectively.

As with any complex survey design, the Visit 3 sampling weights account for non-response under both definitions. The non-response probability at Visit 3 is estimated using a **Classification Tree (CT)** analysis that allows an estimation of non-response profiles using all data collected at either baseline or over the course of follow-up. The idea is to form strata based on factors associated with the probability of returning for Visit 3 examination. To identify these factors, the R package 'rpart' was used to implement the CT. The advantage of the CT is that it takes interactions among factors into consideration and provides estimates for the cutpoints of continuous variables. The baseline factors considered include the following categorical variables: Hispanic/Latino Background, Age, Sex, PSU Strata, Education, Income, Health Insurance, Mental Health Status, Physical Health Status, Alcohol Use, Cigarette Use, Diabetes Status, Employment Status, Physical Activity, Prevalent Hypertension, Prevalent MI, Prevalent Stroke, Born in Mainland US, and Years Lived in US at the baseline, and AFU refusal; and the following continuous variables: Height, Weight, BMI, Cardiac Risk Ratio, eGFR, Triglycerides, HDL, LDL, Glucose, Creatinine, Urine Creatinine, Urine Micro albumin, Albumin/Creatinine Ratio, Cystatin C, and Insulin at baseline, and Log-Distance between V1 address and the last AFU address before V3 (referred to as Mobility Score hereafter).

The CT identified several factors associated with the probability of returning for Visit 3. For the "Exam Only" definition, these factors include AFU refusal, Mobility Score with a cutpoint of 3.94, Age group, Sex, PSU Strata, Cystatin C with cutpoints of 0.795, 1.09, and 1.2, and Income. For the "All" definition, the same factors were identified, except for Income. The CT divided the participants into groups, referred to as CT groups, based on identified factors (used cutpoints for continuous variables). The CT groups were further stratified by Cigarette Use. When forming the final strata for Visit 3 non-response adjustment, we imposed a minimum of 90 participants per stratum to ensure stability and reliability. If a stratum had less than 90 participants, it was combined with an adjacent tree branch that was grown from the same parent branch until sufficient number of participants was reached to form a stratum. Visit 3 non-response rates were then calculated within each of these strata.

Consistent with the approach used for overall sampling weights at baseline and Visit 2, the derivation of the overall sampling weight at Visit 3 follows the following procedure: (1) calculate Visit 3 non-response adjusted sampling weights by multiplying the Visit 1 non-response adjusted sampling weights by the inverse of the Visit 3 non-response rates, calculated for each stratum that is formed from the CT analysis described above; (2) trim extreme weights to control variability of the non-response rates; (3) calibrate to the age, gender and Hispanic/Latino background distributions from the 2010 US Census for the four study centers based on participants' Visit 1 age; (4) normalize to the overall sample.

The two definitions of participation at Visit 3 each have their corresponding overall sampling weights: WEIGHT_NORM_OVERALL_EXAMONLY_V3 for the "Exam Only" definition, and WEIGHT_NORM_OVERALL_ALL_V3 for the "All" definition. Investigators using data from clinic/home exams or biospecimens should use the "Exam Only" dataset with 9,090 participants

and the "Exam Only" sampling weights. However, if they are interested only in measures collected through phone interviews, they can use the larger dataset with 9,864 participants and the "All" sampling weights.

2.2. Comparison of Estimates for Baseline Characteristics

The sampling weights released for Visit 1 and Visit 3 data are both designed for inferences in the HCHS/SOL target population. We compared estimates for some baseline characteristics using Visit 1 sampling weights (WEIGHT_FINAL_NORM_OVERALL) with data from Visit 1 to two scenarios of those using Visit 3 sampling weights with data from Visit 3: (1) using WEIGHT_NORM_OVERALL_EXAMONLY_V3 for Visit 3 participation based on the "Exam Only" definition (**Output 2.2-1**), and (2) using WEIGHT_NORM_OVERALL_ALL_V3 for Visit 3 participation based on the "All" definition (**Output 2.2-2**).

To compare the results, we examined the difference in estimated percentages or means, defined as (value_v3 - value_v1), and the relative difference, defined as the difference divided by value_v1. Comparing the results, we note that most of these estimates have the absolute value of the difference less than 2.7% for percentages and 0.9 units for continuous variables. The absolute values of the relative difference are less than 10%, except for those with very low prevalence (Underweight, CVD, and MI) where the estimates are not stable.

2.3. Visit 3 Cross-Sectional Analysis

For how to conduct **cross-sectional analysis** for HCHS/SOL data involving Visit 3 data only, please refer to *HCHS/SOL Analysis Methods at Baseline* and use Visit 3 sampling weights.

Output 2.2-1

Baseline Characteristics of HCHS/SOL Target Population using Data from Visit 1 (Baseline) and Visit 3 "Exams Only" Participants

Baseline Characteristics	HCHS/SOL Target Population Estimates based on Visit 1 Sample (N=16415 for Visit 1 Data) N	Mean or % (95% CI)	HCHS/SOL Target Population Estimates based on Visit 3 Sample (N=9090 for Visit 3 Data) N	Mean or % (95% CI)	Difference	Relative Difference
Age (years)	16415	41.06 (40.6, 41.5)	9090	41.13 (40.5, 41.7)	0.07	0.00
Sex at birth(%)						
Male	6580	47.87 (46.8, 48.9)	3166	47.87 (46.3, 49.4)	0.00	0.00
Female	9835	52.13 (51.1, 53.2)	5924	52.13 (50.6, 53.7)	0.00	0.00
Education (%)						
Less than high school	6207	32.35 (31.0, 33.7)	3319	30.36 (28.6, 32.1)	-1.99	-0.06
High school graduate	4180	28.20 (27.1, 29.3)	2261	27.51 (26.1, 28.9)	-0.69	-0.02
Greater than high school	5937	39.46 (37.9, 41.1)	3478	42.14 (40.1, 44.1)	2.68	0.07
Hispanic/Latino background(%)						
Cuban	2348	20.02 (16.7, 23.3)	1320	19.81 (16.4, 23.2)	-0.21	-0.01
Dominican	1473	9.94 (8.6, 11.3)	836	9.96 (8.4, 11.5)	0.02	0.00
Mexican	6472	37.37 (34.2, 40.6)	3690	37.13 (33.9, 40.4)	-0.25	-0.01
Puerto Rican	2728	16.15 (14.6, 17.7)	1337	15.98 (14.3, 17.7)	-0.17	-0.01
Central American	1732	7.40 (6.3, 8.5)	984	7.63 (6.3, 8.9)	0.22	0.03
South American	1072	4.98 (4.4, 5.6)	656	4.97 (4.3, 5.7)	-0.02	-0.00
Other	503	4.13 (3.6, 4.7)	245	4.54 (3.7, 5.4)	0.40	0.10
Annual family income(%)						
<$20,000	7207	41.85 (40.1, 43.6)	3932	40.45 (38.2, 42.7)	-1.40	-0.03
$20,000-$50,000	6119	36.88 (35.6, 38.2)	3553	37.64 (35.8, 39.5)	0.76	0.02
>$50,000	1601	11.70 (10.2, 13.2)	898	12.80 (10.8, 14.8)	1.09	0.09
Not reported	1488	9.57 (8.8, 10.3)	707	9.11 (8.1, 10.1)	-0.46	-0.05

Baseline Characteristics	HCHS/SOL Target Population Estimates based on Visit 1 Sample (N=16415 for Visit 1 Data) N	Mean or % (95% CI)	HCHS/SOL Target Population Estimates based on Visit 3 Sample (N=9090 for Visit 3 Data) N	Mean or % (95% CI)	Difference	Relative Difference
Marital status(%)						
Single	4522	34.64 (33.3, 36.0)	2189	33.96 (32.2, 35.7)	-0.67	-0.02
Married or living with partner	8436	48.82 (47.3, 50.4)	5003	50.22 (48.2, 52.3)	1.39	0.03
Separated divorced, or widowed	3369	16.54 (15.6, 17.5)	1869	15.82 (14.5, 17.1)	-0.72	-0.04
Health insurance(%)	8172	50.54 (48.7, 52.4)	4552	52.64 (50.4, 54.9)	2.10	0.04
US residence >= 10 Years(%)	12490	72.34 (70.5, 74.2)	6966	72.82 (70.6, 75.0)	0.48	0.01
Language preference(%)						
Spanish	13119	74.86 (73.0, 76.7)	7545	75.09 (72.9, 77.3)	0.23	0.00
English	3296	25.14 (23.3, 27.0)	1545	24.91 (22.7, 27.1)	-0.23	-0.01
Systolic BP (mmHg)	16401	119.92 (119.4, 120.4)	9085	119.24 (118.7, 119.8)	-0.68	-0.01
Diastolic BP (mmHg)	16394	72.19 (71.9, 72.5)	9080	71.95 (71.5, 72.3)	-0.24	-0.00
Hypertension (%)	4937	24.19 (23.0, 25.4)	2730	23.80 (22.4, 25.2)	-0.39	-0.02
Treated for hypertension(%)[b]	3464	68.94 (66.8, 71.0)	1962	70.10 (67.6, 72.6)	1.17	0.02
Total cholesterol(mg/dL)	16248	194.32 (193.2, 195.4)	9022	194.52 (193.0, 196.1)	0.20	0.00
LDL-cholesterol(mg/dL)	15918	119.74 (118.8, 120.7)	8866	120.29 (119.0, 121.6)	0.54	0.00
HDL-cholesterol(mg/dL)	16246	48.48 (48.2, 48.8)	9022	48.70 (48.3, 49.1)	0.22	0.00
eGFR	16131	106.92 (106.3, 107.5)	8960	107.78 (107.1, 108.5)	0.86	0.01
Treated for hypercholesterolemia(%)[c]	1629	24.36 (22.6, 26.1)	1119	24.08 (22.1, 26.1)	-0.28	-0.01
BMI kg/m^2	16344	29.36 (29.2, 29.5)	9064	29.27 (29.1, 29.5)	-0.09	-0.00
Obesity Status (%)						
Underweight (BMI<18.5 kg/m^2)	130	1.16 (0.9, 1.4)	47	0.99 (0.6, 1.4)	-0.17	-0.15
Normal (BMI 18.5-25 kg/m^2)	3191	22.07 (21.1, 23.1)	1622	21.58 (20.2, 22.9)	-0.49	-0.02
Overweight (BMI 25-30 kg/m^2)	6116	37.19 (36.0, 38.4)	3539	38.58 (37.1, 40.1)	1.39	0.04
Obese (BM>=30 kg/m^2)	6907	39.58 (38.3, 40.9)	3856	38.85 (37.2, 40.5)	-0.73	-0.02

Baseline Characteristics	HCHS/SOL Target Population Estimates based on Visit 1 Sample (N=16415 for Visit 1 Data) N	Mean or % (95% CI)	HCHS/SOL Target Population Estimates based on Visit 3 Sample (N=9090 for Visit 3 Data) N	Mean or % (95% CI)	Difference	Relative Difference
Fasting glucose(mg/dL)	16220	102.20 (101.4, 103.0)	9010	102.00 (100.9, 103.1)	-0.21	-0.00
Diabetes - definition #2 (%)[d]	3218	14.88 (14.1, 15.7)	1738	14.88 (13.8, 16.0)	-0.00	-0.00
Diabetes - definition #4 (%)[e]	3227	14.85 (14.0, 15.7)	1744	14.83 (13.8, 15.9)	-0.02	-0.00
Treated for diabetes(%)[f]	1836	53.77 (51.3, 56.2)	956	51.77 (48.2, 55.3)	-2.00	-0.04
Waist circumference (cm)	16349	97.37 (96.9, 97.8)	9064	97.16 (96.6, 97.7)	-0.21	-0.00
Current Smoker (%)	3166	21.37 (20.3, 22.5)	1545	20.51 (19.1, 21.9)	-0.86	-0.04
Asthma (%)	2637	17.37 (16.4, 18.4)	1420	17.55 (16.2, 18.9)	0.18	0.01
COPD (%)	488	2.78 (2.4, 3.1)	252	2.65 (2.2, 3.1)	-0.13	-0.05
CVD (%)	858	4.72 (4.2, 5.2)	420	4.14 (3.5, 4.7)	-0.58	-0.12
MI (%)	384	2.34 (2.0, 2.7)	187	1.90 (1.5, 2.3)	-0.44	-0.19
Hearing Loss (%)	2799	15.06 (14.2, 15.9)	1491	14.13 (13.1, 15.2)	-0.93	-0.06

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 at baseline (Unweighted Visit 1: N=4937, Visit 3: N=2730).

[c] Denominator is restricted to participants with hypercholesterolemia at baseline (Unweighted Visit 1: N=5332, Visit 3: N=3775).

[d] ADA guideline plus scanned/transcribed medication use.

[e] ADA guideline plus self-reported medication use.

[f] Denominator is restricted to participants with diabetes (ADA guideline plus self-reported diabetes) at baseline (Unweighted Visit 1: N=3384, Visit 3: N=1833).

Source: HC331511 (18SEP24 using INV2 data)

Output 2.2-2

Baseline Characteristics of HCHS/SOL Target Population using Data from Visit 1 (Baseline) and Visit 3 "All" Participants

Baseline Characteristics	HCHS/SOL Target Population Estimates based on Visit 1 Sample (N=16415 for Visit 1 Data) N	Mean or % (95% CI)	HCHS/SOL Target Population Estimates based on Visit 3 Sample (N=9864 for Visit 3 Data) N	Mean or % (95% CI)	Difference	Relative Difference
Age (years)	16415	41.06 (40.6, 41.5)	9864	41.13 (40.6, 41.7)	0.07	0.00
Sex at birth(%)						
Male	6580	47.87 (46.8, 48.9)	3471	47.87 (46.5, 49.2)	-0.00	-0.00
Female	9835	52.13 (51.1, 53.2)	6393	52.13 (50.8, 53.5)	-0.00	-0.00
Education (%)						
Less than high school	6207	32.35 (31.0, 33.7)	3617	30.62 (28.9, 32.3)	-1.73	-0.05
High school graduate	4180	28.20 (27.1, 29.3)	2465	27.56 (26.2, 28.9)	-0.64	-0.02
Greater than high school	5937	39.46 (37.9, 41.1)	3745	41.82 (40.0, 43.7)	2.36	0.06
Hispanic/Latino background(%)						
Cuban	2348	20.02 (16.7, 23.3)	1392	19.82 (16.5, 23.1)	-0.20	-0.01
Dominican	1473	9.94 (8.6, 11.3)	922	9.95 (8.4, 11.5)	0.01	0.00
Mexican	6472	37.37 (34.2, 40.6)	4033	37.23 (34.0, 40.5)	-0.14	-0.00
Puerto Rican	2728	16.15 (14.6, 17.7)	1477	16.11 (14.4, 17.8)	-0.04	-0.00
Central American	1732	7.40 (6.3, 8.5)	1048	7.63 (6.3, 8.9)	0.23	0.03
South American	1072	4.98 (4.4, 5.6)	699	5.00 (4.3, 5.7)	0.02	0.00
Other	503	4.13 (3.6, 4.7)	264	4.25 (3.5, 5.0)	0.12	0.03
Annual family income(%)						
<$20,000	7207	41.85 (40.1, 43.6)	4294	41.46 (39.4, 43.5)	-0.39	-0.01
$20,000-$50,000	6119	36.88 (35.6, 38.2)	3819	37.22 (35.5, 39.0)	0.34	0.01
>$50,000	1601	11.70 (10.2, 13.2)	976	12.34 (10.6, 14.0)	0.64	0.05
Not reported	1488	9.57 (8.8, 10.3)	775	8.97 (8.0, 9.9)	-0.59	-0.06

Baseline Characteristics	HCHS/SOL Target Population Estimates based on Visit 1 Sample (N=16415 for Visit 1 Data) N	Mean or % (95% CI)	HCHS/SOL Target Population Estimates based on Visit 3 Sample (N=9864 for Visit 3 Data) N	Mean or % (95% CI)	Difference	Relative Difference
Marital status(%)						
Single	4522	34.64 (33.3, 36.0)	2424	33.91 (32.2, 35.6)	-0.72	-0.02
Married or living with partner	8436	48.82 (47.3, 50.4)	5397	50.34 (48.4, 52.3)	1.52	0.03
Separated divorced, or widowed	3369	16.54 (15.6, 17.5)	2008	15.75 (14.5, 17.0)	-0.80	-0.05
Health insurance(%)	8172	50.54 (48.7, 52.4)	4936	51.86 (49.8, 53.9)	1.32	0.03
US residence >= 10 Years(%)	12490	72.34 (70.5, 74.2)	7548	72.32 (70.2, 74.4)	-0.01	-0.00
Language preference(%)						
Spanish	13119	74.86 (73.0, 76.7)	8142	75.72 (73.7, 77.7)	0.86	0.01
English	3296	25.14 (23.3, 27.0)	1722	24.28 (22.3, 26.3)	-0.86	-0.03
Systolic BP (mmHg)	16401	119.92 (119.4, 120.4)	9858	119.33 (118.8, 119.9)	-0.59	-0.00
Diastolic BP (mmHg)	16394	72.19 (71.9, 72.5)	9853	72.00 (71.6, 72.4)	-0.19	-0.00
Hypertension (%)	4937	24.19 (23.0, 25.4)	2951	23.72 (22.3, 25.1)	-0.47	-0.02
Treated for hypertension(%)[b]	3464	68.94 (66.8, 71.0)	2122	70.33 (68.0, 72.7)	1.39	0.02
Total cholesterol(mg/dL)	16248	194.32 (193.2, 195.4)	9787	194.87 (193.5, 196.3)	0.55	0.00
LDL-cholesterol(mg/dL)	15918	119.74 (118.8, 120.7)	9614	120.59 (119.4, 121.8)	0.84	0.01
HDL-cholesterol(mg/dL)	16246	48.48 (48.2, 48.8)	9787	48.59 (48.2, 49.0)	0.10	0.00
eGFR	16131	106.92 (106.3, 107.5)	9717	107.56 (106.8, 108.3)	0.64	0.01
Treated for hypercholesterolemia(%)[c]	1629	24.36 (22.6, 26.1)	1186	23.65 (21.7, 25.6)	-0.71	-0.03
BMI kg/m^2	16344	29.36 (29.2, 29.5)	9835	29.30 (29.1, 29.5)	-0.06	-0.00
Obesity Status (%)						
Underweight (BMI<18.5 kg/m^2)	130	1.16 (0.9, 1.4)	62	1.20 (0.8, 1.6)	0.04	0.03
Normal (BMI 18.5-25 kg/m^2)	3191	22.07 (21.1, 23.1)	1776	21.58 (20.3, 22.9)	-0.49	-0.02
Overweight (BMI 25-30 kg/m^2)	6116	37.19 (36.0, 38.4)	3795	37.77 (36.3, 39.2)	0.58	0.02
Obese (BM>=30 kg/m^2)	6907	39.58 (38.3, 40.9)	4202	39.45 (37.8, 41.1)	-0.13	-0.00

Baseline Characteristics	HCHS/SOL Target Population Estimates based on Visit 1 Sample (N=16415 for Visit 1 Data) N	Mean or % (95% CI)	HCHS/SOL Target Population Estimates based on Visit 3 Sample (N=9864 for Visit 3 Data) N	Mean or % (95% CI)	Difference	Relative Difference
Fasting glucose(mg/dL)	16220	102.20 (101.4, 103.0)	9776	102.10 (101.1, 103.1)	-0.10	-0.00
Diabetes - definition #2 (%)[d]	3218	14.88 (14.1, 15.7)	1897	15.26 (14.2, 16.3)	0.38	0.03
Diabetes - definition #4 (%)[e]	3227	14.85 (14.0, 15.7)	1904	15.19 (14.2, 16.2)	0.34	0.02
Treated for diabetes(%)[f]	1836	53.77 (51.3, 56.2)	1040	51.70 (48.3, 55.1)	-2.07	-0.04
Waist circumference (cm)	16349	97.37 (96.9, 97.8)	9837	97.23 (96.7, 97.7)	-0.13	-0.00
Current Smoker (%)	3166	21.37 (20.3, 22.5)	1685	21.00 (19.6, 22.4)	-0.37	-0.02
Asthma (%)	2637	17.37 (16.4, 18.4)	1525	17.07 (15.8, 18.3)	-0.29	-0.02
COPD (%)	488	2.78 (2.4, 3.1)	273	2.76 (2.3, 3.3)	-0.01	-0.00
CVD (%)	858	4.72 (4.2, 5.2)	454	4.05 (3.5, 4.6)	-0.67	-0.14
MI (%)	384	2.34 (2.0, 2.7)	201	1.86 (1.5, 2.3)	-0.48	-0.20
Hearing Loss (%)	2799	15.06 (14.2, 15.9)	1609	14.10 (13.1, 15.1)	-0.97	-0.06

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 at baseline (Unweighted Visit 1: N=4937, Visit 3: N=2951).

[c] Denominator is restricted to participants with hypercholesterolemia at baseline (Unweighted Visit 1: N=5332, Visit 3: N=4026).

[d] ADA guideline plus scanned/transcribed medication use.

[e] ADA guideline plus self-reported medication use.

[f] Denominator is restricted to participants with diabetes (ADA guideline plus self-reported diabetes) at baseline (Unweighted Visit 1: N=3384, Visit 3: N=1997).

SOURCE: HC331511 (18SEP24 using INV2 data)

3. Longitudinal Analysis: Introduction

In this chapter, we introduce the groundwork for conducting **longitudinal analysis** with repeated measures for HCHS/SOL data involving more than two clinic visits. We begin by discussing missing visits and related methodologies in longitudinal analysis. Next, we explore data management strategies for longitudinal analysis. Finally, we provide guidance on creating an analytic dataset, including sample code for dataset generation. All examples will utilize data from the first three HCHS/SOL clinic visits.

For how to conduct **longitudinal analysis** for HCHS/SOL data involving only two clinic visits, for example, Visit 1 and Visit 3 data only or Visit 2 and Visit 3 data only, focusing on modelling the difference, rate of change, incident event odds ratio, or incidence rate, please refer to *HCHS/SOL Analysis Methods - Visit 2* and use Visit 3 sampling weights.

3.1. Missing Visits

Participants missing follow-up visits is a common phenomenon in any longitudinal study. Data for participants who missed follow-up visit(s) will be missing. It can lead to biased estimates and reduced precision if missing visits are not accounted for properly. The missingness mechanism behind missing visits can be grouped into three categories: **Missing Completely at Random (MCAR)**, **Missing at Random (MAR)**, **Missing Not at Random (MNAR)**.

MCAR occurs when the probability of a participant missing a visit is independent of both observed and unobserved data. In other words, a participant missing a visit is a result of completely random events that are unrelated to any participant characteristics or outcomes of interest, regardless of whether they are observed or unobserved. MCAR can be partially verified if no significant differences are found when comparing the characteristics of participants with complete visits to those with missing visits. However, this verification is limited to observed variables and cannot rule out relationships with unobserved data. When MCAR holds, a complete case analysis which drops the missing records and uses only the data from participants who completed all visits, is expected to provide valid inference of the true population parameters. This approach is the default in most statistical software. However, MCAR is a strong assumption that rarely holds in practice. Moreover, using only the complete cases leads to a loss of efficiency (larger standard errors) with the extent of efficiency loss depending on the proportion of missing data.

MAR occurs when the probability of a participant missing a visit depends on observed data, but not on unobserved data. In other words, a participant missing a visit is a result of factors that are related to observed participant characteristics or outcomes of interest, but not to unobserved characteristics or outcomes that would have been collected at a missing visit. When MAR holds, statistical methods that properly account for the observed data associated with missingness can provide valid inference of the true population parameters. MAR is a less stringent assumption than MCAR and is often more plausible in longitudinal studies.

MNAR, also known as informative or non-ignorable missingness, occurs when the probability of a participant missing a visit depends on unobserved data, even after accounting for the observed data. In other words, a participant missing a visit is a result of factors that are related to unobserved participant characteristics or outcomes of interest, including those that would have been collected at a missing visit. When MNAR holds, standard statistical methods, even those that account for observed data, can provide biased inference of the true population parameters. Handling MNAR often requires more complex approaches that jointly model the outcome and missingness process, such as selection models or pattern mixture models. What approach to use depends on the scientific question of interest. MNAR is the most challenging missing data mechanism to address, and its presence cannot be definitively determined from the observed data alone. Therefore, sensitivity analyses are recommended to assess the robustness of findings under different MNAR scenarios.

In HCHS/SOL, the baseline cohort (N=16,415) has been followed over time. About 71% of the original cohort attended Visit 2 (N=11,623). About 60% of the original cohort attended Visit 3 (N=9,864), out of which 9,090 participated in the in-person exam and 774 had phone interview only. For participants who did not attend Visit 2 or/and Visit 3 or dropped out of the study, they are considered as having missing visits. An overview of missing visits with respect to the baseline cohort is presented in two ways (**Output 3.1-1**): (1) for Visit 3 in-person attendance only, and (2) for ALL Visit 3 attendance, including those with phone interview only. We assume the missing-visit mechanism is MAR and describe appropriate methods to address this type of missingness in each respective chapter.

Output 3.1-1: Missing Visits Overview

Visit 1	Visit 2	Visit 3 Exam Only	N	%
✓			4134	25.2
✓	✓		3191	19.4
✓		✓	658	4.0
✓	✓	✓	8432	51.4
Sum			**16415**	**100**

Visit 1	Visit 2	Visit 3 All	N	%
✓			3905	23.8
✓	✓		2646	16.1
✓		✓	887	5.4
✓	✓	✓	8977	54.7
Sum			**16415**	**100**

3.1.1. Multiple Imputation

Multiple Imputation (MI) is a widely used strategy to handle missingness in both outcome variables and covariates, particularly under the MAR assumption. MI involves creating multiple plausible imputed datasets, analyzing each dataset separately, and then combining the results using specific rules, e.g., Rubin's rules (Rubin, 2018). This approach accounts for the uncertainty

in the imputed values, leading to valid statistical inferences. For a detailed introduction, please refer to *Flexible Imputation of Missing Data* by Stef van Buuren (van Buuren, 2018).

Within the MI framework, various methods can be used to create the imputed datasets. One popular and flexible method is **Fully Conditional Specification (FCS)**, also known as **Multiple Imputation by Chained Equations (MICE)**. FCS operates through a sequence of univariate imputation models, assuming the existence of a joint distribution for all variables. This approach makes FCS suitable for datasets with arbitrary missing patterns. The method works by imputing missing values on a variable-by-variable basis, using iterative cycles to refine imputations. This process preserves relationships between variables in the imputed data and captures complex interdependencies. FCS can accommodate various types of variables (continuous, binary, categorical) within the same imputation model. Additionally, the method allows for the inclusion of auxiliary variables in the imputation model, potentially improving the quality of imputations.

3.2. Data Management: wide-format and long-format

For longitudinal data, there are two ways to format the data for analysis, wide-format and long-format. In the **wide-format data**, each participant has one record with separate variables for repeated measures at each follow-up visit. For example, BMI measurements at Visits 1, 2, 3 would be represented as three distinct variables: BMI_V1, BMI_V2, and BMI_V3. In contrast, in the **long-format data** there is only one variable with the measurement (BMI) and a variable to identify the clinic visit (VISIT), and there are multiple records per participant, one for each visit. For example, a participant would have one record for BMI at Visit 1, another record for BMI at Visit 2, and a third record for BMI at Visit 3.

3.3. Analytic Dataset

The following code generates the analytic dataset "sol_wide.sas7bdat", a wide-format SAS dataset with all participants from the baseline cohort (N=16,415). This dataset will be used for examples in Chapter 4.1. This dataset is created by importing variables needed for the examples from relevant investigator files (e.g., blood pressure measurements from “sbp” files) from each visit, renaming some with visit-specific suffixes (e.g., _V1, _V2, _V3) to accommodate the wide format. The modified 7-level reclassification of Hispanic/Latino background, BKGRD1_C7_NOMISS, is created to incorporate missing data into the "Mixed/Others" category. The Household ID, HH_ID, is available from the multilevel sampling weights files “mlweights” at Visit 1 and Visit 2. The Visit 3 sampling weights for Exam Only participants, WEIGHT_NORM_OVERALL_EXAMONLY_V3, is imported because the examples are based on measures from the in-person exam. A list of variables in the analytic dataset is presented in **Output 3.3-1.**

```
/* Visit 1 */
data analys_v1 (rename = (BMI = BMI_V1 SBPA5 = SBP5_V1));
      merge part_derv_inv4 sbpa_inv4 mlweights_inv4;
      by ID;
      keep PSU_ID HH_ID ID WEIGHT_FINAL_NORM_OVERALL
            CENTERNUM GENDERNUM BKGRD1_C7 AGEGROUP_C6
            US_BORN EMPLOYED EDUCATION_C3 BMI SBPA5;
run;

/* Visit 2 */
data analys_v2 (rename = (SBP5 = SBP5_V2));
      merge part_derv_v2_inv3 sbp_v2_inv3;
      by ID;
      keep ID WEIGHT_NORM_OVERALL_V2 BMI_V2 YRS_BTWN_V1V2 SBP5;
run;

/* Visit 3 */
data analys_v3 (rename = (SBP5 = SBP5_V3));
      merge part_derv_v3_inv2 sbp_v3_inv2;
      by ID;
      keep ID WEIGHT_NORM_OVERALL_EXAMONLY_V3 BMI_V3 YRS_BTWN_V1V3 SBP5;
run;

/* Analytic Dataset (wide-fomrat) */
data sol_wide;
      merge analys_v1 analys_v2 analys_v3;
      by ID;
      BKGRD1_C7NOMISS = BKGRD1_C7;
      if BKGRD1_C7NOMISS < .Z then BKGRD1_C7NOMISS = 6;
      drop BKGRD1_C7;
run;
```

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

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

Output 3.3-1: Variables in the Analytic Dataset

Variable	Description
ID	Participant ID
HH_ID	Secondary Sampling Unit (Household) ID
PSU_ID	Primary Sampling Unit (Block Group) ID
AGEGROUP_C6	Age Groups: 1=Ages 18-24, 2=Ages 25-34, 3=Ages 35-44, 4=Ages 45-54, 5=Ages 55-64, 6=Ages 65+
BKGRD1_C7NOMISS	Hispanic/Latino Background: 0=Dominican, 1=Central American, 2=Cuban, 3=Mexican, 4=Puerto-Rican, 5=South American, 6=More than one heritage/Other, DK/Refused, Missing
CENTERNUM	Participant's Field Center - numeric: 1=Bronx, 2=Chicago, 3=Miami, 4=San Diego
GENDERNUM	Sex: 0=Female, 1=Male
WEIGHT_FINAL_NORM_OVERALL	Overall Sampling Weights, Visit 1
SBP5_V1	Average Systolic (mm Hg), Visit 1
BMI_V1	BMI (kg/m^2), Visit 1
US_BORN	Born in mainland US: 0=Not born in 50 US States/DC, 1=Born in 50 US States/DC Only
EMPLOYED	Employment Status: 1=Retired and not currently employed, 2=Not retired and not currently employed, 3=Employed part-time (<=35 hours/week), 4=Employed full-time (>35 hours/week)
EDUCATION_C3	Education Status: 1=Less Than High School, 2=High School or Equivalent, 3=Greater than High School or Equivalent
WEIGHT_NORM_OVERALL_V2	Overall Sampling Weights, Visit 2
YRS_BTWN_V1V2	Elapsed time between visits 1 and 2 (years)
SBP5_V2	Average Systolic (mm Hg), Visit 2
BMI_V2	BMI (kg/m^2), Visit 2
WEIGHT_NORM_OVERALL_EXAMONLY_V3	Overall Sampling Weights, excluding those with phone interview only, Visit 3
YRS_BTWN_V1V3	Elapsed time between visits 1 and 3 (years)
SBP5_V3	Average Systolic (mm Hg), Visit 3
BMI_V3	BMI (kg/m^2), Visit 3

4. Longitudinal Analysis of Continuous Outcomes

In this chapter, we describe how to conduct **longitudinal analysis** with repeated measures for HCHS/SOL data involving more than two clinic visits, focusing on modeling a continuous outcome over time. We use Body Mass Index (BMI) as an example and provide sample codes in SAS, Stata, and R.

4.1. Marginal (GEE) Approach with MI

4.1.1. Analytic Procedure

For the longitudinal analysis of HCHS/SOL data using a marginal (GEE) approach, the CC recommends GEE combined with MI to handle missing visits, based on extensive simulation studies conducted by the CC. The full results of these simulations will be presented in a separate document. The CC advises applying GEE with MI approach to the HCHS/SOL baseline cohort (N=16,415) with Visit 1 sampling weights, following a three-step process:

Step 1 (Impute): Generate *m* imputed datasets from the wide-format analytic dataset using FCS/MICE; Impute each variable (both covariates and the outcome) with missing values that appears in the main model of interest.

The CC recommends $m = 10$. The imputation model should include all variables from the main model of interest, as well as any variables associated with the probability of missing a clinic visit, even if they are not in the main model. Using wide-format data for imputation preserves relationships between variables across different time points, allowing for a more comprehensive consideration of the longitudinal structure and ensuring that temporal dependencies and associations between variables at different visits are maintained in the imputed datasets.

For the FCS/MICE imputation process, we recommend the following regression methods based on the type of variable being imputed:

- Continuous: Linear regression
- Binary: Logistic regression
- Categorical (ordinal): Ordered logistic regression (proportional odds)
- Categorical (nominal): Multinomial (polytomous) logistic regression

Step 2 (Transform then Fit): Transform each imputed dataset from wide to long-format; apply weighted GEE (weighted with Visit 1 sampling weights) to each transformed dataset to fit the model of interest.

Step 3 (Combine): Combine the results from the *m* separate GEE analyses using Rubin's rules to obtain final estimates and standard errors, accounting for variability both within and between the imputed datasets.

A key point in MI is to appropriately specify the covariates related to the missing mechanism in the imputation model under the MAR assumption. Our simulation results showed: when the imputation model is under-specified, the resulting estimates can be biased, and the inference can be invalid; when the imputation model is correctly specified or over-specified, the resulting estimates are approximately unbiased, and the inference is valid. Since we do not know the correct model in practice, the CC recommends including all the variables in the main model as well as any variables that have potential to be associated with the missingness.

4.1.2. Analytic Example

As an example for illustration, we define the main model of interest as a longitudinal analysis examining the effect of time-varying systolic blood pressure on BMI over time across the three clinic visits (long-format: BMI; wide-format: BMI_V1, BMI_V2, BMI_V3) in the HCHS/SOL target population using the marginal (GEE) approach. The approach is weighted with Visit 1 overall sampling weights (WEIGHT_FINAL_NORM_OVERALL) and considers clustering within households (HH_ID) to account for the complex survey design. Visit 1 sampling weights are used for valid inference because the analysis includes all participants who attended the baseline clinic visit (N = 16,415).

In the main model of interest, the primary predictor of interest is systolic blood pressure over time across the three clinic visits (long-format: SBP5; wide-format: SBP5_V1, SBP5_V2, SBP5_V3), while adjusting for the following covariates:

- Baseline demographic factors: 6-level age group (AGEGROUP_C6), 7-level re-classification of Hispanic/Latino background (BKGRD1_C7NOMISS), field center (CENTERNUM), sex (GENDERNUM), US-born status (US_BORN), 4-level employment status (EMPLOYED), and 3-level education level (EDUCATION_C3)
- Time-related factor: years elapsed from Visit 1 (long-format: TIME; wide-format: YRS_BTWN_V1V2, YRS_BTWN_V1V3)

The main model of interest is:

$$g\left(E\left[Y_{it}\right]\right) = \beta_0 + \beta_1 SBP5_{it} + \beta_2 TIME_{it} + \beta_3 AGEGROUP\_C6_i +$$
$$\beta_4 BKGRD1\_C7NOMISS_i + \beta_5 CENTERNUM_i + \beta_6 GENDERNUM_i + ,$$
$$\beta_7 US\_BORN_i + \beta_8 EMPLOYED_i + \beta_9 EDUCATION\_C3_i$$

where $g(\cdot)$ is the link function appropriate for the distribution of Y_{it} (e.g., use identity link for the continuous outcome BMI) for participant i at visit t (for covariates only at baseline, t is omitted).

The coefficient β_2 for the TIME variable would be interpreted as: On average, among individuals with the same values of the covariates included in the model, the expected BMI

increased by β_2 kg/m² for each year that elapsed since Visit 1, or equivalently, for each year of aging.

We first examine the extent of missingness in the data, then follow the three steps as described in **Section 4.1.1**:

Step 1 (Impute): Generate *10* imputed datasets from the wide-format analytic dataset "sol_wide" created in SAS (see **Section 3.3)** using FCS/MICE; Impute each variable (both covariates and the outcome) with missing values using the following FCS regressions:

- Linear regression: SBP5_V1, BMI_V1, YRS_BTWN_V1V2, SBP5_V2, BMI_V2, YRS_BTWN_V1V3, BMI_V3, SBP5_V3
- Binary logistic regression: US_BORN
- Ordered logistic regression (proportional odds): EMPLOYED
- Multinomial (polytomous) logistic regression: EDUCATION_C3

Based on the simulation results, the CC recommends including the sampling weights for Visit 2 and Visit 3 in the imputation model because they reflect factors that could be related to the probability of missing the respective visit. In this example, the imputation model includes Visit 2 overall sampling weights (WEIGHT_NORM_OVERALL_V2) and Visit 3 overall sampling weights with clinic or home exams only (WEIGHT_NORM_OVERALL_EXAMONLY_V3), i.e., excluding those with phone interview only.

Specifying in the imputation models all variables in the main model of interest and additionally, the Visit 2 sampling weights WEIGHT_NORM_OVERALL_V2 and the Visit 3 sampling weights WEIGHT_NORM_OVERALL_EXAMONLY_V3.

Step 2 (Transform then Fit): Transform each imputed dataset from wide to long-format; apply weighted GEE (weighted with Visit 1 sampling weights) to each transformed dataset to fit the following model of interest:

- Outcome: BMI
- Covariates: AGEGROUP_C6, BKGRD1_C7NOMISS, CENTERNUM, GENDERNUM, US_BORN, EMPLOYED, EDUCATION_C3, SBP5, TIME
- Weight: WEIGHT_FINAL_ NORM_OVERALL
- Clusters: HH_ID
- Working correlation structure: Independent

Note that TIME is the long-format version of YRS_BTWN_V1V2 and YRS_BTWN_V1V3.

Step 3 (Combine): Combine the results from the *m* separate GEE analyses using Rubin's rules to obtain final estimates and standard errors.

The parameter estimates obtained from the process described above mostly agree across software with consistent statistical conclusions. Some minor differences are expected and can be attributed to random variation inherent in the MI process and differences in the technicalities of

method implementation, including variations in handling numerical precision and rounding, across software. Different clustering variables are used for SAS/R versus Stata. Parameter estimates accounting for household clusters cannot be done in Stata as the MI procedure from Stata has the limitation that the specified weights need to be constant within the panel variable, which is not the case if using household clusters. Thus, in Stata example, we provide results using subject clusters instead of household clusters. In the following sections, we present sample codes and results from each software.

4.1.3. SAS

```
/* Extent of Missingness */
proc means data=sol_wide n nmiss;
      var AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM GENDERNUM
            WEIGHT_FINAL_NORM_OVERALL SBP5_V1 BMI_V1
            US_BORN EMPLOYED EDUCATION_C3
            WEIGHT_NORM_OVERALL_V2 YRS_BTWN_V1V2 SBP5_V2 BMI_V2
            WEIGHT_NORM_OVERALL_EXAMONLY_V3 YRS_BTWN_V1V3 SBP5_V3 BMI_V3;
run;
```

The procedure **`proc means`** with options `n` and `nmiss` examines the extent of missingness in the dataset `sol_wide`. **Output 4.1-1** presents the results.

Output 4.1-1: SAS, Extent of Missingness

Variable	N	N Miss
AGEGROUP_C6	16415	0
BKGRD1_C7NOMISS	16415	0
CENTERNUM	16415	0
GENDERNUM	16415	0
WEIGHT_FINAL_NORM_OVERALL	16415	0
SBP5_V1	16401	14
BMI_V1	16344	71
US_BORN	16342	73
EMPLOYED	16109	306
EDUCATION_C3	16324	91
WEIGHT_NORM_OVERALL_V2	11623	4792
YRS_BTWN_V1V2	11623	4792
SBP5_V2	11591	4824
BMI_V2	11245	5170
WEIGHT_NORM_OVERALL_EXAMONLY_V3	9090	7325
YRS_BTWN_V1V3	9864	6551
SBP5_V3	9046	7369
BMI_V3	8758	7657

```
/* Step 1 */
proc mi data=sol_wide nimpute=10 seed=2024 out=sol_mi_wide;
      class AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM GENDERNUM
            US_BORN EMPLOYED EDUCATION_C3;
      var AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM GENDERNUM
            WEIGHT_FINAL_NORM_OVERALL SBP5_V1 BMI_V1
            US_BORN EMPLOYED EDUCATION_C3
            WEIGHT_NORM_OVERALL_V2 YRS_BTWN_V1V2 SBP5_V2 BMI_V2
            WEIGHT_NORM_OVERALL_EXAMONLY_V3 YRS_BTWN_V1V3 SBP5_V3 BMI_V3;
      fcs reg(SBP5_V1 BMI_V1
            WEIGHT_NORM_OVERALL_V2 YRS_BTWN_V1V2 SBP5_V2 BMI_V2
            WEIGHT_NORM_OVERALL_EXAMONLY_V3 YRS_BTWN_V1V3 SBP5_V3 BMI_V3);
      fcs logistic(US_BORN EMPLOYED /* link=logit*/);
      fcs logistic(EDUCATION_C3 / link=glogit);
run;
```

For Step 1, the procedure **`proc mi`** performs MI. The `nimpute` option specifies the number of imputations. The `seed` option sets a random seed for reproducibility (i.e., obtain the same results every time the code is run). The `out` option outputs `sol_mi_wide`, a single dataset that contains all the imputed data stacked, containing an imputation number identifier _IMPUTATION_ = 1, 2,…10 automatically generated by SAS.

The `class` statement specifies the categorical variables. The `var` statement specifies all variables to be used in the imputation model. The `fcs` statement specifies the following FCS regressions: `reg`, linear regression for continuous variables; `logistic` (with the default `logit` link), binary logistic regression for binary variables (US_BORN) and ordered logistic regression for ordinal variables (EMPLOYED); `logistic` specifying `link=glogit`, multinomial logistic regression for nominal variables (EDUCATION_C3).

```
/* Step 2 */
* Reshape data from wide to long;
data sol_mi_widetolong;
      set sol_mi_wide;

      VISIT = 1;
      SBP5 = SBP5_V1;
      TIME = 0;
      BMI = BMI_V1;
      output;

      VISIT = 2;
      SBP5 = SBP5_V2;
      TIME = YRS_BTWN_V1V2;
      BMI = BMI_V2;
      output;

      VISIT = 3;
      SBP5 = SBP5_V3;
      TIME = YRS_BTWN_V1V3;
      BMI = BMI_V3;
      output;
run;

* Fit GEE simultaneously in 10 imputed datasets ;
proc genmod data=sol_mi_widetolong;
      by _IMPUTATION_;
      class HH_ID AGEGROUP_C6(ref = '6') BKGRD1_C7NOMISS(ref = '3')
            CENTERNUM(ref = '4') GENDERNUM(ref = '0') US_BORN(ref = '0')
      EMPLOYED(ref = '1') EDUCATION_C3 (ref = '1');
      weight WEIGHT_FINAL_NORM_OVERALL;
      model BMI = AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM GENDERNUM US_BORN
            EMPLOYED EDUCATION_C3 SBP5 TIME/ dist=normal;
      repeated subject = HH_ID /corr=ind;
      ods output GEEEmpPEst=betas_mi;
run;
```

For Step 2, the `data` step transforms the wide-format imputed dataset `sol_mi_wide` (16415*10 observations because of 10 imputed files) into long format `sol_mi_widetolong` (16415*10*3 observations because of 10 imputed files and 3 visits) by assigning the visit-specific variables to their generic long-format versions (SBP5, TIME, and BMI) and creates an indicator variable VISIT to indicate to which visit an observation belongs. For Visit 1, TIME is set to 0.

The **`proc genmod`** procedure fits GEE to `sol_mi_widetolong`. The analysis is performed separately for each imputation through specifying in the `by` statement the imputation number identifier _IMPUTATION_. Reference levels can be specified in the `class` statement, e.g., `AGEGROUP_C6(ref = '6')` sets level 6 as the reference. The `weight` statement specifies Visit 1 overall sampling weights (WEIGHT_FINAL_NORM_OVERALL) for weighted GEE. The `model` statement specifies BMI as the outcome and includes all covariates of interest, assuming a normal distribution through `dist=normal`. The `repeated` statement defines the clustering variable `subject=HH_ID` for household clusters. `corr=ind` specifies an

independent working correlation structure. The `ods output` outputs the parameter estimates in the output object `GEEEmpPEst` to the dataset `betas_mi`.

```
/* Step 3 */
proc mianalyze parms(classvar=level)=betas_mi;
      class AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM GENDERNUM US_BORN
            EMPLOYED EDUCATION_C3;
      modeleffects INTERCEPT AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM
            GENDERNUM US_BORN EMPLOYED EDUCATION_C3 SBP5 TIME;
run;
```

For Step 3, the **`proc mianalyze`** procedure combines the MI results in `betas_mi` using Rubin's rules. `parms` with the `classvar=level` option is needed to correctly identify the classification levels of variables specified in the `class` statement. The `modeleffects` statement lists all the effects in the model, including the intercept and all covariates specified in **`proc genmod`** from Step 2.

After removing redundant rows and columns (for the classification levels) from the output, parameter estimates with household clusters are displayed in **Output 4.1-2**. Based on the results, the estimate for systolic blood pressure (SBP5) is 0.0293 with a standard error of 0.0045. This positive coefficient suggests that, on average, among individuals with the same values of the covariates included in the model (such as baseline age, sex, background, center, US-born status, employment, and education), and the same amount of time that have elapsed since Visit 1 or equivalently the same amount of aging (represented by 'TIME'), every 10 units (mm Hg) increase in systolic blood pressure within a person is associated with a 0.293 units (kg/m^2) increase in BMI. This effect is statistically significant ($p < .0001$).

Output 4.1-2: SAS, Parameter Estimates from GEE (household clusters) with MI

Parameter	Level	Estimate	Std Error	95% Confidence Limits		t for H0: Parameter= Theta0	Pr > |t|
INTERCEPT		24.8219	0.6732	23.4846	26.1592	36.87	<.0001
AGEGROUP_C6	1	0.0045	0.3704	-0.7237	0.7327	0.01	0.9903
AGEGROUP_C6	2	1.5670	0.3613	0.8539	2.2801	4.34	<.0001
AGEGROUP_C6	3	2.0124	0.3176	1.3870	2.6378	6.34	<.0001
AGEGROUP_C6	4	1.3321	0.3105	0.7174	1.9468	4.29	<.0001
AGEGROUP_C6	5	0.8755	0.2785	0.3286	1.4223	3.14	0.0017
BKGRD1_C7NOMISS	0	0.0404	0.3446	-0.6352	0.7161	0.12	0.9067
BKGRD1_C7NOMISS	1	0.1010	0.2925	-0.4737	0.6756	0.35	0.7301
BKGRD1_C7NOMISS	2	0.1002	0.3414	-0.5704	0.7707	0.29	0.7693
BKGRD1_C7NOMISS	4	0.5157	0.3042	-0.0812	1.1126	1.7	0.0903
BKGRD1_C7NOMISS	5	-0.9325	0.2971	-1.5156	-0.3493	-3.14	0.0018
BKGRD1_C7NOMISS	6	0.1666	0.4281	-0.6731	1.0064	0.39	0.6972
CENTERNUM	1	0.4758	0.2920	-0.0976	1.0493	1.63	0.1037
CENTERNUM	2	0.4377	0.2241	-0.0021	0.8775	1.95	0.0511
CENTERNUM	3	0.2403	0.3345	-0.4166	0.8972	0.72	0.4728
GENDERNUM	1	-1.1170	0.1448	-1.4008	-0.8331	-7.72	<.0001
US_BORN	1	1.4840	0.2409	1.0116	1.9565	6.16	<.0001
EMPLOYED	2	0.0652	0.3007	-0.5280	0.6583	0.22	0.8287
EMPLOYED	3	-0.5238	0.3372	-1.1892	0.1416	-1.55	0.1221
EMPLOYED	4	-0.1047	0.3148	-0.7258	0.5165	-0.33	0.7399
EDUCATION_C3	2	0.0076	0.1741	-0.3339	0.3492	0.04	0.965
EDUCATION_C3	3	-0.3243	0.1743	-0.6661	0.0176	-1.86	0.063
SBP5		0.0293	0.0045	0.0201	0.0384	6.47	<.0001
TIME		0.0834	0.0063	0.0706	0.0961	13.3	<.0001

4.1.4. Stata

Note: in Stata example, we provide results using subject clusters (ID) instead of household clusters (HH_ID) as the MI procedure from Stata has the limitation that the specified weights need to be constant within the panel variable, which is not the case if using household clusters.

```
import sas using "sol_wide.sas7bdat",clear
set seed 2024
rename WEIGHT_NORM_OVERALL_EXAMONLY_V3 WEIGHT_EXAMONLY_V3
mi set flong

** Extent of Missingness **
mi misstable summarize
```

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. *import sas* command loads the SAS dataset "sol_wide.sas7bdat". The *clear* option ensures that any existing data in memory is cleared before importing the new dataset. The *set seed* command sets a specific random seed for reproducibility of any subsequent random processes. The *rename* command is used to shorten the name of the variable WEIGHT_NORM_OVERALL_EXAMONLY_V3 to WEIGHT_EXAMONLY_V3, as a name of an imputation variable is not allowed to contain more than 29 characters in Stata.

The *mi set flong* command sets up the data for MI in the "flong" (full long) style, which is one of Stata's formats for storing multiply imputed data. The *mi misstable summarize* command examines the extent of missingness in the dataset and **Output 4.1-3** presents the results.

Output 4.1-3: Stata, Extent of Missingness

				Obs<.		
Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max
US_BORN	**73**		**16,342**	2	0	1
BMI_V1	**71**		**16,344**	>500	13.8199	70.34503
EMPLOYED	**306**		**16,109**	4	1	4
EDUCATION_C3	**91**		**16,324**	3	1	3
SBP5_V1	**14**		**16,401**	136	74	232
WEIGHT_NOR~2	**4,792**		**11,623**	>500	.0755914	13.82749
BMI_V2	**5,170**		**11,245**	>500	14.04012	70.94644
YRS_BTWN_V~2	**4,792**		**11,623**	>500	3.397673	9.598905
SBP5_V2	**4,824**		**11,591**	131	71	230
YRS_BTWN_V~3	**6,551**		**9,864**	>500	9.377139	15.75907
BMI_V3	**7,657**		**8,758**	>500	15.07812	67.22382
WEIGHT_EXA~3	**7,325**		**9,090**	>500	.0577422	12.61602
SBP5_V3	**7,369**		**9,046**	138	70	235

```
** Step 1 **
mi register imputed SBP5_V1 BMI_V1 US_BORN EMPLOYED EDUCATION_C3
WEIGHT_NORM_OVERALL_V2 YRS_BTWN_V1V2 SBP5_V2 BMI_V2 WEIGHT_EXAMONLY_V3
YRS_BTWN_V1V3 SBP5_V3 BMI_V3

mi register passive AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM GENDERNUM
WEIGHT_FINAL_NORM_OVERALL

mi impute chained (regress) SBP5_V1 BMI_V1 WEIGHT_NORM_OVERALL_V2
YRS_BTWN_V1V2 SBP5_V2 BMI_V2 WEIGHT_EXAMONLY_V3 YRS_BTWN_V1V3 SBP5_V3
BMI_V3 (logit) US_BORN (ologit) EMPLOYED (mlogit) EDUCATION_C3 =
i.AGEGROUP_C6 i.BKGRD1_C7NOMISS i.CENTERNUM i.GENDERNUM
WEIGHT_FINAL_NORM_OVERALL, add(10)
```

For Step 1, The *mi register imputed* command specifies all variables to be imputed. The *mi register passive* command identifies variables that are not imputed but are used in the imputation model. The *mi impute chained* command performs multivariate imputation using FCS methods: linear regression *regress* for continuous variables; logistic regression *logit* for binary variables (US_BORN); ordered logistic regression *ologit* for ordinal variables (EMPLOYED); multinomial logistic regression *mlogit* for nominal variables (EDUCATION_C3). The *add(10)* option specifies that 10 imputed datasets will be created. The non-imputed variables to be included in the imputation model are specified at the end after the = sign, with the *i.* prefix indicating the classification/categorical variables.

```
** Step 2 **
* Renaming BMI variables for easy reshape
rename BMI_V1 BMI1
rename BMI_V2 BMI2
rename BMI_V3 BMI3

* Renaming SBP5 variables for easy reshape
rename SBP5_V1 SBP51
rename SBP5_V2 SBP52
rename SBP5_V3 SBP53

* Renaming years between visits for easy reshape
rename YRS_BTWN_V1V2 TIME2
rename YRS_BTWN_V1V3 TIME3

* Creating a new variable TIME1 and setting it to 0 for all
generate TIME1 = 0

* Reshape data from wide to long;
mi reshape long BMI SBP5 TIME, i(ID) j(VISIT)
```

For Step 2, after MI, visit-specific variables are renamed to facilitate reshaping the data from wide to long format by modifying their suffixes (from _VX to X), so they can be recognized by Stata as to which visit they are referring to. For instance, BMI_V1, BMI_V2, and BMI_V3 are renamed to BMI1, BMI2, and BMI3. The time since Visit 1 variable for Visit 1 (TIME1) is created and set to 0. The *mi reshape long* command transforms the data from wide to long format. The *i(ID)* option specifies that ID is the variable that uniquely identifies subjects across

visits, and the *j(VISIT)* option creates an indicator variable VISIT to indicate to which visit an observation belongs. In Stata, fitting GEE and combining the MI results using Rubin's rules are done with a single command, explained in Step 3.

```
** Step 3 **
encode ID, gen(ID_NUM)
mi xtset ID_NUM

mi estimate: xtgee BMI ib6.AGEGROUP_C6 ib3.BKGRD1_C7NOMISS ib4.CENTERNUM
ib0.GENDERNUM ib0.US_BORN ib1.EMPLOYED ib1.EDUCATION_C3 SBP5 TIME
[pw=WEIGHT_FINAL_NORM_OVERALL], family(gaussian) corr(independent)
```

For Step 3, the *encode* command encodes the ID variable into the numeric format as a new variable ID_NUM. This is necessary so that the *mi xtset* command declares the data to be longitudinal (panel) data, with ID_NUM specified as the panel variable.

The *mi estimate*: *xtgee* command fits GEE and automatically combines the results across imputed datasets using Rubin's rules. Categorical variables are indicated by the prefix *ib*. and numeric values can be appended to indicate the reference level, e.g., ib6.AGEGROUP_C6 sets level 6 as the reference. The *[pw=WEIGHT_FINAL_NORM_OVERALL]* option applies Visit 1 overall sampling weights as probability weights for weighted GEE. The *family(gaussian)* option specifies a Gaussian (normal) distribution for the dependent variable, and *corr(independent)* specifies an independent working correlation structure for the GEE model.

Parameter estimates with subject clusters are displayed in **Output 4.1-4**. The point estimates, standard errors, and confidence intervals with household clusters from other software and those with subject clusters from Stata are similar with only slight differences in this example, and no impact on statistical significance.

Output 4.1-4: Stata, Parameter Estimates from GEE (subject clusters) with MI

BMI	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
AGEGROUP_C6						
1	-.001468	.3598383	-0.00	0.997	-.7073868	.7044507
2	1.562004	.3244852	4.81	0.000	.9259608	2.198048
3	2.000493	.2990399	6.69	0.000	1.414183	2.586803
4	1.341878	.2740565	4.90	0.000	.8046962	1.879061
5	.870342	.2715106	3.21	0.001	.3380931	1.402591
BKGRD1_C7NOMISS						
0	.0620307	.3442508	0.18	0.857	-.6127444	.7368058
1	.1569969	.2621241	0.60	0.549	-.3567926	.6707863
2	.1457897	.3124972	0.47	0.641	-.4668648	.7584442
4	.5274176	.2926485	1.80	0.072	-.0466767	1.101512
5	-.9424203	.279333	-3.37	0.001	-1.490654	-.394187
6	.2118631	.4256011	0.50	0.619	-.6226502	1.046376
CENTERNUM						
1	.4728106	.2684143	1.76	0.078	-.0536103	.9992314
2	.4608246	.2056709	2.24	0.025	.0574876	.8641616
3	.2427272	.3007764	0.81	0.420	-.3468249	.8322792
1.GENDERNUM	-1.109592	.1412209	-7.86	0.000	-1.386561	-.8326228
1.US_BORN	1.492035	.2500461	5.97	0.000	1.001728	1.982343
EMPLOYED						
2	.0531164	.2840879	0.19	0.852	-.5046084	.6108411
3	-.5349388	.3184888	-1.68	0.094	-1.160512	.0906345
4	-.1209047	.2910816	-0.42	0.678	-.6921183	.4503089
EDUCATION_C3						
2	.0531246	.1803948	0.29	0.768	-.3005232	.4067723
3	-.3116274	.1698357	-1.83	0.067	-.6445516	.0212969
SBP5	.0289858	.0035293	8.21	0.000	.0220486	.035923
TIME	.0826351	.0056093	14.73	0.000	.0713753	.093895
_cons	24.8206	.5716055	43.42	0.000	23.69872	25.94248

4.1.5. R

```
## Set up ##
library(haven)
library(dplyr)
library(tidyr)
library(skimr)
library(mice)
library(glmtoolbox)
library(mitml)

sol <- read_sas("sol_wide.sas7bdat")

# Reference levels
sol$GENDERNUM <- relevel(factor(sol$GENDERNUM), ref='0')
sol$CENTERNUM <- relevel(factor(sol$CENTERNUM), ref='4')
sol$AGEGROUP_C6 <- relevel(factor(sol$AGEGROUP_C6), ref='6')
sol$BKGRD1_C7NOMISS <- relevel(factor(sol$BKGRD1_C7NOMISS), ref='3')

sol$US_BORN <- relevel(factor(sol$US_BORN), ref='0')
sol$EMPLOYED <- relevel(factor(sol$EMPLOYED), ref='1')
sol$EDUCATION_C3 <- relevel(factor(sol$EDUCATION_C3), ref='1')

## Examine the extent of missingness ##
skim(sol)
```

In R, necessary libraries need to be loaded first. These include: 'haven' for reading data formats from other software; 'dplyr' and 'tidyr' for data manipulation 'skimr' for data summaries; 'mice' for MI using FCS; 'glmtoolbox' for GEE; 'mitml' for additional MI tools. The *read_sas* function reads the SAS dataset "sol_wide.sas7bdat" into R. The *relevel* function converts categorical variables to factors with specified reference levels, e.g., relevel(factor(sol$AGEGROUP_C6), ref='6') sets level 6 as the reference. This ensures that subsequent analyses use the correct reference categories for these variables. Finally, the *skim* function examines the extent of missingness in the dataset. **Output 4.1-5** presents part of the results.

Output 4.1-5: R, Extent of Missingness

```
── Variable type: factor ──────────────────────────────────
  skim_variable   n_missing complete_rate ordered n_unique
1 US_BORN                73         0.996 FALSE          2
2 EMPLOYED              306         0.981 FALSE          4
3 AGEGROUP_C6             0         1     FALSE          6
4 CENTERNUM               0         1     FALSE          4
5 GENDERNUM               0         1     FALSE          2
6 EDUCATION_C3           91         0.994 FALSE          3
7 BKGRD1_C7NOMISS         0         1     FALSE          7

── Variable type: numeric ─────────────────────────────────
   skim_variable                    n_missing complete_rate
 1 PSU_ID                                   0         1
 2 WEIGHT_FINAL_NORM_OVERALL                0         1
 3 BMI_V1                                  71         0.996
 4 SBP5_V1                                 14         0.999
 5 WEIGHT_NORM_OVERALL_V2                4792         0.708
 6 BMI_V2                                5170         0.685
 7 YRS_BTWN_V1V2                         4792         0.708
 8 SBP5_V2                               4824         0.706
 9 YRS_BTWN_V1V3                         6551         0.601
10 BMI_V3                                7657         0.534
11 WEIGHT_NORM_OVERALL_EXAMONLY_V3       7325         0.554
12 SBP5_V3                               7369         0.551
```

```
## Step 1 ##
# Set up MI using MICE
predMatrix <- quickpred(sol, include = c("AGEGROUP_C6", "BKGRD1_C7NOMISS",
"CENTERNUM", "GENDERNUM", "WEIGHT_FINAL_NORM_OVERALL", "SBP5_V1",
"BMI_V1", "US_BORN", "EMPLOYED", "EDUCATION_C3", "WEIGHT_NORM_OVERALL_V2",
"YRS_BTWN_V1V2", "SBP5_V2", "BMI_V2", "WEIGHT_NORM_OVERALL_EXAMONLY_V3 ",
"YRS_BTWN_V1V3", "SBP5_V3", "BMI_V3"))

methods <- make.method(sol)

# choose imputation methods, default for continuous variables is PMM
for (i in seq_along(methods)) {
  if (methods[i] == "pmm") {
    methods[i] <- "norm"
  }
}

# Modify the method for binary and categorical variables specifically
methods[c("US_BORN")] <- "logreg"
methods[c("EMPLOYED")] <- "polr"
methods[c("EDUCATION_C3")] <- "polyreg"

# Perform MI
imputed_data_wide <- mice(sol, method = methods, predictorMatrix =
predMatrix, m = 10, seed = 2024)
```

For Step 1, the *quickpred* function creates a prediction matrix, with *include* option specifying variables to be included in the imputation model. The *make.method* function sets up the default FCS methods. For continuous variables, the method is changed from the default predictive mean matching *pmm* to linear regression *norm*. In terms of other variable types, specify: logistic regression *logreg* for binary variables (US_BORN); ordered logistic regression *polr* for ordinal variables (EMPLOYED); multinomial logistic regression *polyreg* for nominal variables (EDUCATION_C3). The *mice* function performs MI, with the following options: imputation methods *method*; predictor matrix *predictorMatrix*; number of imputations *m*; random seed for reproducibility *seed*. The process results in a list object, stored as 'imputed_data_wide', that contains all the imputed data with the imputation identifier 'imp'.

```
## Step 2 ##
# Combine all imputed datasets into one data frame
imputed_data_combined <- complete(imputed_data_wide, "long")

# Transform the combined data from wide to long format
imputed_data_long_combined <- imputed_data_combined %>%
  pivot_longer(
    cols = starts_with(c("BMI_", "SBP5_")),
    names_to = c(".value", "VISIT"),
    names_pattern = "(.*)_(V\\d)"
  ) %>%
  mutate(
    VISIT = as.numeric(gsub("V", "", VISIT)),
    TIME = case_when(
      VISIT == 1 ~ 0,
      VISIT == 2 ~ YRS_BTWN_V1V2,
      VISIT == 3 ~ YRS_BTWN_V1V3
    )
  )

# Split the combined long data back into individual imputed datasets
imputed_data_long_list <- split(imputed_data_long_combined,
imputed_data_long_combined$.imp)

# Initialize lists to store GEE results
model_list <- list()

# Fit GEE to each transformed imputed dataset
for (i in 1:10) {

  imputed_data_long_i <- imputed_data_long_list[[i]]

  # Fit GEE
  model_list[[i]] <- glmgee(
    BMI ~ AGEGROUP_C6 + BKGRD1_C7NOMISS + CENTERNUM + GENDERNUM + US_BORN
+ EMPLOYED + EDUCATION_C3 + SBP5 + TIME,
    data = imputed_data_long_i,
    id = HH_ID,
    corstr = "independence",
    weight = WEIGHT_FINAL_NORM_OVERALL,
    family = gaussian(link = "identity")
  )
}
```

For Step 2, the *complete* function combines all items in the list object from Step 1 into a single data frame. The *pivot_longer* function transforms the combined data from wide to long format. This transformation creates separate rows for each visit, with variables like BMI and SBP5 now having a single column each, and a new VISIT column indicating the visit number. The time since Visit 1 (TIME) for Visit 1 is set to 0. The *split* function splits the long-format data back into a list object based on the imputation identifier 'imp'. Within a *for* loop, the *glmgee* function applies GEE to each of the transformed imputed datasets in the list object . The option *id = HH_ID* specifies household (HH_ID) clusters. The *corstr = independence* option sets the working correlation structure to independence. The *weight =*

WEIGHT_FINAL_NORM_OVERALL option applies the Visit 1 overall sampling weights in weighted GEE. The option *family = gaussian(link = identity)* specifies that the model assumes a Gaussian (normal) distribution for the outcome. The results are stored in a list object 'model_list'.

```
## Step 3 ##
pooled_results <- mitml::testEstimates(model_list, fun = summary)

# Create a data frame of coefficients
 coefficients_df <- data.frame(
   name = rownames(model_list[[1]]$coefficients),
   round(pooled_results$estimates,4)
 )
 coefficients_df
```

For Step 3, the *testEstimates* function from the *mitml* package pools the results with Rubin's rules. To include variable names in the output, which are not provided from the *testEstimates* function, a data frame 'coefficients_df' is created. This data frame combines the variable names extracted from the 'model_list' object (from the coefficients in GEE fitting) with the rounded pooled estimates (4 decimal places), providing a more interpretable summary of the results.

Parameter estimates (formatted to include variable names) accounting for household clusters are displayed in **Output 4.1-6**;

Output 4.1-6: R, Parameter Estimates from GEE (household clusters) with MI

name	Estimate	Std.Error	t.value	df	P...t..
(Intercept)	24.9201	0.6546	38.0689	115.1967	0.0000
AGEGROUP_C61	0.0238	0.3484	0.0684	3360.3064	0.9455
AGEGROUP_C62	1.5644	0.3341	4.6831	1102.4538	0.0000
AGEGROUP_C63	1.9944	0.3003	6.6405	1072.2013	0.0000
AGEGROUP_C64	1.3299	0.2730	4.8723	2846.8171	0.0000
AGEGROUP_C65	0.8722	0.2696	3.2347	2939.2093	0.0012
BKGRD1_C7NOMISS0	0.0924	0.3488	0.2650	1606.5337	0.7910
BKGRD1_C7NOMISS1	0.1273	0.2858	0.4455	1661.0571	0.6560
BKGRD1_C7NOMISS2	0.1107	0.3298	0.3358	2118.5111	0.7371
BKGRD1_C7NOMISS4	0.5772	0.3097	1.8638	717.9052	0.0628
BKGRD1_C7NOMISS5	-0.9098	0.2890	-3.1482	3929.9339	0.0017
BKGRD1_C7NOMISS6	0.1904	0.4369	0.4358	505.3188	0.6632
CENTERNUM1	0.4212	0.2864	1.4707	1419.0550	0.1416
CENTERNUM2	0.4366	0.2160	2.0212	23552.5674	0.0433
CENTERNUM3	0.2496	0.3205	0.7787	4298.2745	0.4362
GENDERNUM1	-1.1126	0.1429	-7.7871	4228.9800	0.0000
US_BORN1	1.4793	0.2448	6.0438	1122.0610	0.0000
EMPLOYED2	0.0906	0.2849	0.3178	485.9391	0.7508
EMPLOYED3	-0.4671	0.3143	-1.4863	729.4528	0.1376
EMPLOYED4	-0.0711	0.2932	-0.2426	771.3345	0.8084
EDUCATION_C32	0.0229	0.1809	0.1269	488.2124	0.8991
EDUCATION_C33	-0.3270	0.1755	-1.8627	1196.0340	0.0627
SBP5	0.0281	0.0038	7.3647	143.5960	0.0000
TIME	0.0826	0.0058	14.1730	48.1950	0.0000

References

Lavange, L. M., Kalsbeek, W. D., Sorlie, P. D., Avilés-Santa, L. M., Kaplan, R. C., Barnhart, J., Liu, K., Giachello, A., Lee, D. J., Ryan, J., Criqui, M. H., & Elder, J. P. (2010). Sample Design and Cohort Selection in the Hispanic Community Health Study/Study of Latinos. *Annals of Epidemiology*, *20*(8), 642-649. https://doi.org/10.1016/j.annepidem.2010.05.006

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13-22. https://doi.org/10.1093/biomet/73.1.13

Rubin, D. B. (2018). Multiple imputation. In *Flexible Imputation of Missing Data, Second Edition* (pp. 29-62). Chapman and Hall/CRC.

Sterba, S. K. (2009). Alternative Model-Based and Design-Based Frameworks for Inference From Samples to Populations: From Polarization to Integration. *Multivariate Behavioral Research*, *44*(6), 711-740. https://doi.org/10.1080/00273170903333574

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC Press, Taylor & Francis Group. https://books.google.com/books?id=bLmItgEACAAJ