HCHS/SOL Analysis Methods - Visit 3

February 2026

Version 2.0

Prepared by the

HCHS/SOL Coordinating Center

Collaborative Studies Coordinating Center
UNC Department of Biostatistics

Jianwen Cai
Beibo Zhao
Daniela Sotres-Alvarez
Franklyn Gonzalez
Alex Akushevich
Alvaro Clemente Quijano Angarita
Wenyi Xie

Table of Contents

Table of Contents

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.
- For time to event data using annual follow-up data or incident events, please refer to Chapter 6 of the *HCHS/SOL Analysis Methods – Visit 2* and use Visit 1 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.
- Case sensitivity: In R and Stata, variable names as well as commands are case-sensitive.

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
- HCHS/SOL Baseline Physical Activity Data Overview, Methods & Guidelines
  https://sites9.cscc.unc.edu/hchs/node/415
- 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 Updates

Version 2.0 (February 2026)

- INV file version updates for the illustrative examples:

  Visit 1 → INV5; Visit 2 → INV3; Visit 3 → INV2

- Visit 3 sampling weights:

  Modified Chapter 2 to focus on Visit 3 sampling weights

- Longitudinal methods framework:

  Expanded Chapter 3 to provide a unified framework for handling missing visits, with explicit sections on Multiple Imputation (MI) and Inverse Probability Weighting (IPW), and the addition of general procedures for marginal GEE analyses using MI and IPW.

- Continuous longitudinal outcomes:

  Expanded Chapter 4, Examples for Longitudinal Analysis of Continuous Outcomes with a detailed illustrative example and parallel implementations using MI and IPW.

- Binary longitudinal outcomes:

  Added new Chapter 5, Examples for Longitudinal Analysis of Binary Outcomes.

- Separation of Visit 1 vs. Visit 3 analytic populations:

  Separated analyses based on the Visit 1 sample and the Visit 3 sample, with corresponding examples on appropriate procedures.

List of Abbreviations

AFU	Annual Follow-Up
BG	Block Group
CART	Classification and Regression Tree
CC	Coordinating Center
CVD	Cardiovascular Disease
FCS	Fully Conditional Specification
GEE	Generalized Estimating Equation
GLM	Generalized Linear Model
HCHS	Hispanic Community Health Study
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
SOL	Study of Latinos
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 manual provides guidelines for analyses that include data collected at Visit 3. 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. 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. This manual focuses on analyses that involve data from more than two clinic visits. Specifically, Chapters 3 provides analysis recommendations on **longitudinal analysis** with repeated measures for HCHS/SOL data involving more than two clinic visits. Illustrative examples and sample codes and results using readily available software (e.g., SAS, Stata, R) for the recommended methods are provided for continuous outcomes in Chapter 4 and for binary outcomes in Chapter 5. 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 recommended analysis methods.

1.1. Inferential Framework

In all our analysis, we adopt the following perspective (randomization theory): 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; Lohr 2022). 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, the sampling design (stratification, clustering and sampling weights) needs to be accounted for during the point and variance estimation of finite-population parameters. Analytic techniques that properly incorporate these features are referred to as **design-based complex-survey procedures**. For certain model structures, particularly those involving longitudinal or highly clustered data, such complex survey procedures do not yet exist or have not been fully implemented in some statistical software.

In contrast to complex-survey procedures, **model-based procedures** assume that the observed data arise from an underlying stochastic (superpopulation) model that specifies the probability distribution of the measurements given model parameters. Inferences are drawn based on the likelihood function or estimating equations derived from that model, rather than from the randomization distribution of the sampling design. When applied appropriately, model-based procedures can provide consistent and unbiased estimates of finite-population parameters if they incorporate sampling weights to adjust for unequal selection probabilities and use robust variance estimators to account for intra-cluster correlation and model misspecification.

The **Coordinating Center (CC)** conducted simulation studies to evaluate both complex-survey procedures (when available) and model-based procedures as tools for obtaining finite-population estimates. The simulation results, which will be communicated in a separate document, show that both can yield valid finite-population inferences for HCHS/SOL data. Therefore, in this document, we present the use of both complex-survey and model-based procedures for analyzing HCHS/SOL data, highlighting approaches that have been empirically shown to provide appropriate inference for the target finite population.

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 the covariates in the model 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 (betas) of the covariates 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. Only when the response variable is continuous and the link function is the 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.

2. Visit 3 Sampling Weights

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. Calculation of Visit 3 Sampling Weights

2.1.1. Visit 1 Non-response-adjusted 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.

The Visit 1 non-response-adjusted sampling weight, **WEIGHT_NONRESP**, is released for the first time in **PART_DERV_INV5** and is used throughout this document for Visit 1-based analyses. Unlike the commonly used overall normalized sampling weight, WEIGHT_NONRESP does not incorporate trimming, calibration, or normalization and is therefore appropriate as the baseline weight for constructing additional visit-specific non-response adjustments in longitudinal analyses.

2.1.2. Visit 3 Participation Definitions

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 only phone interviews); 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 only phone interviews. In released datasets, 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.

2.1.3. Visit 3 Non-Response Adjustment and Overall Sampling Weights

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 and Regression Tree (CART)** 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 CART. The advantage of the CART 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 Myocardial Infraction, Prevalent Stroke, Born in Mainland US, and Years Lived in US at the baseline, and annual follow-up (AFU) refusal; and the following continuous variables: Height, Weight, Body Mass Index (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 CART 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 CART divided the participants into groups, referred to as CART groups, based on identified factors (used cutpoints for continuous variables). The CART 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 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 response rates, calculated for each stratum that is formed from the CART analysis described above; (2) trim extreme weights to control variability of the response rates; (3) calibrate to the age, sex 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.

In released datasets, 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. Baseline Characteristics: Visit 1 Sample vs Visit 3 Sample

The **Visit 1 Sample** includes all participants enrolled at baseline (N = 16,415) and incorporates all available data from any visit, regardless of subsequent participation status.

The **Visit 3 Sample** is restricted to participants who completed the most recent follow-up visit and may be defined under two participation definitions, as discussed in 2.1.2, depending on the variables required for the analysis. Specifically, the Visit 3 Sample may include either (1) participants who completed an in-person Visit 3 examination (including home visits), excluding those with only phone interviews (Exam Only; N = 9,090), or (2) all participants who completed any component of Visit 3, including only phone interviews (All; N = 9,864). The choice between these definitions is dictated by data availability at Visit 3: analyses involving variables collected exclusively during the in-person examination must use the Exam Only definition, whereas analyses limited to variables collected during the phone interviews may use the All definition. Throughout this document, unless otherwise noted, the term **Visit 3 Sample** refers to the Exam Only definition, as the illustrative examples use variables collected at the in-person examination.

For either definition, the analytic dataset incorporates all available data from any visit for participants meeting the selected Visit 3 participation criterion. Under the HCHS/SOL inferential framework, both the Visit 1 Sample and either definition of the Visit 3 Sample can be used to estimate parameters for the same finite target population. When missing visits are appropriately addressed using methods tailored to the chosen analytic sample and participation definition, analyses based on these alternative samples are expected to yield similar estimates of the same

finite-population parameters. 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, cardiovascular disease (CVD), and myocardial infraction, where the estimates are not stable.

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
Myocardial Infraction (%)	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.

[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
Myocardial Infraction (%)	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.

[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. Statistical Methods for Longitudinal Analysis

In this chapter, we present general statistical framework and recommendations for conducting **longitudinal analysis** with repeated measures for HCHS/SOL data involving more than two clinic visits. We begin by discussing missing visits in longitudinal analysis and, under a **Missing at Random** assumption, describe appropriate methods for addressing this form of missingness. We then describe the two dataset formats used in longitudinal analysis, wide-format and long-format, and make our recommendations on analysis methods for the study. We also provide SAS code for constructing the analytic dataset used in the illustrative examples in Chapters 4 and 5. 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.

In HCHS/SOL, the baseline cohort (N=16,415) has been followed over time. About 71% of the original cohort participated in Visit 2 (N=11,623). About 60% of the original cohort participated in Visit 3 (N=9,864), out of which 9,090 participated in the in-person exam and 774 had only phone interviews. For participants who did not participate in 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: (1) for Visit 3 with in-person participation, and (2) for ALL Visit 3 participation, including those with only phone interviews.

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. Methods for Addressing Missing Clinic Visits

Participants not returning for follow-up clinic visits is a common phenomenon in any longitudinal study, and their data 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. Unless otherwise specified, the longitudinal analyses of HCHS/SOL data in this document assume MAR for missing visits.

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.

3.1.1. Multiple Imputation

Multiple Imputation (MI) is a widely used strategy to handle missingness in both outcome 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, variables that are not part of the main models of interest, in the imputation model, potentially improving the quality of imputations.

For the FCS/MICE imputation process, based on the type of variable being imputed, the following regression methods can be used:

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

Under the MAR assumption, a key point in MI is to appropriately specify the variables related to the missing mechanism in the imputation model. 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. Because the true missingness model is unknown in practice, we recommend including all covariates from the main analytic model, along with any additional variables that may be related to the probability of missingness in MI. In the HCHS/SOL, the overall sampling weights for Visit 1, Visit 2, and Visit 3 already accounted for nonresponse due to missing the specific visit, and we recommend including these sampling weights in MI. For details on how these weights were generated: for Visit 1, please refer to *HCHS/SOL Analysis Methods at Baseline*; for Visit 2, please refer to *HCHS/SOL Analysis Methods – Visit 2*; for Visit 3, please refer to **Section 2.1** in this document.

3.1.2. Inverse Probability Weighting

Inverse Probability Weighting (IPW) is an alternative approach for handling missing visits that, like MI, provides unbiased estimation under the assumption that visit participation is MAR after conditioning on observed data. Whereas MI replaces missing data with plausible values, IPW adjusts the contribution of the observed data by up-weighting participants who resemble those who did not contribute data at a given visit.

In the HCHS/SOL, the overall sampling weights for Visit 1, Visit 2, and Visit 3 already accounted for nonresponse due to missing the specific visit. The Visit 3 participation probabilities were estimated using a CART model based on baseline and follow-up characteristics, and the inverses of these probabilities multiplied by the corresponding base sampling weights formed the non-response-adjusted weights. These overall weights therefore

serve as **design-based nonresponse-adjusted weights**, accounting simultaneously for unequal selection probabilities at baseline and for the respective differential visit participation. When the true missingness mechanism depends solely on categorical variables that are well represented in the CART-defined strata (see 2.1.3), using overall sampling weights can provide valid finite-population inference for longitudinal analysis. However, when the true missingness mechanism also depends on continuous variables, the CART-based sampling weights may not fully capture variation in visit participation, because the tree algorithm partitions the sample into broad cells and can only approximate continuous effects coarsely. In such cases, an IPW adjustment using a **Generalized Linear Model (GLM)** is a robust alternative. In this approach, **visit data contribution, defined as both participating in the visit and providing complete data on all baseline variables required for the main analytic model**, is modeled using a GLM with a logit link, where predictors are factors considered to be potentially related to the missingness. Separate GLM models are fitted for Visit 2 and Visit 3 to estimate visit-specific probabilities of data contribution, and the inverses of these estimated predicted probabilities are multiplied by Visit 1 non-response-adjusted sampling weights to form Visit 2 and Visit 3 non-response-adjusted IPW weights. Note that the Visit 1 non-response-adjusted sampling weights differ from the Visit 1 overall sampling weights, which were created with additional steps, such as weight trimming, calibration to Census benchmarks, and normalization, beyond the initial nonresponse adjustment. Refer to *HCHS/SOL Analysis Methods at Baseline* for more details.

Because the IPW approach depends on the covariates specified in the model, complete data on those covariates are required for all participants. We therefore recommend imputing any missing covariate values using MI prior to fitting the GLMs used to estimate visit-specific probabilities of data contribution. When MI is used, the visit-specific GLM is fit within each imputed dataset, and the resulting linear predictors are combined by averaging across imputations to obtain a single pooled predicted probability for each participant and visit; this pooling step replaces having multiple IPW, and the inverse of the pooled probability defines the visit-specific IPW. When the proportion of missing data is small (e.g., less than 5%), the IPW approach performs well. Because the true missingness process is typically unknown in practice, we recommend including in the MI model all variables that will be used in the GLMs for predicting visit data contribution, as well as additional variables that may be associated with covariate missingness. For the GLM models themselves, only covariates plausibly related to visit data contribution should be included. The imputed covariate values are used solely for estimating the IPW and are not used in fitting the main analytic models.

3.1.3. Additional Notes on MI and IPW Models

Both MI and IPW are valid approaches for handling missing visits under the MAR assumptions, and each has distinct advantages depending on the research objectives. The primary focus of MI is on modeling the joint relationships among study variables and fully utilizing all available information. When the imputation model includes sampling weights and design variables, MI produces design-consistent estimates while preserving the relationships among outcomes and covariates across visits, under the assumption that missing data are MAR and that the imputation

model is correctly specified**.** Although MI has been shown to yield reliable and unbiased estimates under these assumptions even when outcome missingness is high, investigators may still be concerned about imputing a large amount of data. The IPW approach relies only on the observed data and requires imputation of covariates (if needed) but not outcomes, thereby avoiding concerns about the plausibility of imputed outcome values. This approach directly models data contribution and applies inverse probability weights on top of the Visit 1 non-response-adjusted sampling weights to adjust for differential data contribution, maintaining design-consistent inference. While IPW is typically less efficient than MI, because it does not borrow information across participants or time points, it is straightforward to implement, and more robust when imputation is problematic to justify clinically.

In practice, investigators may tailor the MI and IPW model specifications based on the analytic objectives, data structure, and observed patterns of missingness. For example, the set of covariates included in the IPW models may differ across visits to more accurately capture visit-specific factors influencing data contribution. The overarching goal is to specify imputation and weighting models that adequately reflect the mechanisms driving missingness and visit-level data contribution, thereby ensuring that the resulting analytic weights align appropriately with the main analysis and support unbiased population inference.

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. Recommendations for Marginal Approaches

Assuming the missing-visit mechanism is MAR, we recommend the following marginal (population-averaged) approaches for the analysis of longitudinal HCHS/SOL data. These approaches are organized by outcome type, analysis procedure, method for addressing missing visits appropriate for the specified analytic sample, and software procedure in **Table 3.3-1**. These recommended approaches are grounded in the finite-population inferential framework adopted for HCHS/SOL and are informed by extensive simulation studies conducted by the Coordinating Center, the full results of which will be presented in a separate report. In the subsequent chapters corresponding to different outcome types and analytic settings, we illustrate the recommended procedures through examples, providing software-specific sample code and associated results to demonstrate their application in practice.

Table 3.3-1: Recommendations for Marginal Approaches

Outcome Type	Analysis Procedure	Analytic Sample	Missing-Visit Strategy	Software	Section
Continuous	Complex-survey GEE	Visit 1 Sample	MI + Visit 1 Overall Sampling Weights	SUDAAN	4.2.1.1
			Visit-specific IPW	SUDAAN	4.2.2.1
		Visit 3 Sample	Visit 3 IPW	SUDAAN	4.2.3.1
	Model-based GEE	Visit 1 Sample	MI + Visit 1 Overall Sampling Weights	SAS	4.3.1.1
				Stata	4.3.1.2
				R	4.3.1.3
			Visit-specific IPW	SAS	4.3.2.1
				R	4.3.2.2
		Visit 3 Sample	Visit 3 IPW	SAS	4.3.3.1
				Stata	4.3.3.2
				R	4.3.3.3
Binary	Complex-survey GEE	Visit 1 Sample	Visit-specific IPW	SUDAAN	5.2.1.1
		Visit 3 Sample	Visit 3 IPW	SUDAAN	5.2.2.1
	Model-based GEE	Visit 1 Sample	Visit-specific IPW	SAS	5.3.1.1
		Visit 3 Sample	Visit 3 IPW	SAS	5.3.2.1
				Stata	5.3.2.2

SUDAAN procedures can be run using SAS-callable SUDAAN, but a SUDAAN license is needed.

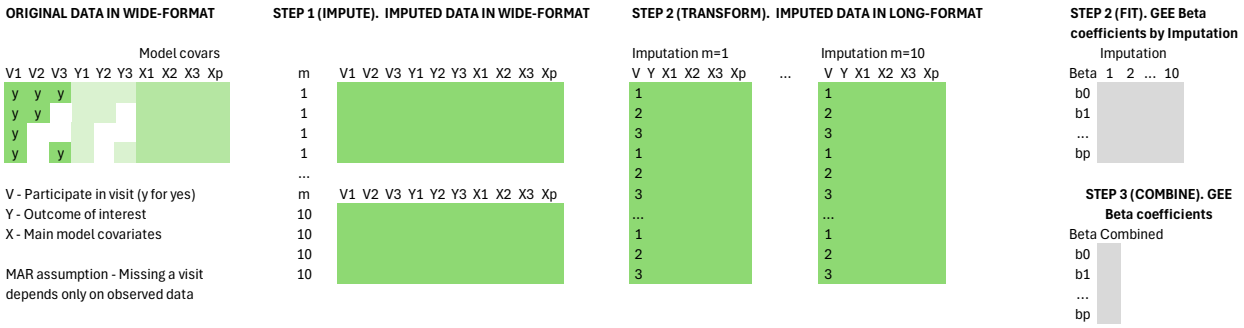
Note: There are no recommendations (1) for using visit-specific non-response-adjusted weights in Stata because Stata's panel/longitudinal procedures require the weight variable to be constant within an individual panel and do not support weights that vary; and (2) for using R for binary outcomes with IPW because R implementations of weighted logistic GEE experience numerical instability that lead to nonconvergence or implausibly large coefficient estimates.

3.3.1. General Procedure for GEE with MI

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

Step 2 (Transform then Fit): Transform each imputed dataset from wide to long-format; apply GEE to each transformed dataset for the longitudinal outcome.

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



We recommend using *m = 10* imputations. The imputation model should include all variables from the main analytic model, along with any additional variables that may be related to the probability of missing a clinic visit, even if they are not part of the main model. Design variables should also be included, as they capture key aspects of the sampling design and may be associated with visit participation. Performing imputation in the wide format helps preserve relationships among variables across visits, providing a more comprehensive representation of the longitudinal structure and maintaining the temporal dependencies and correlations between measurements at different time points.

3.3.2. General Procedure for GEE with GLM-based IPW

Step 1 (Impute): Generate *m* imputed datasets from the wide-format analytic dataset using FCS/MICE; impute all variables to be used in the IPW model if they have any missingness, noting that these imputed covariates will be used solely for estimating the IPW and not for fitting the main analytic models.

Step 2 (Estimate IPW for V2 and V3): Fit a GLM within each imputed dataset to estimate the probability of visit data contribution (i.e., both attending the visit and providing complete data on all variables required for the main analytic model) separately at each visit (e.g., Visit 2 or Visit 3). After fitting the GLM in each imputed dataset, average the linear predictors (logits) across all *m* imputations for each participant to obtain a single pooled mean linear predictor. Transform this pooled linear predictor to the probability scale using the logistic function:

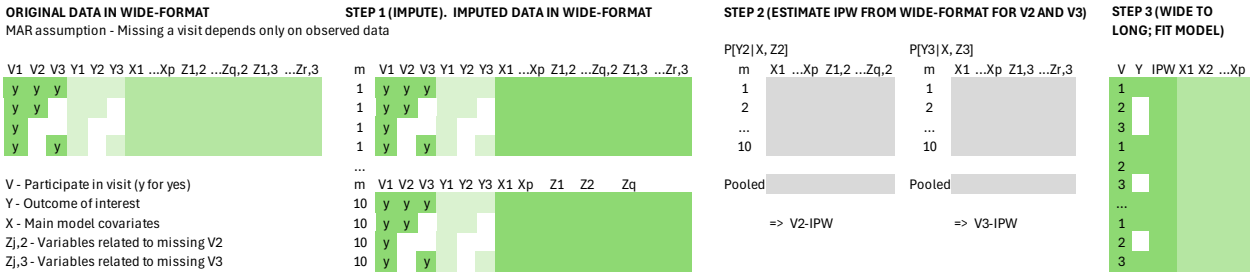
$$\text{Predicted probability of visit-specific data contribution} = \frac{1}{1 + \exp\left(-\text{pooled logit}\right)}.$$

This predicted probability represents each participant's estimated likelihood of contributing usable data at a specific visit. For each participant, the inverse of this probability is then multiplied by their **Visit 1 non-response-adjusted sampling weight** (see 2.1.1) to create their non-response-adjusted IPW weight:

$$\text{Visit-specific non-response-adjusted IPW weight} = \frac{\text{Visit 1 non-response-adjusted sampling weight}}{\text{Predicted probability of visit-specific data contribution}}.$$

The resulting non-response-adjusted IPW weights are then merged back into the wide-format analytic dataset for subsequent steps.

Step 3 (Transform then Fit): Transform the dataset from wide to long-format and apply GEE to the updated dataset for the longitudinal outcome using the non-response-adjusted IPW weights.



We recommend using *m = 10* imputations and applying appropriate imputation methods for continuous, binary, and categorical variables. The purpose of the imputation in this approach differs from that described in **Section 3.3.1**. Here, imputation is performed only for covariates used as predictors in the IPW model, rather than for all variables in the main analytic model (including outcomes). This imputation ensures complete data for estimating visit data contribution probabilities and allows all participants to contribute to the estimation of data-contribution probabilities.

3.4. 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 the illustrative examples, and it 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. Because the illustrative examples are based on measures from the in-person exam, WEIGHT_NORM_OVERALL_EXAMONLY_V3, the Visit 3 sampling weights for Exam Only participants, is imported. In addition, the Visit 1 non-response-adjusted sampling weights, WEIGHT_NONRESP, is imported for the IPW approach.

The modified 7-level reclassification of Hispanic/Latino background, BKGRD1_C7_NOMISS, is created to incorporate missing data into the "Mixed/Others" category. The binary indicator for hypertension using ACC/AHA definition at Visit 1 is renamed to add _V1 suffix for clarity. The binary indicator for Visit 2 participants, PARTICIPANT_V2, is created. The binary indicator for Visit 3 participants with in-person exam component, PARTICIPANT_EXAMONLY_V3, is merged from the Visit 3 dataset, with non–Visit 3 participants coded 0 (missing recoded to 0).

A list of variables in the analytic dataset is presented in **Output 3.4-1**. In **Output 3.4-2**, we summarize the extent of missingness for variables in the analytic dataset based on the Visit 1 Sample. In **Output 3.4-3**, we summarize the extent of missingness for variables in the analytic dataset based on the Visit 3 Sample.

```
%let v1_inv = inv5;  /* version of V1 INV file */
%let v2_inv = inv3;  /* version of V2 INV file */
%let v3_inv = inv2;  /* version of V3 INV file */

/* Visit 1 */
data analys_v1 (rename = (BMI = BMI_V1 SBPA5 = SBP5_V1));
      merge part_derv_&v1_inv. sbpa_&v1_inv.;
      by ID;
      keep PSU_ID HH_ID ID STRAT WEIGHT_FINAL_NORM_OVERALL WEIGHT_NONRESP
            CENTERNUM SEX BKGRD1_C7 AGEGROUP_C6
            US_BORN EMPLOYED EDUCATION_C3 BMI SBPA5 HYPERTENSION2_AHA;
run;

/* Visit 2 */
data analys_v2 (rename = (SBP5 = SBP5_V2));
      merge part_derv_v2_&v2_inv. sbp_v2_&v2_inv.;
      by ID;
      keep ID WEIGHT_NORM_OVERALL_V2
      BMI_V2 YRS_BTWN_V1V2 SBP5 HYPERTENSION2_AHA_V2;
run;

/* Visit 3 */
data analys_v3 (rename = (SBP5 = SBP5_V3));
      merge part_derv_v3_&v3_inv. sbp_v3_&v3_inv.;
      by ID;
      keep ID WEIGHT_NORM_OVERALL_EXAMONLY_V3 PARTICIPANT_EXAMONLY_V3
      BMI_V3 YRS_BTWN_V1V3 SBP5 HYPERTENSION2_AHA_V3;
run;

/* Analytic Dataset (wide-format) */
data sol_wide;
      merge analys_v1 analys_v2 analys_v3;
      by ID;

      /* rename hypertension at V1 to add _V1 suffix */
      HYPERTENSION2_AHA_V1 = HYPERTENSION2_AHA;
      drop HYPERTENSION2_AHA;

      /* recode background */
      BKGRD1_C7NOMISS = BKGRD1_C7;
      if BKGRD1_C7NOMISS < .Z then BKGRD1_C7NOMISS = 6;
      drop BKGRD1_C7;

      /* V2 participant indicator */
      if WEIGHT_NORM_OVERALL_V2 < .Z then PARTICIPANT_V2 = 0;
      else PARTICIPANT_V2 = 1;

      /* Set missing to 0 for non-V3 participants */
      if missing(PARTICIPANT_EXAMONLY_V3) then PARTICIPANT_EXAMONLY_V3 = 0;
run;
```

Output 3.4-1: Variables in the Analytic Dataset

Variable	Description
Design	
PSU_ID	Primary Sampling Unit (Block Group) ID
STRAT	Stratification Variable ID
HH_ID	Secondary Sampling Unit (Household) ID
ID	Participant ID
Baseline/Visit 1	
AGEGROUP_C6	Age Groups, Visit 1: 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, Visit 1: 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, Visit 1: 1=Bronx, 2=Chicago, 3=Miami, 4=San Diego
SEX	Sex, Visit 1: 0=Female, 1=Male
WEIGHT_NONRESP	Non-response-adjusted Sampling Weights, Visit 1
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
HYPERTENSION2_AHA_V1	Hypertension using ACC/AHA definition, Visit 1: 0=No, 1=Yes
US_BORN	Born in mainland US, Visit 1: 0=Not born in 50 US States/DC, 1=Born in 50 US States/DC Only
EMPLOYED	Employment Status, Visit 1: 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, Visit 1: 1=Less Than High School, 2=High School or Equivalent, 3=Greater than High School or Equivalent
Visit 2	
PARTICIPANT_V2	Visit 2 participants Indicator
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
HYPERTENSION2_AHA_V2	Hypertension using ACC/AHA definition, Visit 2: 0=No, 1=Yes
Visit 3	
PARTICIPANT_EXAMONLY_V3	Visit 3 participants with in-person exam Indicator
WEIGHT_NORM_OVERALL_EXAMONLY_V3	Overall Sampling Weights, excluding those with only phone interviews, 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
HYPERTENSION2_AHA_V3	Hypertension using ACC/AHA definition, Visit 3: 0=No, 1=Yes

Output 3.4-2: Visit 1 Sample (N=16,415), Extent of Missingness

Variable	# Observed	# Missing
Baseline/Visit 1		
AGEGROUP_C6	16414	1
BKGRD1_C7NOMISS	16415	0
CENTERNUM	16415	0
SEX	16414	1
WEIGHT_NONRESP	16415	0
WEIGHT_FINAL_NORM_OVERALL	16415	0
SBP5_V1	16400	15
BMI_V1	16343	72
HYPERTENSION2_AHA_V1	16412	3
US_BORN	16341	74
EMPLOYED	16108	307
EDUCATION_C3	16323	92
Visit 2		
WEIGHT_NORM_OVERALL_V2	11623	4792
YRS_BTWN_V1V2	11623	4792
SBP5_V2	11591	4824
BMI_V2	11245	5170
HYPERTENSION2_AHA_V2	11620	4795
Visit 3		
WEIGHT_NORM_OVERALL_EXAMONLY_V3	9090	7325
YRS_BTWN_V1V3	9864	6551
SBP5_V3	9046	7369
BMI_V3	8758	7657
HYPERTENSION2_AHA_V3	9087	7328

Output 3.4-3: Visit 3 Sample (N=9,090), Extent of Missingness

Variable	# Observed	# Missing
Baseline/Visit 1		
AGEGROUP_C6	9089	1
BKGRD1_C7NOMISS	9090	0
CENTERNUM	9090	0
SEX	9089	1
WEIGHT_NONRESP	9090	0
WEIGHT_FINAL_NORM_OVERALL	9090	0
SBP5_V1	9084	6
BMI_V1	9063	27
HYPERTENSION2_AHA_V1	9089	1
US_BORN	9070	20
EMPLOYED	8978	112
EDUCATION_C3	9057	33
Visit 2		
WEIGHT_NORM_OVERALL_V2	8432	658
YRS_BTWN_V1V2	8432	658
SBP5_V2	8416	674
BMI_V2	8246	844
HYPERTENSION2_AHA_V2	8432	658
Visit 3		
WEIGHT_NORM_OVERALL_EXAMONLY_V3	9090	0
YRS_BTWN_V1V3	9090	0
SBP5_V3	9046	44
BMI_V3	8758	332
HYPERTENSION2_AHA_V3	9087	3

4. Examples for Longitudinal Analysis of Continuous Outcomes

In this chapter, we illustrate the recommended MI and IPW methods for conducting longitudinal analysis of HCHS/SOL data for continuous outcomes with repeated measures involving more than two clinic visits and accounting for HCHS/SOL complex survey design. We assume MAR for the missing-visit mechanism. To illustrate the proposed methods, the adjusted association between time-varying covariate BMI and outcome, systolic blood pressure, is used as an example, with sample code provided in SUDAAN, SAS, Stata, and R.

4.1. Illustrative Example

4.1.1. Model Specification and Covariates

As an example for illustration, we define the main model of interest as a longitudinal analysis examining the effect of time-varying BMI on systolic blood pressure over time (long-format: SBP5; wide-format: SBP5_V1, SBP5_V2, SBP5_V3) in the HCHS/SOL target population. The primary predictor of interest is BMI over time across the three clinic visits (long-format: BMI; wide-format: BMI_V1, BMI_V2, BMI_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 (SEX), 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:

$$g(E[Y_{it}|\text{covariates}]) = \beta_0 + \beta_1\text{AGEGROUP\_C6}_i + \beta_2\text{BKGRD1\_C7NOMISS}_i + \beta_3\text{CENTERNUM}_i + \beta_4\text{SEX}_i + \beta_5\text{US\_BORN}_i + \beta_6\text{EMPLOYED}_i + \beta_7\text{EDUCATION\_C3}_i + \beta_8\text{BMI}_{it} + \beta_9\text{TIME}_{it},$$

where $g(\bullet)$ is the link function appropriate for the distribution of Y_{it} for participant i at time t (for covariates only at baseline, t is omitted). We assume **identity link** for the illustrations of continuous outcomes in this chapter.

There are two interpretations for the coefficient $\boldsymbol{\beta_8}$ for the time-varying BMI variable:

(1) WITHIN-PERSON: β_8 represents the expected change in Y (systolic blood pressure) at a given time t with a one-unit (1 kg/m²) increase in an individual's BMI. We adopt this interpretation in this chapter because the focus is longitudinal. Since the comparison is within the same individual over time, all other covariates remain unchanged; therefore, it is not necessary to state "holding all other covariates constant."

(2) BETWEEN-PERSON: β_8 represents the difference in the expected value of Y (systolic blood pressure) at a given time t when comparing individuals whose baseline covariates are identical but whose BMI values differ by one unit.

The coefficient $\boldsymbol{\beta_9}$ for the TIME variable represents the expected within-person change in Y (systolic blood pressure) with one year of aging assuming BMI remains unchanged. Note that all other covariates are exactly the same because the individual is being compared to themselves over time; therefore, it is not necessary to state "holding all other covariates constant."

4.1.2. Implementation of MI

Following the procedure in **Section 3.3.1**, use FCS/MICE to generate *10* imputed datasets from the wide-format analytic dataset sol_wide (N=16,415). Impute each variable (including 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, SEX
- Ordered logistic regression (proportional odds): EMPLOYED, AGEGROUP_C6
- Multinomial (polytomous) logistic regression: EDUCATION_C3

Covariates without missing values but included in the main model (e.g., BKGRD1_C7NOMISS) are also specified in the MI process to preserve their associations with other variables. In addition, the imputation model includes the following design variables: center (CENTERNUM), the Visit 1 overall sampling weights (WEIGHT_FINAL_NORM_OVERALL), Visit 2 overall sampling weights (WEIGHT_NORM_OVERALL_V2), and Visit 3 overall sampling weights for clinic or home exams only (WEIGHT_NORM_OVERALL_EXAMONLY_V3).

4.1.3. Implementation of IPW

Following the procedure in **Section 3.3.2**, use FCS/MICE to generate *10* imputed datasets from the wide-format analytic dataset sol_wide (N=16,415). This will impute all variables with missing values and these will be used as predictors in the IPW models. The following FCS regression models are used:

- Linear regression: BMI_V1
- Binary logistic regression: US_BORN, SEX
- Ordered logistic regression (proportional odds): EMPLOYED, AGEGROUP_C6
- Multinomial (polytomous) logistic regression: EDUCATION_C3

Covariates without missing values but included in the main model (e.g., BKGRD1_C7NOMISS) are also specified in the MI process to preserve their associations with other variables. In addition, the imputation model includes the following design variables: center (CENTERNUM), and the Visit 1 overall sampling weights (WEIGHT_FINAL_NORM_OVERALL). The imputed

covariate values are used solely for estimating the IPW and are not used in fitting the main analytic models.

Within each imputed dataset, fit a logistic regression model to estimate the probability of Visit 2 data contribution (PARTICIPANT_V2_NOMISS) among all participants. Data contribution at Visit 2 is defined as attending the visit (PARTICIPANT_V2 = 1) and having complete data on all baseline covariates used in the main model (AGEGROUP_C6, BKGRD1_C7NOMISS, CENTERNUM, SEX, US_BORN, EMPLOYED, EDUCATION_C3, BMI_V1), as well as the Visit 2 analytic variables SBP5_V2, BMI_V2, and YRS_BTWN_V1V2. Consistent with the imputation model, the same set of baseline covariates are included as predictors. For each participant, average the fitted linear predictors across imputations and transform this pooled value to the probability scale to obtain predicted probability of Visit 2 data contribution (RR_V2).

Within each imputed dataset, fit a logistic regression model to estimate the probability of Visit 3 Exam-Only data contribution (PARTICIPANT_EXAMONLY_V3_NOMISS), using the same set of baseline covariates as in the Visit 2 model, plus Visit 2 participation status (PARTICIPANT_V2) to reflect the sequential nature of visit processes. Data contribution at Visit 3 (Exam-Only) is defined as completing the in-person exam (PARTICIPANT_EXAMONLY_V3 = 1) and having complete data on all baseline covariates used in the main model (AGEGROUP_C6, BKGRD1_C7NOMISS, CENTERNUM, SEX, US_BORN, EMPLOYED, EDUCATION_C3, BMI_V1), along with the Visit 3 analytic variables SBP5_V3, BMI_V3, and YRS_BTWN_V1V3. Average the fitted linear predictors across imputations and transform them to the probability scale to obtain the predicted probability of Visit 3 data contribution (RR_V3).

As noted in **Output 3.4-2** and **Output 3.4-3**, all variables imputed in this illustrative example have less than 2% missingness.

Next, combine the pooled Visit 2 and Visit 3 response probabilities with the Visit 1 non-response-adjusted sampling weights (WEIGHT_NONRESP) to construct the Visit 2 and Visit 3 non-response-adjusted IPW weights.

For Visit 2, assign the non-response-adjusted IPW weight to participants who contributed data at Visit 2 (PARTICIPANT_V2_NOMISS = 1) as

$$\text{WEIGHT\_IPW\_V2} = \frac{\text{WEIGHT\_NONRESP}}{\text{RR\_V2}},$$

where WEIGHT_NONRESP is the Visit 1 non-response-adjusted sampling weight, and RR_V2 is the estimated probability of Visit 2 data contribution obtained from the model.

For Visit 3 (Exam-Only), assign the non-response-adjusted IPW weight to participants who contributed data for the in-person exam (PARTICIPANT_EXAMONLY_V3_NOMISS = 1) as

$$\text{WEIGHT\_EXAMONLY\_IPW\_V3} = \frac{\text{WEIGHT\_NONRESP}}{\text{RR\_V3}},$$

where RR_V3 is the estimated probability of Visit 3 Exam-Only data contribution obtained from the model.

4.2. Complex-Survey GEE

4.2.1. Visit 1 Sample, MI + Visit 1 Overall Sampling Weights

4.2.1.1. SUDAAN

MI Specification

```
proc mi data=sol_wide nimpute=10 seed=2024 out=sol_mi_wide;
      class AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX
            US_BORN EMPLOYED EDUCATION_C3;
      var AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX
            US_BORN EMPLOYED EDUCATION_C3
            WEIGHT_FINAL_NORM_OVERALL 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 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 SEX EMPLOYED AGEGROUP_C6 /* link=logit*/);
      fcs logistic(EDUCATION_C3 / link=glogit);
run;
```

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_ 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 and ordered logistic regression for ordinal variables; logistic specifying link=glogit, multinomial logistic regression for nominal variables.

Note: For details on FCS logistic regression methods in SAS, please refer to

https://documentation.sas.com/doc/en/pgmsascdc/v_069/statug/statug_mi_details13.htm

Wide-to-Long Data Transformation

```
data sol_mi_long;
  set sol_mi_wide;

  /* Visit 1 */
  VISIT = 1;
  SBP5  = SBP5_V1;
  BMI   = BMI_V1;
  TIME  = 0;
  output;

  /* Visit 2 */
  VISIT = 2;
  SBP5  = SBP5_V2;
  BMI   = BMI_V2;
  TIME  = YRS_BTWN_V1V2;
  output;

  /* Visit 3 */
  VISIT = 3;
  SBP5  = SBP5_V3;
  BMI   = BMI_V3;
  TIME  = YRS_BTWN_V1V3;
  output;

run;
```

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

Design-Based GEE with MI

```
* Convert cluster ID to a numerical variable;
data db_sudaan_mi;
      set sol_mi_long;
      hh_id_num=input(substr(hh_id, 2),8.);
run;

* Call GEE_MI_SUDAAN macro;
%GEE_MI_SUDAAN(data=db_sudaan_mi,
  strata=strat,
  psu=hh_id_num,
  weight=weight_final_norm_overall,
  response=sbp5,
  covars=bmi agegroup_c6 bkgrd1_c7nomiss centernum sex us_born employed
education_c3 time,
  class= agegroup_c6 bkgrd1_c7nomiss centernum sex us_born employed
education_c3,
  class_ref= agegroup_c6=6 bkgrd1_c7nomiss=3 centernum=4 sex=0 us_born=0
employed=1 education_c3=1);
```

Design-based estimates are obtained using SUDAAN but it does not have a native BY statement. Therefore, producing combined estimates from multiple imputed datasets involves looping over each dataset and consequently, using the MIANALYZE procedure to generate the final pooled values. The **GEE_MI_SUDAAN** macro was developed to allow analysts to perform both model fitting and pooling of estimates using a unified workflow run on the long-format sol_mi_long dataset.

This macro produces estimates for continuous responses and requires an input dataset (data) containing the _imputation_ index as a column, along with design variables specifying stratification (strata), primary sampling units (psu), and sampling weights (weight). The user defines the continuous dependent variable (response), independent covariates (covars), and categorical predictors (class) with their respective reference categories (class_ref). The optional argument nimpute (default = 10) indicates the number of imputed datasets. For each imputed dataset, the macro fits the model, extracts estimates, and produces pooled coefficients and standard errors that account for both within- and between-imputation variability. Note that before executing the macro, the PSU information must be transformed into a numerical variable to prevent issues with SUDAAN execution.

Within the macro, the PROC REGRESS procedure is employed to fit a GEE that accounts for the complex survey design. The option filetype=SAS enables SUDAAN to read and write standard SAS datasets. The correlation structure is specified as r=independent. Robust variance estimation is conducted using semethod=zeger, applying the Zeger “sandwich” estimator to ensure consistent standard errors even if the working correlation model is incorrect. The NOTSORTED option eliminates the need for prior sorting by design variables. Finally, the statement

output beta sebeta / filename=<filename> filetype=sas replace;

saves the estimated regression coefficients and standard errors into a SAS dataset for subsequent multiple-imputation pooling and analysis, and replaces it if it already exists.

Because the SUDAAN output dataset does not retain user-friendly parameter labels, we developed a companion macro, **var_levels**, which reconstructs the mapping from each pooled parameter (V_1, V_2, …) to its corresponding covariate and (for categorical predictors) level, so the final results can be displayed with clear variable names and category labels.

```
%macro var_levels(data=, covars=, class=);
       data data_out;
       length Variable $32 ClassVal 8;
       Variable="Intercept";
       ClassVal=.;
       output;
       run;

       %let i=1;
       %let cont=1;

       %do %while(%length(%scan(&covars.,&i.,%str( )))>0);
       %let var=%scan(&covars.,&i.,%str( ));

       %if %sysfunc(findw(&class.,&var.))>0 %then %do;
       proc sql noprint;
       create table _temp as
       select distinct "&var." as Variable length=32,
              &var. as ClassVal
       from &data.
       where not missing(&var.);
       quit;

       proc append base=data_out data=_temp force; run;
       %let cont=%eval(&cont.+&sqlobs);
       %end;

       %else %do;
       data _temp;
       length Variable $32 ClassVal 8;
       Variable="&var.";
       ClassVal=.;
       run;

       proc append base=data_out data=_temp force; run;
       %let cont=%eval(&cont.+1);
       %end;

       proc datasets library=work nolist;
       delete _temp;
       quit;

       data data_out;
       set data_out;
       Parm=cats('V_',_N_);
       Variable=upcase(Variable);
       run;

       %let i=%eval(&i.+1);
       %end;

%mend var_levels;
```

```
%macro GEE_MI_SUDAAN(data, strata, psu, weight, response, covars, class, class_ref,
nimpute=10);
  * Loop over each imputed dataset;
  %do m = 1 %to &nimpute.;
    data db;
      set &data.(where=(_imputation_ = &m.));
    run;
    * Fit the GEE using SUDAAN REGRESS procedure;
    proc regress data=db filetype=sas r=independent semethod=zeger notsorted;
      nest &strata. &psu.;
      weight &weight.;
      class &class.;
      model &response. = &covars.;
      reflevel &class_ref.;
      output beta sebeta / filename=est_mi_&m. filetype=sas replace;
    run;
    data betas_mi_&m.(rename=(beta=Estimate sebeta=StdErr));
      set est_mi_&m.;
      _imputation_ = &m.;
      Parm = cats('V_', modelrhs);
      format beta sebeta 12.4;
    run;
  %end;
  * Combine and arrange dataset to use in MIANALYZE;
  data outparms;
    set betas_mi_:;
  run;
  proc sql noprint;
    select max(modelrhs) into :maxrhs
    from outparms;
  quit;
  %let vlist=;
  %do i = 1 %to &maxrhs.;
    %let vlist = &vlist. V_&i.;
  %end;
  data outparms;
    set outparms;
    drop modelrhs procnum modelno;
  run;
  proc sort data=outparms;
    by _imputation_;
  run;
  * Obtain pooled estimates;
  proc mianalyze parms=outparms;
    modeleffects &vlist.;
    ods output ParameterEstimates =
      betas_mi(keep=Parm Estimate StdErr tValue Probt);
  run;
  * Use var_levels macro to produce variable names and labels;
  %var_levels(data=&data., covars=&covars., class=&class.);
  * Sort and merge estimates with labels;
  proc sort data=betas_mi; by Parm; run;
  proc sort data=data_out; by Parm; run;
  data betas_out; merge data_out betas_mi; by Parm; drop Parm; run;
  proc sort data=betas_out; by Variable ClassVal; run;
  * Print estimates;
  proc print data=betas_out noobs;
    title 'Pooled Beta Estimates using SUDAAN';
  run;
%mend GEE_MI_SUDAAN;
```

Estimates

The **Output 4.2-1** displays the pooled estimates across 10 imputed datasets.

The estimated coefficient for BMI is 0.197 with a standard error of 0.026. This positive association indicates that, a 1 kg/m² increase in an individual's BMI is associated with on average a 0.197 mmHg increase in their systolic blood pressure. This effect is statistically significant ($p < .0001$).

The estimated coefficient for TIME is 0.398 with a standard error of 0.023. TIME is the number of years since the baseline visit. It is equivalent to aging. This result indicates that each additional year of aging of an individual is associated with on average a 0.398 mmHg increase in their systolic blood pressure given that their BMI did not change during the additional year.

Output 4.2-1

Variable	Class Val	Estimate	StdErr	tValue	Probt
INTERCEPT	.	128.113806	1.162781	110.18	<.0001
AGEGROUP_C6	1	-27.237640	0.891231	-30.56	<.0001
AGEGROUP_C6	2	-24.900034	0.878139	-28.36	<.0001
AGEGROUP_C6	3	-19.653613	0.904229	-21.74	<.0001
AGEGROUP_C6	4	-13.298408	0.906271	-14.67	<.0001
AGEGROUP_C6	5	-7.093930	0.834566	-8.50	<.0001
AGEGROUP_C6	6	0	0	.	.
BKGRD1_C7NOMISS	0	2.309152	0.779266	2.96	0.0035
BKGRD1_C7NOMISS	1	1.559880	0.633601	2.46	0.0144
BKGRD1_C7NOMISS	2	0.407996	0.742664	0.55	0.5834
BKGRD1_C7NOMISS	3	0	0	.	.
BKGRD1_C7NOMISS	4	0.869791	0.638490	1.36	0.1741
BKGRD1_C7NOMISS	5	-1.880510	0.701913	-2.68	0.0083
BKGRD1_C7NOMISS	6	1.918489	1.060808	1.81	0.0793
BMI	.	0.196995	0.025862	7.62	<.0001
CENTERNUM	1	1.149023	0.657448	1.75	0.0820
CENTERNUM	2	-1.810191	0.478265	-3.78	0.0002
CENTERNUM	3	4.646581	0.754652	6.16	<.0001
CENTERNUM	4	0	0	.	.
EDUCATION_C3	1	0	0	.	.
EDUCATION_C3	2	-1.080721	0.401703	-2.69	0.0078
EDUCATION_C3	3	-2.422101	0.392110	-6.18	<.0001
EMPLOYED	1	0	0	.	.
EMPLOYED	2	-0.490358	0.920357	-0.53	0.5963
EMPLOYED	3	-0.684703	0.961289	-0.71	0.4794
EMPLOYED	4	0.229406	0.912882	0.25	0.8025
SEX	0	0	0	.	.
SEX	1	5.546536	0.318894	17.39	<.0001
TIME	.	0.397621	0.023166	17.16	<.0001
US_BORN	0	0	0	.	.
US_BORN	1	-0.026674	0.491953	-0.05	0.9568

4.2.2. Visit 1 Sample, Visit-specific IPW

4.2.2.1. SUDAAN

Construction of IPW Indicators

```
data sol_wide;
  set sol_wide;

  array basecov AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX US_BORN
                EMPLOYED EDUCATION_C3 BMI_V1;

  if PARTICIPANT_V2 = 1 and
     nmiss(of basecov[*], SBP5_V2, BMI_V2, YRS_BTWN_V1V2) = 0 then
    PARTICIPANT_V2_NOMISS = 1;
  else
    PARTICIPANT_V2_NOMISS = 0;

  if PARTICIPANT_EXAMONLY_V3 = 1 and
     nmiss(of basecov[*], SBP5_V3, BMI_V3, YRS_BTWN_V1V3) = 0 then
    PARTICIPANT_EXAMONLY_V3_NOMISS = 1;
  else
    PARTICIPANT_EXAMONLY_V3_NOMISS = 0;
run;
```

In this **data** step, basecov collects all baseline covariates used in the main model of interest. The following indicators are created to serve as the outcomes in the IPW models:

PARTICIPANT_V2_NOMISS = 1 identifies participants who both attended Visit 2 (PARTICIPANT_V2 = 1) and have complete data on all baseline covariates plus SBP5_V2, BMI_V2, and YRS_BTWN_V1V2. This defines Visit 2 data contribution.

PARTICIPANT_EXAMONLY_V3_NOMISS = 1 similarly identifies participants who both attended the Visit 3 Exam-Only component and have complete data on all baseline covariates plus SBP5_V3, BMI_V3, and YRS_BTWN_V1V3. This defines Visit 3 (Exam-Only) data contribution.

MI for IPW Estimation

```
proc mi data=sol_wide nimpute=10 seed=2024 out=sol_mi_for_ipw;
  class AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX US_BORN
           EMPLOYED EDUCATION_C3;
  var   AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX
        US_BORN EMPLOYED EDUCATION_C3 BMI_V1
           WEIGHT_FINAL_NORM_OVERALL;
  fcs reg(BMI_V1);
  fcs logistic(US_BORN SEX EMPLOYED AGEGROUP_C6/* link=logit */);
  fcs logistic(EDUCATION_C3 / link=glogit);
run;
```

The procedure **proc mi** performs MI with nimpute=**10** generates 10 imputed datasets. 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_for_ipw, a single dataset that contains all the imputed data stacked, containing an imputation number identifier _IMPUTATION_ 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 and ordered logistic regression for ordinal variables; logistic specifying link=glogit, multinomial logistic regression for nominal variables. These imputed covariate values are used solely to estimate the IPW models.

Note: For details on FCS logistic regression methods in SAS, please refer to:

https://documentation.sas.com/doc/en/pgmsascdc/v_069/statug/statug_mi_details13.htm

Estimation of IPW Models by Visit

```
/* = Visit 2: logistic on PARTICIPANT_V2_NOMISS with baseline covariates =
*/
proc logistic data=sol_mi_for_ipw descending noprint;
  by _Imputation_;
  class AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX
        US_BORN EMPLOYED EDUCATION_C3 PARTICIPANT_V2_NOMISS;
  model PARTICIPANT_V2_NOMISS =
        AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX
        US_BORN EMPLOYED EDUCATION_C3 BMI_V1
            WEIGHT_FINAL_NORM_OVERALL;
  output out=pred_v2_imp(keep=_Imputation_ ID xb_v2) xbeta=xb_v2;
run;

proc means data=pred_v2_imp nway noprint;
  class ID;
  var xb_v2;
  output out=pred_v2_bar(drop=_type_ _freq_) mean=xb_v2_bar;
run;

/* = Visit 3: logistic on PARTICIPANT_EXAMONLY_V3_NOMISS
     with baseline + PARTICIPANT_V2 = */
proc logistic data=sol_mi_for_ipw descending noprint;
  by _Imputation_;
  class AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX
        US_BORN EMPLOYED EDUCATION_C3
        PARTICIPANT_V2
        PARTICIPANT_EXAMONLY_V3_NOMISS;
  model PARTICIPANT_EXAMONLY_V3_NOMISS =
        AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX
        US_BORN EMPLOYED EDUCATION_C3 BMI_V1
            WEIGHT_FINAL_NORM_OVERALL PARTICIPANT_V2;
  output out=pred_v3_imp(keep=_Imputation_ ID xb_v3) xbeta=xb_v3;
run;

proc means data=pred_v3_imp nway noprint;
  class ID;
  var xb_v3;
  output out=pred_v3_bar(drop=_type_ _freq_) mean=xb_v3_bar;
run;
```

To estimate Visit 2 data-contribution probabilities, a logistic regression model is fit using **proc logistic** within each imputed dataset (by _Imputation_) with outcome PARTICIPANT_V2_NOMISS and baseline predictors and the Visit 1 overall sampling weight (WEIGHT_FINAL_NORM_OVERALL). The linear predictor (xb_v2) is saved for each participant and then averaged across imputations within ID using **proc means**, yielding xb_v2_bar, the pooled logit of Visit 2 data contribution. Visit 3 Exam-Only data contribution is modeled analogously, with the outcome PARTICIPANT_EXAMONLY_V3_NOMISS and the same baseline covariates used in the Visit 2 model, additionally with Visit 2 participation (PARTICIPANT_V2) to reflect the sequential nature of study visits. The resulting linear predictors (xb_v3) are again averaged across imputations to produce xb_v3_bar, the pooled logit for Visit 3 data contribution.

Construction of IPW Weights by Visit

```
/* = Merge Response Rates (RR) back to sol_wide and compute IPWs = */
proc sort data=sol_wide;     by ID; run;
proc sort data=pred_v2_bar;  by ID; run;
proc sort data=pred_v3_bar;  by ID; run;

data sol_ipw_wide;
  merge sol_wide(in=a)
        pred_v2_bar(rename=(xb_v2_bar=lp_v2))
        pred_v3_bar(rename=(xb_v3_bar=lp_v3));
  by ID;
  if a;

  /* Predicted response rates (probability of being in analytic set) */
  if not missing(lp_v2) then RR_V2 = 1/(1+exp(-lp_v2));
  if not missing(lp_v3) then RR_V3 = 1/(1+exp(-lp_v3));

  /* non-response-adjusted IPW weights (only for analytic attendees) */
  if PARTICIPANT_V2_NOMISS = 1 then WEIGHT_IPW_V2 = WEIGHT_NONRESP / RR_V2;
  else WEIGHT_IPW_V2 = .;

  if PARTICIPANT_EXAMONLY_V3_NOMISS = 1 then WEIGHT_EXAMONLY_IPW_V3 =
WEIGHT_NONRESP / RR_V3;
  else WEIGHT_EXAMONLY_IPW_V3 = .;
run;
```

In this **data** step, lp_v2 and lp_v3 are the pooled logits for Visits 2 and 3, respectively; these values are transformed using the inverse logit to obtain the predicted probabilities of visit-level data contribution (RR_V2 and RR_V3). These probabilities represent each participant's estimated likelihood of contributing complete analytic data at the corresponding visit.

For participants who contribute data at Visit 2 (PARTICIPANT_V2_NOMISS = 1), the non-response-adjusted IPW weight is computed as the Visit 1 non-response-adjusted sampling weight (WEIGHT_NONRESP) divided by RR_V2. For Visit 3 Exam-Only contributors, the corresponding non-response-adjusted IPW weight is WEIGHT_NONRESP divided by RR_V3. Participants with PARTICIPANT_V2_NOMISS = 0 or PARTICIPANT_EXAMONLY_V3_NOMISS = 0 have their visit-specific IPW weight variables set to missing, because they do not contribute analytic data at those visits and therefore are not included in the main analysis.

Creation of IPW Long-Format Dataset

```
data sol_ipw_long;
  set sol_ipw_wide;

  length VISIT 8 TIME 8 SBP5 8 BMI 8 WEIGHT_IPW_BY_VISIT 8;

  /* Visit 1 row */
  VISIT = 1;
  SBP5  = SBP5_V1;
  BMI   = BMI_V1;
  TIME  = 0;
  WEIGHT_IPW_BY_VISIT  = WEIGHT_FINAL_NORM_OVERALL;
  output;

  /* Visit 2 row */
  VISIT = 2;
  SBP5  = SBP5_V2;
  BMI   = BMI_V2;
  TIME  = YRS_BTWN_V1V2;
  WEIGHT_IPW_BY_VISIT  = WEIGHT_IPW_V2;
  output;

  /* Visit 3 row (Exam-only definition) */
  VISIT = 3;
  SBP5  = SBP5_V3;
  BMI   = BMI_V3;
  TIME  = YRS_BTWN_V1V3;
  WEIGHT_IPW_BY_VISIT  = WEIGHT_EXAMONLY_IPW_V3;
  output;
run;
```

In this **data** step, the wide-format dataset is reshaped into a long-format structure by creating one record per participant for each clinic visit and assigning the appropriate visit identifier and time-varying variables. For sampling weights, the Visit 1 row carries the Visit 1 overall sampling weight (WEIGHT_FINAL_NORM_OVERALL), while the Visit 2 and Visit 3 rows use the newly constructed non-response-adjusted IPW weights (WEIGHT_IPW_V2 and WEIGHT_EXAMONLY_IPW_V3). This reshaping produces the final analytic dataset, sol_ipw_long, with one record per subject–visit and all visit-specific variables aligned for the main analysis.

Note that Visit 2 and Visit 3 non-response-adjusted IPW weights are set to be missing for participants who do not contribute data at those visits.

Variable	VISIT 1		VISIT 2		VISIT 3	
	# Observed	# Missing	# Observed	# Missing	# Observed	# Missing
SBP5	16400	15	11591	4824	9046	7369
BMI	16343	72	11245	5170	8758	7657
TIME	16415	0	11623	4792	9864	6551
WEIGHT_IPW_BY_VISIT	16415	0	11033	5382	8580	7835

Design-Based GEE with IPW

```
data db_ipw_byvisit;
      set sol_ipw_long ;
      hh_id_num=input(substr(hh_id, 2),8.);
run;

* Call SUDAAN procedure;
proc regress data=db_ipw_byvisit filetype=sas r=independent semethod=zeger
      notsorted;
 nest strat hh_id_num;
 weight weight_ipw_by_visit;
 class agegroup_c6 bkgrd1_c7nomiss centernum sex us_born employed
           education_c3;
 model sbp5=bmi agegroup_c6 bkgrd1_c7nomiss centernum sex us_born employed
           education_c3 time;
 reflevel agegroup_c6=6 bkgrd1_c7nomiss=3 centernum=4 sex=0 us_born=0
           employed=1 education_c3=1;
 setenv labwidth=25 decwidth=3;
 print beta="Estimate" sebeta="(S.E)" t_beta="t value" p_beta="p-value";
run;
```

The **proc regress** procedure fits the linear GEE to the long-format dataset sol_ipw_long. The analysis incorporates the complex survey design by specifying stratification and clustering through nest strat hh_id_num. The visit-specific non-response-adjusted IPW weights (weight_ipw_by_visit) are used in this case to obtain weighted population-representative estimates. Any records with missing weights are not used in analysis, ensuring that the IPW-defined analysis is consistent with the GEE main analysis. The class statement identifies categorical covariates, and the reflevel statement sets the corresponding reference groups (e.g., agegroup_c6=6). The model statement specifies SBP5 as the continuous outcome and includes BMI, demographic covariates, and follow-up time (time) as predictors. The working correlation structure is specified as independent using r=independent, and robust standard errors are computed using Zeger's sandwich variance estimator (semethod=zeger). The setenv and print statements are used to format and display parameter estimates, standard errors, test statistics, and p-values and its corresponding labels.

Estimates

Output 4.2-2 displays the estimates. The estimated association between BMI and systolic blood pressure is 0.167 with a standard error of 0.033. This positive association indicates that, a 1 kg/m² increase in an individual's BMI is associated with on average a 0.167 mmHg increase in their systolic blood pressure. This effect is statistically significant ($p < .0001$). The estimated coefficient for TIME is 0.648 with a standard error of 0.055. TIME is the number of years since the baseline visit. It is equivalent to aging. This result indicates that each additional year of aging of an individual is associated with on average a 0.648 mmHg increase in their systolic blood pressure given that their BMI did not change during the additional year. This association is statistically significant ($p < .0001$).

Output 4.2-2

Independent Variables and Effects	Estimate	(S.E)	t value	p-value
Intercept	124.782	1.701	73.338	0.000
BMI	0.167	0.033	5.009	0.000
6-level Age Sub-groups				
1	-28.183	1.398	-20.160	0.000
2	-24.834	1.359	-18.279	0.000
3	-18.824	1.331	-14.148	0.000
4	-12.121	1.295	-9.361	0.000
5	-6.563	1.265	-5.189	0.000
6	0.000	0.000	.	.
BKGRD1_C7NOMISS				
0	2.020	1.014	1.992	0.046
1	1.195	0.816	1.465	0.143
2	0.982	0.997	0.986	0.324
3	0.000	0.000	.	.
4	0.810	1.038	0.780	0.435
5	-2.169	0.871	-2.491	0.013
6	1.501	0.999	1.504	0.133
Participant's Field Center - numeric				
1	1.557	0.963	1.616	0.106
2	-0.775	0.663	-1.168	0.243
3	6.391	0.925	6.909	0.000
4	0.000	0.000	.	.
Sex				
0	0.000	0.000	.	.
1	4.648	0.497	9.354	0.000
Born in mainland US (50 States + DC)				
0	0.000	0.000	.	.
1	0.378	0.661	0.572	0.568
Employment Status (includes retirees)				
1	0.000	0.000	.	.
2	0.632	1.275	0.496	0.620
3	0.054	1.321	0.041	0.967
4	1.069	1.273	0.840	0.401
Education Status (3 levels)				
1	0.000	0.000	.	.
2	-1.148	0.610	-1.884	0.060
3	-3.458	0.600	-5.764	0.000
TIME	0.648	0.055	11.871	0.000

4.2.3. Visit 3 Sample, Visit 3 IPW

4.2.3.1. SUDAAN

The steps for constructing the dataset sol_ipw_long are the same as those described in 4.2.2.1.

Design-Based GEE with IPW

```
data db_cont_ipw;
  set sol_ipw_long ;
  hh_id_num=input(substr(hh_id, 2),8.);
run;

* Call SUDAAN procedure;
proc regress data=db_cont_ipw filetype=sas r=independent semethod=zeger
  notsorted;
  nest strat hh_id_num;
  weight weight_examonly_ipw_v3;
  subpopn participant_examonly_v3_nomiss=1;
  class agegroup_c6 bkgrd1_c7nomiss centernum sex us_born employed
education_c3;
  model sbp5=bmi agegroup_c6 bkgrd1_c7nomiss centernum sex us_born employed
    education_c3 time;
  reflevel agegroup_c6=6 bkgrd1_c7nomiss=3 centernum=4 sex=0 us_born=0
    employed=1 education_c3=1;
  setenv labwidth=25 decwidth=3;
  print beta="Estimate" sebeta="(S.E)" t_beta="t value" p_beta="p-value";
run;
```

The **proc regress** procedure in SUDAAN fits GEE to the derived dataset from sol_ipw_long after converting the psu identifier to a numeric variable. The survey design is accounted for through nest strat hh_id_num, which specifies the stratification and clustering structure, and the weight statement applies the Visit 3 non-response-adjusted IPW weights (weight_examonly_ipw_v3). The subpopn statement restricts the analysis to participants who contribute data to Visit 3 (participant_examonly_v3_nomiss=1), ensuring consistency between the IPW-defined analysis and the GEE main analysis.

The class statement identifies categorical covariates, and reference categories are explicitly set with the reflevel statement (e.g., agegroup_c6=6). The model statement specifies systolic blood pressure (sbp5) as the continuous outcome and includes BMI, demographic variables, and time as predictors. The working correlation structure is specified as independent using r=independent, and robust standard errors are calculated through semethod=zeger. The setenv statement controls formatting of printed output, and the print statement outputs regression coefficients, standard errors, test statistics, and p-values with the specified labels.

Estimates

Output 4.2-3 displays the results.

The estimated association between BMI and systolic blood pressure is 0.183 with a standard error of 0.034. This positive association indicates that, a 1 kg/m² increase in an individual's BMI is associated with on average a 0.183 mmHg increase in their systolic blood pressure. This effect is statistically significant ($p < .0001$).

The estimated coefficient for TIME is 0.464 with a standard error of 0.029. TIME is the number of years since the baseline visit. It is equivalent to aging. This result indicates that each additional year of aging of an individual is associated with on average a 0.464 mmHg increase in their systolic blood pressure given that their BMI did not change during the additional year. This association is statistically significant ($p < .0001$).

Output 4.2-3

Independent Variables and Effects	Estimate	(S.E)	t value	p-value
Intercept	126.440	1.637	77.241	0.000
BMI	0.183	0.034	5.416	0.000
6-level Age Sub-groups				
1	-26.020	1.496	-17.396	0.000
2	-24.049	1.432	-16.795	0.000
3	-18.958	1.415	-13.399	0.000
4	-12.154	1.378	-8.818	0.000
5	-6.494	1.340	-4.847	0.000
6	0.000	0.000	.	.
BKGRD1_C7NOMISS				
0	2.009	0.966	2.079	0.038
1	1.164	0.821	1.418	0.156
2	0.878	0.973	0.903	0.367
3	0.000	0.000	.	.
4	1.367	0.979	1.397	0.163
5	-1.519	0.941	-1.613	0.107
6	1.062	0.958	1.108	0.268
Participant's Field Center - numeric				
1	1.329	0.934	1.423	0.155
2	-1.035	0.634	-1.632	0.103
3	6.049	0.915	6.610	0.000
4	0.000	0.000	.	.
Sex				
0	0.000	0.000	.	.
1	5.293	0.499	10.617	0.000
Born in mainland US (50 States + DC)				
0	0.000	0.000	.	.
1	0.108	0.641	0.168	0.866
Employment Status (includes retirees)				
1	0.000	0.000	.	.
2	-0.167	1.330	-0.125	0.900
3	-1.226	1.348	-0.909	0.363
4	0.627	1.314	0.477	0.633
Education Status (3 levels)				
1	0.000	0.000	.	.
2	-1.504	0.594	-2.532	0.011
3	-3.220	0.589	-5.469	0.000
TIME	0.464	0.029	16.130	0.000

4.3. Model-Based GEE

4.3.1. Visit 1 Sample, MI + Visit 1 Overall Sampling Weights

4.3.1.1. SAS

The steps for constructing the dataset sol_mi_long are the same as those described in 4.2.1.1.

Model-Based GEE with MI

```
proc genmod data=sol_mi_long;
      by _IMPUTATION_;
      class HH_ID AGEGROUP_C6(ref = '6') BKGRD1_C7NOMISS(ref = '3')
            CENTERNUM(ref = '4') SEX(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 SEX US_BORN
            EMPLOYED EDUCATION_C3 SBP5 TIME/ dist=normal;
      repeated subject = HH_ID /corr=ind;
      ods output GEEEmpPEst=betas_mi;
run;
```

The **proc genmod** procedure fits GEE to the long-format sol_mi_long dataset. 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.

```
proc mianalyze parms(classvar=level)=betas_mi;
      class AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX US_BORN
            EMPLOYED EDUCATION_C3;
      modeleffects INTERCEPT AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM
            SEX US_BORN EMPLOYED EDUCATION_C3 SBP5 TIME;
run;
```

The **proc mianalyze** procedure combines the MI results in betas_mi using Rubin's rules. Specifying 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**.

Estimates

The **Output 4.3-1** displays the pooled estimates across the 10 imputed datasets.

The estimated association between BMI and systolic blood pressure is 0.197 with a standard error of 0.026. This positive association indicates that, a 1 kg/m² increase in an individual's BMI is associated with on average a 0.197 mmHg increase in their systolic blood pressure. This effect is statistically significant ($p < .0001$).

The estimated coefficient for TIME is 0.398 with a standard error of 0.023. TIME is the number of years since the baseline visit. It is equivalent to aging. This result indicates that each additional year of aging of an individual is associated with on average a 0.398 mmHg increase in their systolic blood pressure given that their BMI did not change during the additional year. This association is statistically significant ($p < .0001$).

Output 4.3-1

Parm	Estimate	StdErr	LCLMean	UCLMean	DF	Min	Max	Theta0	tValue	Probt
INTERCEPT	128.113806	1.165832	125.8079	130.4198	133.1	127.153716	128.916737	0	109.89	<.0001
AGEGROUP_C6	-27.237640	0.892549	-28.9951	-25.4802	262	-27.872611	-26.731039	0	-30.52	<.0001
AGEGROUP_C6	-24.900034	0.878811	-26.6301	-23.1699	272.79	-25.648433	-24.496195	0	-28.33	<.0001
AGEGROUP_C6	-19.653613	0.905195	-21.4446	-17.8626	128.87	-20.215989	-19.154434	0	-21.71	<.0001
AGEGROUP_C6	-13.298408	0.907144	-15.0974	-11.4994	103.53	-13.987948	-12.758080	0	-14.66	<.0001
AGEGROUP_C6	-7.093930	0.835301	-8.7379	-5.4500	293.98	-7.564398	-6.590541	0	-8.49	<.0001
AGEGROUP_C6	0	0	.	.	.	0	0	0	.	.
BKGRD1_C7NOMISS	2.309152	0.778789	0.7725	3.8458	180.85	1.620832	2.820831	0	2.97	0.0034
BKGRD1_C7NOMISS	1.559880	0.633549	0.3131	2.8067	298.72	1.281495	2.067030	0	2.46	0.0144
BKGRD1_C7NOMISS	0.407996	0.742513	-1.0567	1.8727	188.38	-0.025439	1.020668	0	0.55	0.5833
BKGRD1_C7NOMISS	0	0	.	.	.	0	0	0	.	.
BKGRD1_C7NOMISS	0.869791	0.638404	-0.3865	2.1261	302.32	0.513386	1.193687	0	1.36	0.1741
BKGRD1_C7NOMISS	-1.880510	0.701836	-3.2688	-0.4923	132.52	-2.411372	-1.434368	0	-2.68	0.0083
BKGRD1_C7NOMISS	1.918489	1.060630	-0.2364	4.0734	34.25	0.843473	2.840148	0	1.81	0.0793
CENTERNUM	1.149023	0.657828	-0.1480	2.4461	203.08	0.590357	1.557088	0	1.75	0.0822
CENTERNUM	-1.810191	0.478158	-2.7562	-0.8641	128.89	-2.276543	-1.451493	0	-3.79	0.0002
CENTERNUM	4.646581	0.755204	3.1481	6.1450	99.222	4.025419	5.256075	0	6.15	<.0001
CENTERNUM	0	0	.	.	.	0	0	0	.	.
SEX	0	0	.	.	.	0	0	0	.	.
SEX	5.546536	0.319102	4.9194	6.1736	451.27	5.419051	5.731815	0	17.38	<.0001
US_BORN	0	0	.	.	.	0	0	0	.	.
US_BORN	-0.026674	0.491665	-0.9984	0.9450	146.03	-0.398659	0.294760	0	-0.05	0.9568
EMPLOYED	0	0	.	.	.	0	0	0	.	.
EMPLOYED	-0.490358	0.920625	-2.3347	1.3539	55.908	-1.347340	0.432861	0	-0.53	0.5964
EMPLOYED	-0.684703	0.961015	-2.6121	1.2427	53.194	-1.713506	0.171637	0	-0.71	0.4793
EMPLOYED	0.229406	0.912742	-1.5983	2.0571	57.11	-0.403208	1.143198	0	0.25	0.8025
EDUCATION_C3	0	0	.	.	.	0	0	0	.	.
EDUCATION_C3	-1.080721	0.402164	-1.8743	-0.2872	180.65	-1.451704	-0.835194	0	-2.69	0.0079
EDUCATION_C3	-2.422101	0.392303	-3.1937	-1.6505	343.15	-2.665036	-2.224432	0	-6.17	<.0001
BMI	0.196995	0.025929	0.1459	0.2481	223.39	0.180236	0.211221	0	7.60	<.0001
TIME	0.397621	0.023302	0.3503	0.4449	35.379	0.362856	0.412475	0	17.06	<.0001

4.3.1.2. 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.

MI Specification

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

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.

```
mi register imputed AGEGROUP_C6 SEX 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 BKGRD1_C7NOMISS CENTERNUM 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 SEX (ologit) EMPLOYED AGEGROUP_C6 (mlogit)
EDUCATION_C3 = i.BKGRD1_C7NOMISS i.CENTERNUM WEIGHT_FINAL_NORM_OVERALL,
add(10)
```

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, SEX); ordered logistic regression *ologit* for ordinal variables (EMPLOYED, AGEGROUP_C6); 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.

Wide-to-Long Data Transformation

```
rename BMI_V1 BMI1
rename BMI_V2 BMI2
rename BMI_V3 BMI3

rename SBP5_V1 SBP51
rename SBP5_V2 SBP52
rename SBP5_V3 SBP53

rename YRS_BTWN_V1V2 TIME2
rename YRS_BTWN_V1V3 TIME3

generate TIME1 = 0

mi reshape long BMI SBP5 TIME, i(ID) j(VISIT)
```

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.

Model-Based GEE with MI

```
encode ID, gen(ID_NUM)
mi xtset ID_NUM

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

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 GEE.

Estimates

Output 4.3-2 displays the estimates.

The estimated association between BMI and systolic blood pressure is 0.188 with a standard error of 0.026. This positive association indicates that, a 1 kg/m² increase in an individual's BMI is associated with on average a 0.188 mmHg increase in their systolic blood pressure. This effect is statistically significant ($p < .0001$).

The estimated coefficient for TIME is 0.397 with a standard error of 0.022. TIME is the number of years since the baseline visit. It is equivalent to aging. This result indicates that each additional year of aging of an individual is associated with on average a 0.397 mmHg increase in their systolic blood pressure given that their BMI did not change during the additional year. This association is statistically significant ($p < .0001$).

Output 4.3-2

SBP5	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
AGEGROUP_C6						
1	-27.474	.889441	-30.89	0.000	-29.22178	-25.72622
2	-25.06745	.9424164	-26.60	0.000	-26.93029	-23.20461
3	-19.95036	.8644649	-23.08	0.000	-21.64896	-18.25176
4	-13.53793	.8510557	-15.91	0.000	-15.21082	-11.86505
5	-7.299951	.8087194	-9.03	0.000	-8.885696	-5.714206
BKGRD1_C7NOMISS						
0	.8197336	.9564349	0.86	0.393	-1.073417	2.712884
1	.0604927	.8738588	0.07	0.945	-1.661497	1.782483
2	-1.066906	.9602611	-1.11	0.269	-2.972663	.8388515
3	-1.524637	.9201255	-1.66	0.102	-3.360283	.3110094
4	-.5934599	.8560731	-0.69	0.489	-2.28469	1.09777
5	-3.232706	.9143502	-3.54	0.001	-5.037652	-1.427759
CENTERNUM						
1	1.142296	.661812	1.73	0.086	-.1646633	2.449256
2	-1.824794	.5165084	-3.53	0.001	-2.853939	-.7956487
3	4.552927	.7817503	5.82	0.000	2.995234	6.110619
1.SEX	5.525015	.3207325	17.23	0.000	4.89371	6.156321
1.US_BORN	-.1406076	.4784537	-0.29	0.769	-1.08543	.8042152
EMPLOYED						
2	-.2627698	.8563836	-0.31	0.759	-1.957136	1.431596
3	-.3384737	.8670551	-0.39	0.697	-2.049808	1.372861
4	.3959017	.8388349	0.47	0.638	-1.262056	2.053859
EDUCATION_C3						
2	-1.12912	.4248632	-2.66	0.009	-1.97089	-.2873505
3	-2.499465	.4079538	-6.13	0.000	-3.302662	-1.696268
BMI	.1880254	.0263713	7.13	0.000	.1356768	.240374
TIME	.3967365	.0218214	18.18	0.000	.3526628	.4408102
_cons	129.9944	1.527096	85.13	0.000	126.9509	133.0379

4.3.1.3. R

Data Setup

```
## 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$SEX <- relevel(factor(sol$SEX), 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')
```

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.

MI Specification

```
# Set up MI using MICE
predMatrix <- quickpred(sol, include = c("AGEGROUP_C6", "BKGRD1_C7NOMISS",
"CENTERNUM", "SEX", "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","SEX)] <- "logreg"
methods[c("EMPLOYED","AGEGROUP_C6")] <- "polr"
methods[c("EDUCATION_C3")] <- "polyreg"

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

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, SEX); ordered logistic regression *polr* for ordinal variables (EMPLOYED, AGEGROUP_C6); 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'.

Model-Based GEE with MI

```
# 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(
    SBP5 ~ AGEGROUP_C6 + BKGRD1_C7NOMISS + CENTERNUM + SEX + US_BORN +
EMPLOYED + EDUCATION_C3 + BMI + TIME,
    data = imputed_data_long_i,
    id = HH_ID,
    corstr = "independence",
    weight = WEIGHT_FINAL_NORM_OVERALL,
    family = gaussian(link = "identity")
  )
}
```

The *complete* function combines all items in the list object 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'.

```
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
```

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.

Estimates

Output 4.3-3 displays the estimates.

The estimated association between BMI and systolic blood pressure is 0.192 with a standard error of 0.025. This positive association indicates that, a 1 kg/m² increase in an individual's BMI is associated with on average a 0.192 mmHg increase in their systolic blood pressure. This effect is statistically significant ($p < .0001$).

The estimated coefficient for TIME is 0.383 with a standard error of 0.032. TIME is the number of years since the baseline visit. It is equivalent to aging. This result indicates that each additional year of aging of an individual is associated with on average a 0.383 mmHg increase in their systolic blood pressure given that their BMI did not change during the additional year. This association is statistically significant ($p < .0001$).

Output 4.3-3

name	Estimate	Std.Error	t.value	df	P...t..
(Intercept)	128.3728	1.1027	116.4117	235.6975	0.0000
AGEGROUP_C61	-27.2606	0.9248	-29.4758	145.2213	0.0000
AGEGROUP_C62	-24.8551	0.8770	-28.3413	266.4549	0.0000
AGEGROUP_C63	-19.7686	0.8437	-23.4309	361.2890	0.0000
AGEGROUP_C64	-13.4759	0.8277	-16.2816	341.1863	0.0000
AGEGROUP_C65	-7.2584	0.8089	-8.9735	570.6038	0.0000
BKGRD1_C7NOMISS0	2.5772	0.8478	3.0399	74.2927	0.0033
BKGRD1_C7NOMISS1	1.6871	0.7168	2.3537	71.2485	0.0213
BKGRD1_C7NOMISS2	0.6119	0.7733	0.7912	100.7344	0.4307
BKGRD1_C7NOMISS4	0.9041	0.7281	1.2416	65.6437	0.2188
BKGRD1_C7NOMISS5	-1.6706	0.7082	-2.3588	135.9049	0.0198
BKGRD1_C7NOMISS6	1.7163	0.8569	2.0030	125.3378	0.0473
CENTERNUM1	0.9697	0.8080	1.2002	40.9470	0.2370
CENTERNUM2	-1.8702	0.4899	-3.8178	105.8392	0.0002
CENTERNUM3	4.5143	0.7898	5.7156	64.3619	0.0000
SEX1	5.5536	0.3732	14.8822	63.4101	0.0000
US_BORN1	-0.0325	0.4954	-0.0656	134.7921	0.9478
EMPLOYED2	-0.3422	0.7868	-0.4350	275.0483	0.6639
EMPLOYED3	-0.5016	0.8571	-0.5853	137.7983	0.5593
EMPLOYED4	0.3395	0.8082	0.4201	167.8230	0.6750
EDUCATION_C32	-1.1855	0.4807	-2.4660	45.3703	0.0175
EDUCATION_C33	-2.4708	0.4096	-6.0320	181.5131	0.0000
BMI	0.1922	0.0252	7.6359	255.1608	0.0000
TIME	0.3829	0.0318	12.0343	17.0428	0.0000

4.3.2. Visit 1 Sample, Visit-specific IPW

4.3.2.1. SAS

The steps for constructing the dataset sol_ipw_long are the same as those described in 4.2.2.1.

Model-Based GEE with IPW

```
proc genmod data=sol_ipw_long;
      class HH_ID AGEGROUP_C6(ref = '6') BKGRD1_C7NOMISS(ref = '3')
            CENTERNUM(ref = '4') SEX(ref = '0') US_BORN(ref = '0')
      EMPLOYED(ref = '1') EDUCATION_C3(ref = '1');
      weight WEIGHT_IPW_BY_VISIT;
      model SBP5 = AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX US_BORN
            EMPLOYED EDUCATION_C3 BMI TIME / dist=normal;
      repeated subject = HH_ID / corr=ind;
run;
```

This **proc genmod** procedure fits GEE to the long-format dataset sol_ipw_long. The syntax used in this analysis closely parallels that used in 4.3.1.1, with modifications reflecting the use of WEIGHT_IPW_BY_VISIT, which is used to supply the appropriate sampling weight for each visit data contribution and the omission of the by _IMPUTATION_ option, since the analysis is no longer conducted on imputed datasets. The WEIGHT_IPW_BY_VISIT variable equals the overall sampling weight for Visit 1 and the corresponding non-response-adjusted IPW weights for Visits 2 and 3. **proc genmod** automatically excludes records with missing weights, thereby ensuring consistency between the IPW-defined analysis and the GEE main analysis.

Estimates

Output 4.3-4 displays the estimates.

The estimated association between BMI and systolic blood pressure is 0.167 with a standard error of 0.033. This positive association indicates that, a 1 kg/m² increase in an individual's BMI is associated with on average a 0.167 mmHg increase in their systolic blood pressure. This effect is statistically significant ($p < .0001$).

The estimated coefficient for TIME is 0.648 with a standard error of 0.055. TIME is the number of years since the baseline visit. It is equivalent to aging. This result indicates that each additional year of aging of an individual is associated with on average a 0.648 mmHg increase in their systolic blood pressure given that their BMI did not change during the additional year. This association is statistically significant ($p < .0001$).

Output 4.3-4

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		124.7821	1.7030	121.4443	128.1200	73.27	<.0001
AGEGROUP_C6	1	-28.1828	1.4000	-30.9268	-25.4389	-20.13	<.0001
AGEGROUP_C6	2	-24.8336	1.3602	-27.4995	-22.1676	-18.26	<.0001
AGEGROUP_C6	3	-18.8240	1.3327	-21.4361	-16.2119	-14.12	<.0001
AGEGROUP_C6	4	-12.1214	1.2979	-14.6652	-9.5775	-9.34	<.0001
AGEGROUP_C6	5	-6.5631	1.2687	-9.0497	-4.0765	-5.17	<.0001
AGEGROUP_C6	6	0.0000	0.0000	0.0000	0.0000	.	.
BKGRD1_C7NOMISS	0	2.0197	1.0131	0.0341	4.0054	1.99	0.0462
BKGRD1_C7NOMISS	1	1.1951	0.8154	-0.4031	2.7934	1.47	0.1428
BKGRD1_C7NOMISS	2	0.9825	0.9951	-0.9680	2.9329	0.99	0.3235
BKGRD1_C7NOMISS	4	0.8099	1.0365	-1.2216	2.8414	0.78	0.4346
BKGRD1_C7NOMISS	5	-2.1689	0.8698	-3.8737	-0.4642	-2.49	0.0126
BKGRD1_C7NOMISS	6	1.5014	0.9992	-0.4569	3.4597	1.50	0.1329
BKGRD1_C7NOMISS	3	0.0000	0.0000	0.0000	0.0000	.	.
CENTERNUM	1	1.5565	0.9619	-0.3288	3.4418	1.62	0.1056
CENTERNUM	2	-0.7751	0.6631	-2.0747	0.5246	-1.17	0.2425
CENTERNUM	3	6.3911	0.9256	4.5770	8.2052	6.90	<.0001
CENTERNUM	4	0.0000	0.0000	0.0000	0.0000	.	.
SEX	1	4.6480	0.4974	3.6730	5.6230	9.34	<.0001
SEX	0	0.0000	0.0000	0.0000	0.0000	.	.
US_BORN	1	0.3778	0.6602	-0.9162	1.6717	0.57	0.5672
US_BORN	0	0.0000	0.0000	0.0000	0.0000	.	.
EMPLOYED	2	0.6316	1.2730	-1.8634	3.1266	0.50	0.6198
EMPLOYED	3	0.0541	1.3185	-2.5301	2.6382	0.04	0.9673
EMPLOYED	4	1.0694	1.2716	-1.4229	3.5617	0.84	0.4004
EMPLOYED	1	0.0000	0.0000	0.0000	0.0000	.	.
EDUCATION_C3	2	-1.1483	0.6102	-2.3443	0.0476	-1.88	0.0598
EDUCATION_C3	3	-3.4578	0.6003	-4.6344	-2.2811	-5.76	<.0001
EDUCATION_C3	1	0.0000	0.0000	0.0000	0.0000	.	.
BMI		0.1666	0.0333	0.1013	0.2319	5.00	<.0001
TIME		0.6479	0.0549	0.5403	0.7555	11.80	<.0001

4.3.2.2. R

Construction of IPW Indicators

```
library(dplyr)
library(mice)
library(glmtoolbox)

sol <- read_sas("sol_wide.sas7bdat")

sol_ipw_base_cont <- sol %>%
  mutate(
    PARTICIPANT_V2_NOMISS =
      ifelse(
        PARTICIPANT_V2 == 1 &
          complete.cases(
            AGEGROUP_C6, BKGRD1_C7NOMISS, CENTERNUM, SEX, US_BORN,
            EMPLOYED, EDUCATION_C3, BMI_V1,
            SBP5_V2, BMI_V2, YRS_BTWN_V1V2
          ),
        1, 0
      ),
    PARTICIPANT_EXAMONLY_V3_NOMISS =
      ifelse(
        PARTICIPANT_EXAMONLY_V3 == 1 &
          complete.cases(
            AGEGROUP_C6, BKGRD1_C7NOMISS, CENTERNUM, SEX, US_BORN,
            EMPLOYED, EDUCATION_C3, BMI_V1,
            SBP5_V3, BMI_V3, YRS_BTWN_V1V3
          ),
        1, 0
      )
  )
```

This section reads the wide-format analytic dataset (sol_wide.sas7bdat) as sol and then uses mutate() with ifelse() and complete.cases() to construct two visit-specific data contribution indicators for the IPW models:

PARTICIPANT_V2_NOMISS is set to 1 for participants who attended Visit 2 (PARTICIPANT_V2 = 1) and have complete data on all baseline covariates used in the main model, as well as SBP5_V2, BMI_V2, and YRS_BTWN_V1V2; otherwise it is set to 0.

PARTICIPANT_EXAMONLY_V3_NOMISS is defined analogously using PARTICIPANT_EXAMONLY_V3, requiring complete baseline covariates along with SBP5_V3, BMI_V3, and YRS_BTWN_V1V3.

Together, these binary indicators define visit-specific eligibility for contributing follow-up data, combining participation status and required data completeness at Visit 2 and Visit 3 (exam-only).

MI for IPW Estimation

```
mi_vars_ipw_cont <- c(
  "ID",
  "AGEGROUP_C6", "BKGRD1_C7NOMISS", "CENTERNUM", "SEX", "US_BORN",
  "EMPLOYED", "EDUCATION_C3", "BMI_V1",
  "WEIGHT_FINAL_NORM_OVERALL",
  "PARTICIPANT_V2", "PARTICIPANT_V2_NOMISS",
  "PARTICIPANT_EXAMONLY_V3_NOMISS"
)

dat_ipw_mi_cont <- sol_ipw_base_cont %>%
  select(all_of(mi_vars_ipw_cont)) %>%
  mutate(
    AGEGROUP_C6 = factor(AGEGROUP_C6),
    BKGRD1_C7NOMISS = factor(BKGRD1_C7NOMISS),
    CENTERNUM = factor(CENTERNUM),
    SEX = factor(SEX),
    US_BORN = factor(US_BORN),
    EMPLOYED = factor(EMPLOYED),
    EDUCATION_C3 = factor(EDUCATION_C3),
    PARTICIPANT_V2 = factor(PARTICIPANT_V2),
    PARTICIPANT_V2_NOMISS = factor(PARTICIPANT_V2_NOMISS),
    PARTICIPANT_EXAMONLY_V3_NOMISS =
factor(PARTICIPANT_EXAMONLY_V3_NOMISS)
  )

pred_ipw_cont <- quickpred(
  dat_ipw_mi_cont,
  include = c(
    "AGEGROUP_C6", "BKGRD1_C7NOMISS", "CENTERNUM", "SEX", "US_BORN",
    "EMPLOYED", "EDUCATION_C3", "BMI_V1", "WEIGHT_FINAL_NORM_OVERALL",
    "PARTICIPANT_V2"
  )
)

meth_ipw_cont <- make.method(dat_ipw_mi_cont)
meth_ipw_cont["BMI_V1"] <- "norm"
meth_ipw_cont[c("US_BORN","SEX")] <- "logreg"
meth_ipw_cont[c("EMPLOYED","AGEGROUP_C6")] <- "polr"
meth_ipw_cont[c("EDUCATION_C3")] <- "polyreg"


imp_ipw_cont <- mice(
  dat_ipw_mi_cont,
  m = 10,
  seed = 2024,
  predictorMatrix = pred_ipw_cont,
  method = meth_ipw_cont,
  printFlag = FALSE
)
```

This section performs MI for variables to be used in the response models that define IPW.

First, it defines the imputation variable set (mi_vars_ipw_cont) and constructs the imputation dataset (dat_ipw_mi_cont) by using select(all_of()) and mutate() to retain the baseline covariates, the overall Visit 1 weight (WEIGHT_FINAL_NORM_OVERALL), and the visit-specific participation indicators. Categorical variables (e.g., AGEGROUP_C6, EDUCATION_C3, PARTICIPANT_V2, and the *_NOMISS indicators) are explicitly converted to factors.

Next, it creates the predictor matrix (pred_ipw_cont) using quickpred(), specifying an include set so that imputation of missing baseline variables conditions on the baseline covariates, baseline BMI (BMI_V1), the overall sampling weight, and Visit 2 participation (PARTICIPANT_V2). Imputation methods are then specified via make.method() and meth_ipw_cont, using "norm" for continuous baseline BMI, "logreg" for binary variables (US_BORN, SEX), "polr" for ordinal variables (EMPLOYED, AGEGROUP_C6), and "polyreg" for nominal variables (EDUCATION_C3).

Finally, mice() is used to run FCS MI (m = 10, seed = 2024) with predictorMatrix = pred_ipw_cont and method = meth_ipw_cont, producing the imputed object imp_ipw_cont containing the completed baseline datasets for the IPW response models.

Estimation of IPW Models by Visit

```
get_xbeta <- function(df, formula) {
  stopifnot("ID" %in% names(df))
  fit <- glm(formula, data = df, family = binomial(link = "logit"),
na.action = na.exclude)
  xb  <- predict(fit, newdata = df, type = "link")  # aligned to df rows
  tibble::tibble(ID = df$ID, xb = as.numeric(xb))
}

v2_formula_cont <- PARTICIPANT_V2_NOMISS ~
  AGEGROUP_C6 + BKGRD1_C7NOMISS + CENTERNUM + SEX +
  US_BORN + EMPLOYED + EDUCATION_C3 + BMI_V1 + WEIGHT_FINAL_NORM_OVERALL

v3_formula_cont <- PARTICIPANT_EXAMONLY_V3_NOMISS ~
  AGEGROUP_C6 + BKGRD1_C7NOMISS + CENTERNUM + SEX +
  US_BORN + EMPLOYED + EDUCATION_C3 + BMI_V1 +
  WEIGHT_FINAL_NORM_OVERALL + PARTICIPANT_V2

xb_v2_all_cont <- vector("list", length = imp_ipw_cont$m)
xb_v3_all_cont <- vector("list", length = imp_ipw_cont$m)

for (m in 1:imp_ipw_cont$m) {
  df_m <- complete(imp_ipw_cont, action = m)
  xb_v2_all_cont[[m]] <- get_xbeta(df_m, v2_formula_cont) %>% mutate(.imp
= m)
  xb_v3_all_cont[[m]] <- get_xbeta(df_m, v3_formula_cont) %>% mutate(.imp
= m)
}

xb_v2_all_cont <- bind_rows(xb_v2_all_cont)
xb_v3_all_cont <- bind_rows(xb_v3_all_cont)

pred_v2_bar_cont <- xb_v2_all_cont %>%
  group_by(ID) %>%
  summarise(lp_v2 = mean(xb, na.rm = TRUE), .groups = "drop") %>%
  mutate(RR_V2 = plogis(lp_v2))

pred_v3_bar_cont <- xb_v3_all_cont %>%
  group_by(ID) %>%
  summarise(lp_v3 = mean(xb, na.rm = TRUE), .groups = "drop") %>%
  mutate(RR_V3 = plogis(lp_v3))
```

This section fits the response models within each imputed dataset and extracts the corresponding linear predictors, then averages predictions across imputations to obtain participant-specific response probabilities.

First, a function get_xbeta() is defined to (i) fit a logistic regression model using glm(..., family = binomial(link="logit")) and (ii) compute the aligned linear predictor (type = "link") using predict(), returning a two-column tibble with ID and xb. Two model formulas are then specified: v2_formula_cont for PARTICIPANT_V2_NOMISS and v3_formula_cont for PARTICIPANT_EXAMONLY_V3_NOMISS, where the Visit 3 model additionally includes PARTICIPANT_V2 to reflect prior participation.

Next, a loop over imputations (m in 1:imp_ipw_cont$m) uses complete(imp_ipw_cont, action = m) to extract each completed dataset, applies get_xbeta() to both formulas, and stores the results in xb_v2_all_cont and xb_v3_all_cont (tagged by .imp). These are then combined using bind_rows().

Finally, the linear predictors are averaged across imputations by ID using group_by() and summarise(mean(xb)), producing lp_v2 and lp_v3. These averaged linear predictors are transformed to response probabilities via plogis() to yield RR_V2 and RR_V3, stored in pred_v2_bar_cont and pred_v3_bar_cont.

Construction of IPW Weights by Visit

```
sol_ipw_wide_cont <- sol_ipw_base_cont %>%
  left_join(pred_v2_bar_cont %>% select(ID, RR_V2), by = "ID") %>%
  left_join(pred_v3_bar_cont %>% select(ID, RR_V3), by = "ID") %>%
  mutate(
    WEIGHT_IPW_V2 =
      ifelse(PARTICIPANT_V2_NOMISS == 1, WEIGHT_NONRESP / RR_V2,
NA_real_),
    WEIGHT_EXAMONLY_IPW_V3 =
      ifelse(PARTICIPANT_EXAMONLY_V3_NOMISS == 1, WEIGHT_NONRESP / RR_V3,
NA_real_)
  )
```

This section merges the imputation-averaged response probabilities back into the analysis dataset and computes the final visit-specific IPW weights.

First, left_join() is used to merge RR_V2 and RR_V3 (from pred_v2_bar_cont and pred_v3_bar_cont) into the wide-format dataset, creating sol_ipw_wide_cont. Then, mutate() with ifelse() defines two visit-specific IPW weights: WEIGHT_IPW_V2 is computed as WEIGHT_NONRESP / RR_V2 for participants with PARTICIPANT_V2_NOMISS == 1, and WEIGHT_EXAMONLY_IPW_V3 is computed as WEIGHT_NONRESP / RR_V3 for participants with PARTICIPANT_EXAMONLY_V3_NOMISS == 1. For participants who do not contribute visit-specific data (i.e., the corresponding *_NOMISS indicator is 0), the weight is set to missing (NA_real_).

Creation of IPW Long-Format Dataset

```
sol_ipw_long_cont <- dplyr::bind_rows(
  sol_ipw_wide_cont %>%
    transmute(
      ID, HH_ID,
      AGEGROUP_C6, BKGRD1_C7NOMISS, CENTERNUM, SEX, US_BORN, EMPLOYED,
EDUCATION_C3,
      VISIT = 1L,
      TIME  = 0,
      BMI   = BMI_V1,
      SBP5  = SBP5_V1,
      WEIGHT_IPW_BY_VISIT = WEIGHT_FINAL_NORM_OVERALL,
      WEIGHT_EXAMONLY_IPW_V3
    ),
  sol_ipw_wide_cont %>%
    transmute(
      ID, HH_ID,
      AGEGROUP_C6, BKGRD1_C7NOMISS, CENTERNUM, SEX, US_BORN, EMPLOYED,
EDUCATION_C3,
      VISIT = 2L,
      TIME  = YRS_BTWN_V1V2,
      BMI   = BMI_V2,
      SBP5  = SBP5_V2,
      WEIGHT_IPW_BY_VISIT = WEIGHT_IPW_V2,
      WEIGHT_EXAMONLY_IPW_V3
    ),
  sol_ipw_wide_cont %>%
    transmute(
      ID, HH_ID,
      AGEGROUP_C6, BKGRD1_C7NOMISS, CENTERNUM, SEX, US_BORN, EMPLOYED,
EDUCATION_C3,
      VISIT = 3L,
      TIME  = YRS_BTWN_V1V3,
      BMI   = BMI_V3,
      SBP5  = SBP5_V3,
      WEIGHT_IPW_BY_VISIT = WEIGHT_EXAMONLY_IPW_V3,
      WEIGHT_EXAMONLY_IPW_V3
    )
) %>%
  arrange(ID, VISIT)
```

This section reshapes the IPW-enhanced wide dataset into a long (person-visit) format and assigns the appropriate weight for each visit.

First, dplyr::bind_rows() is used to stack three visit-specific datasets created via transmute() from sol_ipw_wide_cont, one row per participant per visit. In each transmute() call, the baseline covariates and identifiers (ID, HH_ID, and the categorical covariates) are carried forward, and a standardized set of longitudinal variables is created: VISIT indicates the visit number (1, 2, or 3), TIME is set to 0 at Visit 1 and to YRS_BTWN_V1V2 or YRS_BTWN_V1V3 at Visits 2 and 3, and the visit-specific measures are mapped into common columns (BMI from BMI_V1/BMI_V2/BMI_V3 and SBP5 from SBP5_V1/SBP5_V2/SBP5_V3).

Next, the visit-specific analysis weight is stored in a single column, WEIGHT_IPW_BY_VISIT, using the appropriate source by visit: WEIGHT_FINAL_NORM_OVERALL for Visit 1, WEIGHT_IPW_V2 for Visit 2, and WEIGHT_EXAMONLY_IPW_V3 for Visit 3. The variable WEIGHT_EXAMONLY_IPW_V3 is also retained as a separate column across all visits for reference.

Finally, arrange(ID, VISIT) orders the resulting long dataset (sol_ipw_long_cont) by participant and visit.

```
sol_ipw_long_cont <- sol_ipw_long_cont %>%
  mutate(
    AGEGROUP_C6        = factor(AGEGROUP_C6, levels =
sort(unique(sol$AGEGROUP_C6))),
    BKGRD1_C7NOMISS    = factor(BKGRD1_C7NOMISS, levels =
sort(unique(sol$BKGRD1_C7NOMISS))),
    CENTERNUM          = factor(CENTERNUM, levels =
sort(unique(sol$CENTERNUM))),
    SEX                = factor(SEX, levels = sort(unique(sol$SEX))),
    US_BORN            = factor(US_BORN, levels =
sort(unique(sol$US_BORN))),
    EMPLOYED           = factor(EMPLOYED, levels =
sort(unique(sol$EMPLOYED))),
    EDUCATION_C3       = factor(EDUCATION_C3, levels =
sort(unique(sol$EDUCATION_C3)))
  ) %>%
  mutate(
    AGEGROUP_C6     = relevel(AGEGROUP_C6, ref = "6"),
    BKGRD1_C7NOMISS = relevel(BKGRD1_C7NOMISS, ref = "3"),
    CENTERNUM       = relevel(CENTERNUM, ref = "4"),
    SEX             = relevel(SEX, ref = "0"),
    US_BORN         = relevel(US_BORN, ref = "0"),
    EMPLOYED        = relevel(EMPLOYED, ref = "1"),
    EDUCATION_C3    = relevel(EDUCATION_C3, ref = "1")
  )
```

This section finalizes the coding of categorical covariates in the long-format dataset prior to model fitting.

First, mutate() with factor() is used to explicitly define each categorical variable (AGEGROUP_C6, BKGRD1_C7NOMISS, CENTERNUM, SEX, US_BORN, EMPLOYED, EDUCATION_C3) as a factor with levels set to the sorted unique values observed in the original dataset sol. This step ensures consistent and complete factor level definitions across all visits and imputations.

Next, a second mutate() call applies relevel() to each factor to assign the desired reference categories for the regression models. Specifically, reference levels are set for age group ("6"), background ("3"), center ("4"), sex ("0"), U.S. birth status ("0"), employment status ("1"), and education ("1"). This guarantees that all model coefficients are interpreted relative to the predefined reference groups.

Model-Based GEE with IPW

```
glmgee(
  SBP5 ~ AGEGROUP_C6 + BKGRD1_C7NOMISS + CENTERNUM + SEX +
    US_BORN + EMPLOYED + EDUCATION_C3 + BMI + TIME,
  data   = sol_ipw_long_cont,
  id     = HH_ID,
  corstr = "independence",
  weight = WEIGHT_IPW_BY_VISIT,
  family = gaussian(link = "identity")
)
```

This section fits a weighted GEE for the continuous outcome using the Visit 1 analytic sample with inverse probability weights that vary by visit to account for follow-up non-response.

Specifically, glmgee() is used to regress SBP5 on baseline covariates (AGEGROUP_C6, BKGRD1_C7NOMISS, CENTERNUM, SEX, US_BORN, EMPLOYED, EDUCATION_C3) and the time-varying covariates BMI and TIME. Repeated measurements within households are handled by specifying HH_ID as the clustering variable and using an independence working correlation structure (corstr = "independence").

Inverse probability weighting is incorporated through WEIGHT_IPW_BY_VISIT, which applies the overall Visit 1 sampling weight at baseline and visit-specific non-response-adjusted IPW weights at follow-up visits. The model assumes a Gaussian mean structure with an identity link (family = gaussian(link = "identity")).

Estimates

Output 4.3-5 displays the estimates.

The estimated association between BMI and systolic blood pressure is 0.167 with a standard error of 0.033. This positive association indicates that, a 1 kg/m² increase in an individual's BMI is associated with on average a 0.167 mmHg increase in their systolic blood pressure. This effect is statistically significant ($p < .0001$).

The estimated coefficient for TIME is 0.648 with a standard error of 0.055. TIME is the number of years since the baseline visit. It is equivalent to aging. This result indicates that each additional year of aging of an individual is associated with on average a 0.648 mmHg increase in their systolic blood pressure given that their BMI did not change during the additional year. This association is statistically significant ($p < .0001$).

Output 4.3-5

Coefficients

	Estimate	Std.Error	z-value	Pr(>|z|)
(Intercept)	124.78027	1.70341	73.25343	< 2.22e-16
AGEGROUP_C61	-28.17990	1.40036	-20.12328	< 2.22e-16
AGEGROUP_C62	-24.83236	1.36069	-18.24987	< 2.22e-16
AGEGROUP_C63	-18.82357	1.33324	-14.11868	< 2.22e-16
AGEGROUP_C64	-12.11949	1.29824	-9.33535	< 2.22e-16
AGEGROUP_C65	-6.55985	1.26905	-5.16911	2.3521e-07
BKGRD1_C7NOMISS0	2.02005	1.01357	1.99301	0.046260
BKGRD1_C7NOMISS1	1.19530	0.81574	1.46529	0.142841
BKGRD1_C7NOMISS2	0.98163	0.99545	0.98611	0.324080
BKGRD1_C7NOMISS4	0.81299	1.03740	0.78367	0.433231
BKGRD1_C7NOMISS5	-2.16971	0.87004	-2.49381	0.012638
BKGRD1_C7NOMISS6	1.50399	0.99904	1.50543	0.132213
CENTERNUM1	1.55707	0.96258	1.61760	0.105750
CENTERNUM2	-0.77440	0.66340	-1.16731	0.243084
CENTERNUM3	6.39404	0.92595	6.90540	5.0062e-12
SEX1	4.64752	0.49785	9.33527	< 2.22e-16
US_BORN1	0.37838	0.66037	0.57298	0.566660
EMPLOYED2	0.63326	1.27203	0.49783	0.618601
EMPLOYED3	0.05765	1.31755	0.04376	0.965099
EMPLOYED4	1.07191	1.27076	0.84352	0.398935
EDUCATION_C32	-1.14804	0.61047	-1.88059	0.060028
EDUCATION_C33	-3.45853	0.60082	-5.75636	8.5945e-09
BMI	0.16650	0.03334	4.99475	5.8912e-07
TIME	0.64794	0.05493	11.79610	< 2.22e-16

4.3.3. Visit 3 Sample, Visit 3 IPW

4.3.3.1. SAS

The steps for constructing the dataset sol_ipw_long are the same as those described in 4.2.2.1.

Model-Based GEE with IPW

```
proc genmod data=sol_ipw_long;
      class HH_ID AGEGROUP_C6(ref = '6') BKGRD1_C7NOMISS(ref = '3')
            CENTERNUM(ref = '4') SEX(ref = '0') US_BORN(ref = '0')
      EMPLOYED(ref = '1') EDUCATION_C3(ref = '1');
      weight WEIGHT_EXAMONLY_IPW_V3;
      model SBP5 = AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX US_BORN
            EMPLOYED EDUCATION_C3 BMI TIME / dist=normal;
      repeated subject = HH_ID / corr=ind;
run;
```

The analysis restricted to participants who attended Visit 3. In this setting, the Visit 3 non-response-adjusted weight WEIGHT_EXAMONLY_IPW_V3 is available in sol_ipw_long. **proc genmod** automatically excludes records with missing weights, ensuring that only participants who contributed Visit 3 exam data are included in the estimation and that the IPW-defined analysis is consistent with the GEE main analysis. All other aspects of the modeling framework are identical to those described for the Visit 1 sample IPW analysis.

Estimates

Output 4.3-6 displays the estimates.

The estimated association between BMI and systolic blood pressure is 0.183 with a standard error of 0.034. This positive association indicates that, a 1 kg/m² increase in an individual's BMI is associated with on average a 0.183 mmHg increase in their systolic blood pressure. This effect is statistically significant ($p < .0001$).

The estimated coefficient for TIME is 0.464 with a standard error of 0.029. TIME is the number of years since the baseline visit. It is equivalent to aging. This result indicates that each additional year of aging of an individual is associated with on average a 0.464 mmHg increase in their systolic blood pressure given that their BMI did not change during the additional year. This association is statistically significant ($p < .0001$).

Output 4.3-6

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		126.4398	1.6384	123.2286	129.6510	77.17	<.0001
AGEGROUP_C6	1	-26.0197	1.4993	-28.9582	-23.0812	-17.36	<.0001
AGEGROUP_C6	2	-24.0490	1.4332	-26.8580	-21.2401	-16.78	<.0001
AGEGROUP_C6	3	-18.9584	1.4200	-21.7416	-16.1753	-13.35	<.0001
AGEGROUP_C6	4	-12.1544	1.3813	-14.8617	-9.4471	-8.80	<.0001
AGEGROUP_C6	5	-6.4937	1.3439	-9.1277	-3.8597	-4.83	<.0001
AGEGROUP_C6	6	0.0000	0.0000	0.0000	0.0000	.	.
BKGRD1_C7NOMISS	0	2.0085	0.9646	0.1179	3.8992	2.08	0.0373
BKGRD1_C7NOMISS	1	1.1642	0.8206	-0.4441	2.7725	1.42	0.1560
BKGRD1_C7NOMISS	2	0.8781	0.9713	-1.0257	2.7819	0.90	0.3660
BKGRD1_C7NOMISS	4	1.3669	0.9771	-0.5482	3.2820	1.40	0.1618
BKGRD1_C7NOMISS	5	-1.5189	0.9400	-3.3613	0.3234	-1.62	0.1061
BKGRD1_C7NOMISS	6	1.0616	0.9602	-0.8204	2.9437	1.11	0.2689
BKGRD1_C7NOMISS	3	0.0000	0.0000	0.0000	0.0000	.	.
CENTERNUM	1	1.3288	0.9335	-0.5007	3.1584	1.42	0.1546
CENTERNUM	2	-1.0348	0.6338	-2.2769	0.2074	-1.63	0.1025
CENTERNUM	3	6.0493	0.9163	4.2535	7.8452	6.60	<.0001
CENTERNUM	4	0.0000	0.0000	0.0000	0.0000	.	.
SEX	1	5.2929	0.4981	4.3167	6.2691	10.63	<.0001
SEX	0	0.0000	0.0000	0.0000	0.0000	.	.
US_BORN	1	0.1079	0.6397	-1.1458	1.3617	0.17	0.8660
US_BORN	0	0.0000	0.0000	0.0000	0.0000	.	.
EMPLOYED	2	-0.1669	1.3297	-2.7730	2.4393	-0.13	0.9001
EMPLOYED	3	-1.2256	1.3469	-3.8655	1.4144	-0.91	0.3629
EMPLOYED	4	0.6268	1.3162	-1.9529	3.2064	0.48	0.6339
EMPLOYED	1	0.0000	0.0000	0.0000	0.0000	.	.
EDUCATION_C3	2	-1.5042	0.5945	-2.6695	-0.3390	-2.53	0.0114
EDUCATION_C3	3	-3.2201	0.5891	-4.3747	-2.0655	-5.47	<.0001
EDUCATION_C3	1	0.0000	0.0000	0.0000	0.0000	.	.
BMI		0.1831	0.0339	0.1167	0.2495	5.41	<.0001
TIME		0.4639	0.0289	0.4072	0.5206	16.03	<.0001

4.3.3.2. Stata

Construction of IPW Indicators

```
local num 10

import sas using "sol_wide.sas7bdat", clear
set seed 2024

* If Stata shortened a long name, adapt here
capture confirm variable WEIGHT_NORM_OVERALL_EXAMONLY_V3
if _rc==0 rename WEIGHT_NORM_OVERALL_EXAMONLY_V3 WEIGHT_EXAMONLY_V3


local basecov AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX US_BORN EMPLOYED
EDUCATION_C3 BMI_V1

egen miss_v2 = rowmiss(`basecov' SBP5_V2 BMI_V2 YRS_BTWN_V1V2)
gen byte PARTICIPANT_V2_NOMISS = (PARTICIPANT_V2 == 1 & miss_v2 == 0)
drop miss_v2

egen miss_v3 = rowmiss(`basecov' SBP5_V3 BMI_V3 YRS_BTWN_V1V3)
gen byte PARTICIPANT_EXAMONLY_V3_NOMISS = (PARTICIPANT_EXAMONLY_V3 == 1 &
miss_v3 == 0)
drop miss_v3
```

This section sets up the Stata workflow, imports the wide-format dataset, and constructs the visit-specific analytic contribution indicators used as outcomes in the response (RR) models.

First, it defines the number of imputations (local num 10), reads the SAS dataset into memory using import sas, and sets a fixed random seed (set seed 2024) for reproducibility. It then checks whether Stata truncated a long weight variable name using capture confirm variable ...; if the truncated version exists, it is harmonized via rename so the expected weight variable name (WEIGHT_EXAMONLY_V3) is available for downstream steps.

Next, it defines the baseline covariate set in a macro (local basecov ...). For Visit 2, it uses egen ... = rowmiss() to count missing values across the baseline covariates plus SBP5_V2, BMI_V2, and YRS_BTWN_V1V2, then creates the binary flag PARTICIPANT_V2_NOMISS with gen byte to equal 1 only when the participant attended Visit 2 (PARTICIPANT_V2 == 1) and has no missing values in the required variables (miss_v2 == 0); otherwise it is 0. The temporary missingness counter is removed with drop.

The same logic is applied to Visit 3 exam-only participation: rowmiss() is used to evaluate completeness across baseline covariates plus SBP5_V3, BMI_V3, and YRS_BTWN_V1V3, and PARTICIPANT_EXAMONLY_V3_NOMISS is generated to equal 1 only for participants who completed the exam-only component (PARTICIPANT_EXAMONLY_V3 == 1) and have complete data on all required variables, with the temporary counter dropped afterward.

MI for IPW Estimation

```
mi set flong

mi register imputed AGEGROUP_C6 SEX BMI_V1 US_BORN EMPLOYED EDUCATION_C3
mi register passive BKGRD1_C7NOMISS CENTERNUM WEIGHT_FINAL_NORM_OVERALL

mi impute chained (regress) BMI_V1 (logit) US_BORN SEX (ologit) EMPLOYED
AGEGROUP_C6 (mlogit) EDUCATION_C3 = i.BKGRD1_C7NOMISS i.CENTERNUM
WEIGHT_FINAL_NORM_OVERALL, add(`num') rseed(2024)
```

This section sets up MI in Stata and imputes baseline variables needed for the response (RR) models using chained equations.

First, it initializes the MI framework in long format using mi set flong, which stores imputations in a stacked ("flong") structure. It then declares which variables are to be imputed versus carried through unchanged. Specifically, mi register imputed identifies baseline variables that may have missingness and will be imputed (including BMI_V1 and selected baseline categorical covariates such as AGEGROUP_C6, SEX, US_BORN, EMPLOYED, and EDUCATION_C3). In contrast, mi register passive designates variables that are retained as-is and not directly imputed (e.g., BKGRD1_C7NOMISS, CENTERNUM, and WEIGHT_FINAL_NORM_OVERALL).

Next, mi impute chained performs FCS MI with variable-type–appropriate models: (regress) for continuous BMI_V1, (logit) for binary variables (US_BORN, SEX), (ologit) for ordinal variables (EMPLOYED, AGEGROUP_C6), and (mlogit) for nominal variables (EDUCATION_C3). The imputation models condition on background and center (i.BKGRD1_C7NOMISS, i.CENTERNUM) and the overall baseline weight (WEIGHT_FINAL_NORM_OVERALL). The number of imputations is set by add(\num), and reproducibility is ensured via rseed(2024).

Estimation of IPW Models by Visit

```
tempfile base xb_v2_all xb_v3_all
save `base', replace

clear
set obs 0
gen ID = ""
gen double xb = .
gen int m = .
save `xb_v2_all', replace
save `xb_v3_all', replace

forvalues m = 1/`num' {
    use `base', clear
    mi extract `m', clear

    quietly logit PARTICIPANT_V2_NOMISS i.AGEGROUP_C6 i.BKGRD1_C7NOMISS
i.CENTERNUM i.SEX i.US_BORN i.EMPLOYED i.EDUCATION_C3 BMI_V1
WEIGHT_FINAL_NORM_OVERALL, nolog
    predict double xb_v2, xb

    keep ID xb_v2
    gen int m = `m'
    rename xb_v2 xb

    append using `xb_v2_all'
    save `xb_v2_all', replace

    use `base', clear
    mi extract `m', clear

    quietly logit PARTICIPANT_EXAMONLY_V3_NOMISS i.AGEGROUP_C6
i.BKGRD1_C7NOMISS i.CENTERNUM i.SEX i.US_BORN i.EMPLOYED i.EDUCATION_C3
BMI_V1 WEIGHT_FINAL_NORM_OVERALL i.PARTICIPANT_V2, nolog
    predict double xb_v3, xb

    keep ID xb_v3
    gen int m = `m'
    rename xb_v3 xb

    append using `xb_v3_all'
    save `xb_v3_all', replace
}

use `xb_v2_all', clear
collapse (mean) lp_v2=xb, by(ID)
gen double RR_V2 = invlogit(lp_v2)
tempfile rr2
save `rr2', replace

use `xb_v3_all', clear
collapse (mean) lp_v3=xb, by(ID)
gen double RR_V3 = invlogit(lp_v3)
tempfile rr3
save `rr3', replace
```

This section fits the visit-specific response (RR) models within each imputed dataset, saves the fitted linear predictors by participant, and then averages predictions across imputations to obtain participant-level response probabilities.

First, the current MI dataset is saved to a temporary file (tempfile base) to serve as a clean starting point for repeated extraction. Two empty "container" datasets are then initialized and saved (xb_v2_all and xb_v3_all) to accumulate the linear predictors (xb) across imputations for Visit 2 and Visit 3, respectively.

Next, the loop forvalues m = 1/num iterates over imputations. Within each iteration, mi extract m loads the m-th completed dataset. A logistic regression (quietly logit) is fit for the Visit 2 response indicator PARTICIPANT_V2_NOMISS using the baseline covariates and WEIGHT_FINAL_NORM_OVERALL, and predict generates the fitted linear predictor (xb_v2). Only ID and the linear predictor are retained, the imputation index m is added, and the results are appended to the cumulative file xb_v2_all using append and save.

The same process is repeated for Visit 3: the same imputation is reloaded via mi extract m, a logistic regression is fit for PARTICIPANT_EXAMONLY_V3_NOMISS with the same baseline covariates and WEIGHT_FINAL_NORM_OVERALL, additionally conditioning on i.PARTICIPANT_V2, and the fitted linear predictor (xb_v3) is generated and appended to xb_v3_all.

After looping over all imputations, the accumulated linear predictors are averaged across imputations by participant using collapse (mean) ... , by(ID), yielding lp_v2 and lp_v3. These averaged linear predictors are transformed to response probabilities using invlogit() to produce RR_V2 and RR_V3, which are saved as temporary datasets (rr2 and rr3) for downstream weight construction.

Construction of IPW Weights by Visit

```
use `base', clear
mi extract 0, clear

merge 1:1 ID using `rr2', nogen
merge 1:1 ID using `rr3', nogen

gen double WEIGHT_IPW_V2 = .
replace WEIGHT_IPW_V2 = WEIGHT_NONRESP / RR_V2 if PARTICIPANT_V2_NOMISS ==
1

gen double WEIGHT_EXAMONLY_IPW_V3 = .
replace WEIGHT_EXAMONLY_IPW_V3 = WEIGHT_NONRESP / RR_V3 if
PARTICIPANT_EXAMONLY_V3_NOMISS == 1
```

This section merges the imputation-averaged response probabilities back into the main wide dataset and then computes the visit-specific inverse probability weights.

First, it reloads the saved dataset (use \base') and extracts the observed ("m = 0") version of the MI data using mi extract 0. The participant-level response probabilities are then merged in by ID using two one-to-one merges (merge 1:1 ID using `rr2' and merge 1:1 ID using `rr3'), adding RR_V2 and RR_V3 to the working dataset.

Next, the Visit 2 IPW weight WEIGHT_IPW_V2 is initialized to missing (gen double ... = .) and then filled in using replace as WEIGHT_NONRESP / RR_V2 for participants who contribute Visit 2 data (PARTICIPANT_V2_NOMISS == 1). All other records retain missing values for this weight.

Similarly, the Visit 3 exam-only IPW weight WEIGHT_EXAMONLY_IPW_V3 is generated and set to WEIGHT_NONRESP / RR_V3 for participants who contribute Visit 3 exam-only data (PARTICIPANT_EXAMONLY_V3_NOMISS == 1), leaving the weight missing otherwise.

Creation of IPW Long-Format Dataset

```
gen double TIME_1 = 0
gen double TIME_2 = YRS_BTWN_V1V2
gen double TIME_3 = YRS_BTWN_V1V3

gen double WBY_1 = WEIGHT_FINAL_NORM_OVERALL
gen double WBY_2 = WEIGHT_IPW_V2
gen double WBY_3 = WEIGHT_EXAMONLY_IPW_V3

rename SBP5_V1 SBP5_1
rename SBP5_V2 SBP5_2
rename SBP5_V3 SBP5_3

rename BMI_V1 BMI_1
rename BMI_V2 BMI_2
rename BMI_V3 BMI_3

reshape long SBP5_ BMI_ TIME_ WBY_, i(ID) j(VISIT)

rename SBP5_ SBP5
rename BMI_ BMI
rename TIME_ TIME
rename WBY_ WEIGHT_IPW_BY_VISIT

sort ID VISIT
```

This section converts the wide-format dataset into a person-visit long format and consolidates the visit-specific time variables and weights into single columns.

First, it creates visit-indexed time variables (TIME_1, TIME_2, TIME_3), setting baseline time to 0 and using the elapsed-time variables for Visits 2 and 3. It also creates visit-indexed weight variables (WBY_1, WBY_2, WBY_3), assigning the overall baseline sampling weight to Visit 1, the Visit 2 IPW weight to Visit 2, and the Visit 3 exam-only IPW weight to Visit 3.

Next, it harmonizes the naming of visit-specific outcome and covariate variables by renaming SBP5_V1/SBP5_V2/SBP5_V3 to SBP5_1/SBP5_2/SBP5_3 and BMI_V1/BMI_V2/BMI_V3 to BMI_1/BMI_2/BMI_3, so they match the expected structure for reshaping.

Then, reshape long SBP5_ BMI_ TIME_ WBY_, i(ID) j(VISIT) stacks the data into long format, producing one record per participant per visit and creating the visit indicator VISIT. After reshaping, the suffix-based variable names are cleaned by renaming SBP5_ to SBP5, BMI_ to BMI, TIME_ to TIME, and WBY_ to WEIGHT_IPW_BY_VISIT.

Finally, sort ID VISIT orders the long-format dataset by participant and visit.

Model-Based GEE with IPW

```
encode ID, gen(ID_NUM)
xtset ID_NUM

xtgee SBP5 ib6.AGEGROUP_C6 ib6.BKGRD1_C7NOMISS ib4.CENTERNUM ib0.SEX
ib0.US_BORN ib1.EMPLOYED ib1.EDUCATION_C3 BMI TIME
[pw=WEIGHT_EXAMONLY_IPW_V3], family(gaussian) corr(independent)
vce(robust)
```

This section fits the IPW-weighted GEE for the continuous outcome using the long-format Visit 3 exam-only analytic sample.

First, encode ID, gen(ID_NUM) converts the participant identifier ID into a numeric variable (ID_NUM) suitable for panel-data commands. The panel structure is then declared with xtset ID_NUM, indicating repeated measurements at the participant level.

Next, xtgee is used to regress SBP5 on the baseline covariates (AGEGROUP_C6, BKGRD1_C7NOMISS, CENTERNUM, SEX, US_BORN, EMPLOYED, EDUCATION_C3) and the time-varying covariates BMI and TIME. Reference categories for the categorical covariates are explicitly specified using the ib#. prefix to ensure consistent interpretation of regression coefficients.

IPW is incorporated via the probability weight option [pw=WEIGHT_EXAMONLY_IPW_V3], which applies the Visit 3 exam-only IPW weights. Observations with missing weights are automatically excluded. The model assumes a Gaussian outcome (family(gaussian)), uses an independence working correlation structure (corr(independent)), and reports robust sandwich standard errors (vce(robust)).

Estimates

Output 4.3-7 displays the estimates.

The estimated association between BMI and systolic blood pressure is 0.183 with a standard error of 0.033. This positive association indicates that, a 1 kg/m² increase in an individual's BMI is associated with on average a 0.183 mmHg increase in their systolic blood pressure. This effect is statistically significant ($p < .0001$).

The estimated coefficient for TIME is 0.464 with a standard error of 0.029. TIME is the number of years since the baseline visit. It is equivalent to aging. This result indicates that each additional year of aging of an individual is associated with on average a 0.464 mmHg increase in their systolic blood pressure given that their BMI did not change during the additional year. This association is statistically significant ($p < .0001$).

Output 4.3-7

SBP5	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
AGEGROUP_C6						
1	-26.02275	1.511517	-17.22	0.000	-28.98527	-23.06023
2	-24.05363	1.465389	-16.41	0.000	-26.92573	-21.18152
3	-18.96177	1.439225	-13.17	0.000	-21.7826	-16.14094
4	-12.15714	1.400995	-8.68	0.000	-14.90304	-9.411241
5	-6.494883	1.361549	-4.77	0.000	-9.16347	-3.826295
BKGRD1_C7NOMISS						
0	.9445665	1.049947	0.90	0.368	-1.113293	3.002426
1	.1012654	1.151709	0.09	0.930	-2.156042	2.358573
2	-.1895191	1.226169	-0.15	0.877	-2.592766	2.213728
3	-1.066579	.9629004	-1.11	0.268	-2.953829	.8206714
4	.3049308	1.100012	0.28	0.782	-1.851053	2.460915
5	-2.583895	1.176412	-2.20	0.028	-4.88962	-.2781694
CENTERNUM						
1	1.327012	.9510371	1.40	0.163	-.5369863	3.191011
2	-1.033849	.649209	-1.59	0.111	-2.306275	.2385776
3	6.050946	.9233268	6.55	0.000	4.241259	7.860633
1.SEX	5.29367	.4888088	10.83	0.000	4.335623	6.251718
1.US_BORN	.1085473	.6404192	0.17	0.865	-1.146651	1.363746
EMPLOYED						
2	-.163685	1.3468	-0.12	0.903	-2.803364	2.475994
3	-1.221483	1.351643	-0.90	0.366	-3.870653	1.427688
4	.6288978	1.324579	0.47	0.635	-1.967229	3.225025
EDUCATION_C3						
2	-1.504629	.5938139	-2.53	0.011	-2.668483	-.3407753
3	-3.220366	.5970173	-5.39	0.000	-4.390498	-2.050234
BMI	.1832006	.0328382	5.58	0.000	.1188389	.2475624
TIME	.4638432	.0293361	15.81	0.000	.4063455	.521341
_cons	127.5021	1.88721	67.56	0.000	123.8032	131.2009

4.3.3.3. R

The steps for constructing the dataset sol_ipw_long_cont are the same as those described in 4.3.2.2.

Model-Based GEE with IPW

```
glmgee(
  SBP5 ~ AGEGROUP_C6 + BKGRD1_C7NOMISS + CENTERNUM + SEX +
    US_BORN + EMPLOYED + EDUCATION_C3 + BMI + TIME,
  data   = sol_ipw_long_cont,
  id     = HH_ID,
  corstr = "independence",
  weight = WEIGHT_EXAMONLY_IPW_V3,
  family = gaussian(link = "identity")
)
```

This section fits a weighted GEE for the continuous outcome using the Visit 3 exam-only analytic sample with IPW weights.

The model regresses SBP5 on the baseline covariates (AGEGROUP_C6, BKGRD1_C7NOMISS, CENTERNUM, SEX, US_BORN, EMPLOYED, EDUCATION_C3) and time-varying covariates (BMI and TIME). Correlation among repeated observations is accounted for by specifying the household identifier HH_ID as the clustering variable and using an independence working correlation structure (corstr = "independence").

IPW is applied through WEIGHT_EXAMONLY_IPW_V3. The model is fit assuming a Gaussian outcome with an identity link (family = gaussian(link = "identity")). Observations with missing weights are automatically excluded from model fitting, ensuring that the IPW-defined analysis is consistent with the GEE main analysis.

Estimates

Output 4.3-8 displays the estimates.

The estimated association between BMI and systolic blood pressure is 0.183 with a standard error of 0.034. This positive association indicates that, a 1 kg/m² increase in an individual's BMI is associated with on average a 0.183 mmHg increase in their systolic blood pressure. This effect is statistically significant ($p < .0001$).

The estimated coefficient for TIME is 0.464 with a standard error of 0.029. TIME is the number of years since the baseline visit. It is equivalent to aging. This result indicates that each additional year of aging of an individual is associated with on average a 0.464 mmHg increase in their systolic blood pressure given that their BMI did not change during the additional year. This association is statistically significant ($p < .0001$).

Output 4.3-8

Coefficients

	Estimate	Std.Error	z-value	Pr(>|z|)
(Intercept)	126.43404	1.63941	77.12156	< 2.22e-16
AGEGROUP_C61	-26.01880	1.49997	-17.34616	< 2.22e-16
AGEGROUP_C62	-24.05153	1.43381	-16.77452	< 2.22e-16
AGEGROUP_C63	-18.95989	1.42063	-13.34609	< 2.22e-16
AGEGROUP_C64	-12.15506	1.38176	-8.79680	< 2.22e-16
AGEGROUP_C65	-6.49345	1.34438	-4.83006	1.3649e-06
BKGRD1_C7NOMISS0	2.01091	0.96505	2.08373	0.037185
BKGRD1_C7NOMISS1	1.16595	0.82104	1.42009	0.155581
BKGRD1_C7NOMISS2	0.87747	0.97175	0.90297	0.366540
BKGRD1_C7NOMISS4	1.37102	0.97800	1.40187	0.160955
BKGRD1_C7NOMISS5	-1.51898	0.94062	-1.61488	0.106337
BKGRD1_C7NOMISS6	1.06558	0.96021	1.10973	0.267116
CENTERNUM1	1.32681	0.93422	1.42024	0.155537
CENTERNUM2	-1.03415	0.63414	-1.63078	0.102938
CENTERNUM3	6.05205	0.91681	6.60118	4.0790e-11
SEX1	5.29344	0.49854	10.61783	< 2.22e-16
US_BORN1	0.10688	0.63980	0.16706	0.867326
EMPLOYED2	-0.16152	1.32927	-0.12151	0.903284
EMPLOYED3	-1.21855	1.34665	-0.90487	0.365533
EMPLOYED4	0.63144	1.31586	0.47987	0.631323
EDUCATION_C32	-1.50482	0.59486	-2.52971	0.011416
EDUCATION_C33	-3.22020	0.58963	-5.46139	4.7243e-08
BMI	0.18311	0.03389	5.40224	6.5813e-08
TIME	0.46388	0.02894	16.02789	< 2.22e-16

4.4. Results Summary

In **Table 4.4-1**, we summarize the key results from the illustrative example in this chapter, organized by analysis procedure, analytic sample, approach for addressing missing visits, and software implementation. The estimates from the various methods are similar as expected.

Table 4.4-1

Analysis Procedure	Analytic Sample	Missing-Visit Strategy	Software	Section	BMI Estimate (SE)	TIME Estimate (SE)
Complex-survey GEE	Visit 1 Sample	MI + Visit 1 Overall Sampling Weights	SUDAAN	4.2.1.1	0.197 (0.026)	0.398 (0.023)
		Visit-specific IPW	SUDAAN	4.2.2.1	0.167 (0.033)	0.648 (0.055)
	Visit 3 Sample	Visit 3 IPW	SUDAAN	4.2.3.1	0.183 (0.034)	0.464 (0.029)
Model-based GEE	Visit 1 Sample	MI + Visit 1 Overall Sampling Weights	SAS	4.3.1.1	0.197 (0.026)	0.398 (0.023)
			Stata	4.3.1.2	0.188 (0.026)	0.397 (0.022)
			R	4.3.1.3	0.192 (0.025)	0.383 (0.032)
		Visit-specific IPW	SAS	4.3.2.1	0.167 (0.033)	0.648 (0.055)
			R	4.3.2.2	0.167 (0.033)	0.648 (0.055)
	Visit 3 Sample	Visit 3 IPW	SAS	4.3.3.1	0.183 (0.034)	0.464 (0.029)
			Stata	4.3.3.2	0.183 (0.033)	0.464 (0.029)
			R	4.3.3.3	0.183 (0.034)	0.464 (0.029)

5. Examples for Longitudinal Analysis of Binary Outcomes

In this chapter, we illustrate the recommended MI and IPW methods for conducting longitudinal analysis of HCHS/SOL data for binary outcomes with repeated measures involving more than two clinic visits and accounting for HCHS/SOL complex survey design. We assume MAR for the missing-visit mechanism. To illustrate the proposed methods, the adjusted association between time-varying covariate BMI and outcome, a binary hypertension status using ACC/AHA definition, is used as an example, with sample code provided in SUDAAN, SAS, Stata, and R.

5.1. Illustrative Example

5.1.1. Model Specification and Covariates

As an example for illustration, we define the main model of interest as a longitudinal analysis examining the effect of time-varying BMI on a binary hypertension status over time (long-format: HYPERTENSION2_AHA; wide-format: HYPERTENSION2_AHA_V1, HYPERTENSION2_AHA_V2, HYPERTENSION2_AHA_V3) in the HCHS/SOL target population. Hypertension status is defined using the ACC/AHA criteria: if the systolic or diastolic BP is greater than or equal to 130/80 or if the participant self-reported during clinic visit as currently taking antihypertensive mediations.

The primary predictor of interest is BMI over time, which was measured during the three clinic visits (long-format: BMI; wide-format: BMI_V1, BMI_V2, BMI_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 (SEX), 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)

Formulaically, the main model of interest is:

$$g(E[Y_{it}|\text{covariates}]) = \beta_0 + \beta_1 \text{AGEGROUP\_C6}_i + \beta_2 \text{BKGRD1\_C7NOMISS}_i + \beta_3 \text{CENTERNUM}_i + \beta_4 \text{SEX}_i + \beta_5 \text{US\_BORN}_i + \beta_6 \text{EMPLOYED}_i + \beta_7 \text{EDUCATION\_C3}_i + \beta_8 \text{BMI}_{it} + \beta_9 \text{TIME}_{it},$$

where $g(\bullet)$ is the link function appropriate for the distribution of Y_{it} for participant i at time t (for covariates only at baseline, t is omitted). We assume **logit link** for the illustrations of binary outcomes in this chapter.

There are two interpretations for the coefficient β_8 for the time-varying BMI variable:

(1) WITHIN-PERSON: $\exp(\beta_8)$ represents the expected odds ratio for Y (hypertension status) at a given time t with a one-unit (1 kg/m²) increase in an individual's BMI. We adopt this

interpretation in this chapter because the focus is longitudinal. Since the comparison is within the same individual over time, all other covariates remain unchanged; therefore, it is not necessary to state "holding all other covariates constant."

(2) BETWEEN-PERSON: $\exp(\beta_8)$ represents the expected odds ratio for Y (hypertension status) at a given time t when comparing individuals whose baseline covariates are identical but whose BMI values differ by one unit.

The coefficient $\exp(\boldsymbol{\beta_9})$ for the TIME variable represents the expected within-person odds ratio for Y (hypertension status) with one year of aging assuming BMI remains unchanged. Note that all other covariates are exactly the same because the individual is being compared to themselves over time; therefore, it is not necessary to state "holding all other covariates constant."

5.1.2. Implementation of MI

Following the procedure in **Section 3.3.1**, use FCS/MICE to generate *10* imputed datasets from the wide-format analytic dataset sol_wide (N=16,415). Impute each variable (including outcome) with missing values using the following FCS regressions:

- Linear regression: BMI_V1, YRS_BTWN_V1V2, BMI_V2, YRS_BTWN_V1V3, BMI_V3
- Binary logistic regression: US_BORN, SEX, HYPERTENSION2_AHA_V1, HYPERTENSION2_AHA_V2, HYPERTENSION2_AHA_V3
- Ordered logistic regression (proportional odds): EMPLOYED, AGEGROUP_C6
- Multinomial (polytomous) logistic regression: EDUCATION_C3

Covariates without missing values but included in the main model (e.g., BKGRD1_C7NOMISS) are also specified in the MI process to preserve their associations with other variables. In addition, the imputation model includes the following design variables: center (CENTERNUM), the Visit 1 overall sampling weights (WEIGHT_FINAL_NORM_OVERALL), Visit 2 overall sampling weights (WEIGHT_NORM_OVERALL_V2), and Visit 3 overall sampling weights for clinic or home exams only (WEIGHT_NORM_OVERALL_EXAMONLY_V3).

5.1.3. Implementation of IPW

Following the procedure in **Section 3.3.2**, use FCS/MICE to generate *10* imputed datasets from the wide-format analytic dataset sol_wide (N=16,415). This will impute all variables with missing values and these will be used as predictors in the IPW models. The following FCS regression models are used:

- Linear regression: BMI_V1
- Binary logistic regression: US_BORN, SEX
- Ordered logistic regression (proportional odds): EMPLOYED, AGEGROUP_C6

- Multinomial (polytomous) logistic regression: EDUCATION_C3

Covariates without missing values but included in the main model (e.g., BKGRD1_C7NOMISS) are also specified in the MI process to preserve their associations with other variables. In addition, the imputation model includes the following design variables: center (CENTERNUM), and the Visit 1 overall sampling weights (WEIGHT_FINAL_NORM_OVERALL). The imputed covariate values are used solely for estimating the IPW and are not used in fitting the main analytic models.

Within each imputed dataset, fit a logistic regression model to estimate the probability of Visit 2 data contribution (PARTICIPANT_V2_NOMISS) among all participants. Data contribution at Visit 2 is defined as attending the visit (PARTICIPANT_V2 = 1) and having complete data on all baseline covariates used in the main model (AGEGROUP_C6, BKGRD1_C7NOMISS, CENTERNUM, SEX, US_BORN, EMPLOYED, EDUCATION_C3, BMI_V1), as well as the Visit 2 analytic variables HYPERTENSION2_AHA_V2, BMI_V2, and YRS_BTWN_V1V2. Consistent with the imputation model, the same set of baseline covariates are included as predictors. For each participant, average the fitted linear predictors across imputations and transform this pooled value to the probability scale to obtain predicted probability of Visit 2 data contribution (RR_V2).

Within each imputed dataset, fit a logistic regression model to estimate the probability of Visit 3 Exam-Only data contribution (PARTICIPANT_EXAMONLY_V3_NOMISS), using the same set of baseline covariates as in the Visit 2 model, plus Visit 2 participation status (PARTICIPANT_V2) to reflect the sequential nature of visit processes. Data contribution at Visit 3 (Exam-Only) is defined as completing the in-person exam (PARTICIPANT_EXAMONLY_V3 = 1) and having complete data on all baseline covariates used in the main model (AGEGROUP_C6, BKGRD1_C7NOMISS, CENTERNUM, SEX, US_BORN, EMPLOYED, EDUCATION_C3, BMI_V1), along with the Visit 3 analytic variables HYPERTENSION2_AHA_V3, BMI_V3, and YRS_BTWN_V1V3. Average the fitted linear predictors across imputations and transform them to the probability scale to obtain the predicted probability of Visit 3 data contribution (RR_V3).

As noted in **Output 3.4-2** and **Output 3.4-3**, all variables imputed in this illustrative example have less than 2% missingness.

Next, combine the pooled Visit 2 and Visit 3 response probabilities with the Visit 1 non-response-adjusted sampling weights (WEIGHT_NONRESP) to construct the Visit 2 and Visit 3 non-response-adjusted IPW weights.

For Visit 2, assign the non-response-adjusted IPW weight to participants who contributed data at Visit 2 (PARTICIPANT_V2_NOMISS = 1) as

$$\text{WEIGHT\_IPW\_V2} = \frac{\text{WEIGHT\_NONRESP}}{\text{RR\_V2}},$$

where WEIGHT_NONRESP is the Visit 1 non-response-adjusted sampling weight, and RR_V2 is the estimated probability of Visit 2 data contribution obtained from the GLM-based model.

For Visit 3 (Exam-Only), assign the non-response-adjusted IPW weight to participants who contributed data for the in-person exam (PARTICIPANT_EXAMONLY_V3_NOMISS = 1) as

$$\text{WEIGHT\_EXAMONLY\_IPW\_V3} = \frac{\text{WEIGHT\_NONRESP}}{\text{RR\_V3}},$$

where RR_V3 is the estimated probability of Visit 3 Exam-Only data contribution obtained from the GLM-based model.

5.2. Complex-Survey GEE

5.2.1. Visit 1 Sample, Visit-specific IPW

5.2.1.1. SUDAAN

Construction of IPW Indicators

```
data sol_wide;
  set sol_wide;

  array basecov AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX US_BORN
                EMPLOYED EDUCATION_C3 BMI_V1;
  if PARTICIPANT_V2 = 1 and
     nmiss(of basecov[*], HYPERTENSION2_AHA_V2, BMI_V2, YRS_BTWN_V1V2) = 0
then
    PARTICIPANT_V2_NOMISS = 1;
  else
    PARTICIPANT_V2_NOMISS = 0;

  if PARTICIPANT_EXAMONLY_V3 = 1 and
     nmiss(of basecov[*], HYPERTENSION2_AHA_V3, BMI_V3, YRS_BTWN_V1V3) = 0
then
    PARTICIPANT_EXAMONLY_V3_NOMISS = 1;
  else
    PARTICIPANT_EXAMONLY_V3_NOMISS = 0;
run;
```

In this **data** step, basecov collects all baseline covariates used in the main model of interest. The following indicators are created to serve as the outcomes in the IPW models:

PARTICIPANT_V2_NOMISS = 1 identifies participants who both attended Visit 2 (PARTICIPANT_V2 = 1) and have complete data on all baseline covariates plus HYPERTENSION2_AHA_V2, BMI_V2, and YRS_BTWN_V1V2. This defines Visit 2 data contribution.

PARTICIPANT_EXAMONLY_V3_NOMISS = 1 similarly identifies participants who both attended the Visit 3 Exam-Only component and have complete data on all baseline covariates plus HYPERTENSION2_AHA_V3, BMI_V3, and YRS_BTWN_V1V3. This defines Visit 3 (Exam-Only) data contribution.

MI for IPW Estimation

```
proc mi data=sol_wide nimpute=10 seed=2024 out=sol_mi_for_ipw;
  class AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX US_BORN
           EMPLOYED EDUCATION_C3;
  var   AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX
        US_BORN EMPLOYED EDUCATION_C3 BMI_V1
           WEIGHT_FINAL_NORM_OVERALL;
  fcs reg(BMI_V1);
  fcs logistic(US_BORN SEX EMPLOYED AGEGROUP_C6/* link=logit */);
  fcs logistic(EDUCATION_C3 / link=glogit);
run;
```

The procedure **proc mi** performs MI with nimpute=**10** generates 10 imputed datasets. 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_for_ipw, a single dataset that contains all the imputed data stacked, containing an imputation number identifier _IMPUTATION_ 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 and ordered logistic regression for ordinal variables; logistic specifying link=glogit, multinomial logistic regression for nominal variables. These imputed covariate values are used solely to estimate the IPW models.

Note: For details on FCS logistic regression methods in SAS, please refer to https://documentation.sas.com/doc/en/pgmsascdc/v_069/statug/statug_mi_details13.htm

Estimation of IPW Models by Visit

```
proc logistic data=sol_mi_for_ipw descending noprint;
  by _Imputation_;
  class AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX
        US_BORN EMPLOYED EDUCATION_C3 PARTICIPANT_V2_NOMISS;
  model PARTICIPANT_V2_NOMISS =
        AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX
        US_BORN EMPLOYED EDUCATION_C3 BMI_V1
            WEIGHT_FINAL_NORM_OVERALL;
  output out=pred_v2_imp(keep=_Imputation_ ID xb_v2) xbeta=xb_v2;
run;

proc means data=pred_v2_imp nway noprint;
  class ID;
  var xb_v2;
  output out=pred_v2_bar(drop=_type_ _freq_) mean=xb_v2_bar;
run;

proc logistic data=sol_mi_for_ipw descending noprint;
  by _Imputation_;
  class AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX
        US_BORN EMPLOYED EDUCATION_C3
        PARTICIPANT_V2
        PARTICIPANT_EXAMONLY_V3_NOMISS;
  model PARTICIPANT_EXAMONLY_V3_NOMISS =
        AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX
        US_BORN EMPLOYED EDUCATION_C3 BMI_V1
            WEIGHT_FINAL_NORM_OVERALL PARTICIPANT_V2;
  output out=pred_v3_imp(keep=_Imputation_ ID xb_v3) xbeta=xb_v3;
run;

proc means data=pred_v3_imp nway noprint;
  class ID;
  var xb_v3;
  output out=pred_v3_bar(drop=_type_ _freq_) mean=xb_v3_bar;
run;
```

To estimate Visit 2 data-contribution probabilities, a logistic regression model is fit using **proc logistic** within each imputed dataset (by _Imputation_) with outcome PARTICIPANT_V2_NOMISS and baseline predictors and the Visit 1 overall sampling weight (WEIGHT_FINAL_NORM_OVERALL). The linear predictor (xb_v2) is saved for each participant and then averaged across imputations within ID using **proc means**, yielding xb_v2_bar, the pooled logit of Visit 2 data contribution. Visit 3 Exam-Only data contribution is modeled analogously, with the outcome PARTICIPANT_EXAMONLY_V3_NOMISS and the same baseline covariates used in the Visit 2 model, additionally with Visit 2 participation (PARTICIPANT_V2) to reflect the sequential nature of study visits. The resulting linear predictors (xb_v3) are again averaged across imputations to produce xb_v3_bar, the pooled logit for Visit 3 data contribution.

Construction of IPW Weights by Visit

```
proc sort data=sol_wide;     by ID; run;
proc sort data=pred_v2_bar;  by ID; run;
proc sort data=pred_v3_bar;  by ID; run;

data sol_ipw_wide;
  merge sol_wide(in=a)
        pred_v2_bar(rename=(xb_v2_bar=lp_v2))
        pred_v3_bar(rename=(xb_v3_bar=lp_v3));
  by ID;
  if a;

  if not missing(lp_v2) then RR_V2 = 1/(1+exp(-lp_v2));
  if not missing(lp_v3) then RR_V3 = 1/(1+exp(-lp_v3));

  if PARTICIPANT_V2_NOMISS = 1 then WEIGHT_IPW_V2 = WEIGHT_NONRESP / RR_V2;
  else WEIGHT_IPW_V2 = .;

  if PARTICIPANT_EXAMONLY_V3_NOMISS = 1 then WEIGHT_EXAMONLY_IPW_V3 =
WEIGHT_NONRESP / RR_V3;
  else WEIGHT_EXAMONLY_IPW_V3 = .;
run;
```

In this **data** step, lp_v2 and lp_v3 are the pooled logits for Visits 2 and 3, respectively; these values are transformed using the inverse logit to obtain the predicted probabilities of visit-level data contribution (RR_V2 and RR_V3). These probabilities represent each participant's estimated likelihood of contributing complete analytic data at the corresponding visit.

For participants who contribute data at Visit 2 (PARTICIPANT_V2_NOMISS = 1), the non-response-adjusted IPW weight is computed as the Visit 1 non-response-adjusted sampling weight (WEIGHT_NONRESP) divided by RR_V2. For Visit 3 Exam-Only contributors, the corresponding non-response-adjusted IPW weight is WEIGHT_NONRESP divided by RR_V3. Participants with PARTICIPANT_V2_NOMISS = 0 or PARTICIPANT_EXAMONLY_V3_NOMISS = 0 have their visit-specific IPW weight variables set to missing, because they do not contribute analytic data at those visits and therefore are not included in the main analysis.

Creation of IPW Long-Format Dataset

```
data sol_ipw_long;
  set sol_ipw_wide;

  length VISIT 8 TIME 8 BMI 8 HYPERTENSION2_AHA 8 WEIGHT_IPW_BY_VISIT 8;

  VISIT = 1;
  HYPERTENSION2_AHA = HYPERTENSION2_AHA_V1;
  BMI   = BMI_V1;
  TIME  = 0;
  WEIGHT_IPW_BY_VISIT  = WEIGHT_FINAL_NORM_OVERALL;
  output;

  VISIT = 2;
  HYPERTENSION2_AHA = HYPERTENSION2_AHA_V2;
  BMI   = BMI_V2;
  TIME  = YRS_BTWN_V1V2;
  WEIGHT_IPW_BY_VISIT  = WEIGHT_IPW_V2;
  output;

  VISIT = 3;
  HYPERTENSION2_AHA = HYPERTENSION2_AHA_V3;
  BMI   = BMI_V3;
  TIME  = YRS_BTWN_V1V3;
  WEIGHT_IPW_BY_VISIT  = WEIGHT_EXAMONLY_IPW_V3;
  output;
run;
```

In this **data** step, the wide-format dataset is reshaped into a long-format structure by creating one record per participant for each clinic visit and assigning the appropriate visit identifier and time-varying variables. For sampling weights, the Visit 1 row carries the Visit 1 overall sampling weight (WEIGHT_FINAL_NORM_OVERALL), while the Visit 2 and Visit 3 rows use the newly constructed non-response-adjusted IPW weights (WEIGHT_IPW_V2 and WEIGHT_EXAMONLY_IPW_V3). This reshaping produces the final analytic dataset, sol_ipw_long, with one record per subject–visit and all visit-specific variables aligned for the main analysis.

Note that Visit 2 and Visit 3 non-response-adjusted IPW weights are set to be missing for participants who do not contribute data at those visits.

Variable	VISIT 1		VISIT 2		VISIT 3	
	# Observed	# Missing	# Observed	# Missing	# Observed	# Missing
HYPERTENSION2_AHA	16412	3	11620	4795	9087	7328
BMI	16343	72	11245	5170	8758	7657
TIME	16415	0	11623	4792	9864	6551
WEIGHT_IPW_BY_VISIT	16415	0	11056	5359	8615	7800

Design-Based GEE with IPW

```
data db_ipw_byvisit;
      set sol_ipw_long ;
      hh_id_num=input(substr(hh_id, 2),8.);
      if ^missing(weight_ipw_by_visit);
run;

proc rlogist data=db_ipw_byvisit filetype=sas r=independent semethod=zeger
      notsorted;
 nest strat hh_id_num;
  weight weight_ipw_by_visit;
  class agegroup_c6 bkgrd1_c7nomiss centernum sex us_born employed
education_c3;
  model hypertension2_aha=bmi agegroup_c6 bkgrd1_c7nomiss centernum sex
us_born
            employed education_c3 time;
  reflevel agegroup_c6=6 bkgrd1_c7nomiss=3 centernum=4 sex=0 us_born=0
            employed=1 education_c3=1;
  setenv labwidth=25 decwidth=3;
  print beta="Estimate" sebeta="(S.E)" t_beta="t value" p_beta="p-value";
run;
```

The **proc rlogist** procedure fits the logistic GEE regression model for the binary hypertension indicator (HYPERTENSION2_AHA) to the sol_ipw_long. It accounts for complex sampling using the nest statement, which handles stratification and clustering. The weight statement applies visit-specific non-response-adjusted IPW weights to produce population-representative estimates and accurate standard errors that account for the missingness. Any records with missing weights are not used in analysis, ensuring that the IPW-defined analysis is consistent with the GEE main analysis. The class statement specifies categorical predictors, while the reflevel statement sets the reference categories (e.g., agegroup_c6=6). The model statement employs the logistic link and includes BMI, demographic factors, and follow-up time (TIME) as covariates. An independent working correlation structure is assumed with r=independent, and robust standard errors are calculated using Zeger's sandwich variance estimator (semethod=zeger). The print statement displays parameter estimates, standard errors, Wald t-statistics, and p-values, and their corresponding labels.

Estimates

Output 5.2-1 displays the estimates. The estimated coefficient β_8 for BMI is 0.085 with a standard error of 0.007. This result indicates that 1 kg/m² increase in BMI within an individual is associated with a 9% increase in the odds of hypertension (odds ratio is exp(0.085)=1.09). This association is statistically significant ($p < .0001$).

The estimated coefficient β_9 for TIME is 0.096 with a standard error of 0.010. This result indicates that an additional year of aging is associated with a 10% increase in the odds of hypertension (odds ratio is exp(0.096)=1.10) given that their BMI value did not change during the additional year. This association is statistically significant ($p < .0001$).

Output 5.2-1

Independent Variables and Effects	Estimate	(S.E)	t value	p-value
Intercept	-1.833	0.291	-6.292	0.000
BMI	0.085	0.007	12.334	0.000
6-level Age Sub-groups				
1	-3.788	0.239	-15.880	0.000
2	-3.012	0.209	-14.400	0.000
3	-2.240	0.190	-11.815	0.000
4	-1.303	0.183	-7.111	0.000
5	-0.632	0.191	-3.305	0.001
6	0.000	0.000	.	.
BKGRD1_C7NOMISS				
0	0.383	0.206	1.857	0.063
1	-0.054	0.150	-0.362	0.717
2	0.333	0.172	1.939	0.053
3	0.000	0.000	.	.
4	0.431	0.188	2.294	0.022
5	-0.362	0.167	-2.170	0.030
6	0.915	0.361	2.536	0.011
Participant's Field Center - numeric				
1	0.130	0.202	0.644	0.519
2	0.036	0.117	0.304	0.761
3	0.609	0.168	3.626	0.000
4	0.000	0.000	.	.
Sex				
0	0.000	0.000	.	.
1	0.419	0.082	5.088	0.000
Born in mainland US (50 States + DC)				
0	0.000	0.000	.	.
1	-0.255	0.142	-1.798	0.072
Employment Status (includes retirees)				
1	0.000	0.000	.	.
2	0.024	0.178	0.137	0.891
3	-0.011	0.202	-0.056	0.956
4	-0.091	0.183	-0.497	0.619
Education Status (3 levels)				
1	0.000	0.000	.	.
2	-0.126	0.102	-1.231	0.219
3	-0.215	0.104	-2.068	0.039
TIME	0.096	0.010	9.450	0.000

5.2.2. Visit 3 Sample, Visit 3 IPW

5.2.2.1. SUDAAN

The steps for constructing the dataset sol_ipw_long are the same as those described in 5.2.1.1.

Design-Based GEE with IPW

```
data db_bin_ipw;
      set sol_ipw_long ;
      hh_id_num=input(substr(hh_id, 2),8.);
run;

* Call SUDAAN procedure;
proc rlogist data=db_bin_ipw filetype=sas r=independent semethod=zeger
      notsorted;
  nest strat hh_id_num;
  weight weight_examonly_ipw_v3;
  subpopn participant_examonly_v3_nomiss=1;
  class agegroup_c6 bkgrd1_c7nomiss centernum sex us_born employed
education_c3;
  model hypertension2_aha=bmi agegroup_c6 bkgrd1_c7nomiss centernum sex
us_born
            employed education_c3 time;
  reflevel agegroup_c6=6 bkgrd1_c7nomiss=3 centernum=4 sex=0 us_born=0
            employed=1 education_c3=1;
  setenv labwidth=25 decwidth=3;
  print beta="Estimate" sebeta="(S.E)" t_beta="t value" p_beta="p-value";
run;
```

The **proc rlogist** procedure in SUDAAN fits a logistic regression model to the dataset db_bin_ipw, which is created from sol_ipw_long dataset after converting the household identifier to a numeric format. The analysis accounts for the complex survey design by specifying nest strat hh_id_num, which identifies the stratification variable and household-level clustering. weight_examonly_ipw_v3 is included in the weight statement to apply the Visit 3 non-response-adjusted IPW weights. The subpopn statement restricts the analysis to participants who contribute data to Visit 3 (participant_examonly_v3_nomiss=1), ensuring consistency between the IPW-defined analysis and the GEE main analysis.

The class statement identifies the categorical covariates, and the reflevel statement explicitly sets the reference levels (e.g., agegroup_c6=6 sets level 6 as the reference group). The model statement specifies hypertension status (HYPERTENSION2_AHA) as the binary outcome and includes BMI, demographics, and time as predictors. The working correlation structure is specified as independent through r=independent, and semethod=zeger is specified to obtain a robust (sandwich) standard error. The setenv statement formats output display characteristics, and the print statement outputs parameter estimates, standard errors, test statistics, and p-values with their corresponding labels.

Estimates

Output 5.2-2 displays the estimates.

The estimated coefficient β_8 for BMI is 0.089 with a standard error of 0.006. This result indicates that 1 kg/m² increase in BMI within an individual is associated with a 9% increase in the odds of hypertension (odds ratio is exp(0.089)=1.09). This association is statistically significant ($p < .0001$).

The estimated coefficient β_9 for TIME is 0.087 with a standard error of 0.009. This result indicates that an additional year of aging is associated with a 9% increase in the odds of hypertension (odds ratio is exp(0.087)=1.09) given that their BMI value did not change during the additional year. This association is statistically significant ($p < .0001$).

Output 5.2-2

Independent Variables and Effects	Estimate	(S.E)	t value	p-value
Intercept	-1.563	0.277	-5.653	0.000
BMI	0.089	0.006	14.287	0.000
6-level Age Sub-groups				
1	-3.690	0.270	-13.672	0.000
2	-2.974	0.232	-12.846	0.000
3	-2.236	0.216	-10.354	0.000
4	-1.292	0.209	-6.170	0.000
5	-0.642	0.213	-3.008	0.003
6	0.000	0.000	.	.
BKGRD1_C7NOMISS				
0	0.539	0.186	2.892	0.004
1	0.048	0.151	0.317	0.751
2	0.404	0.168	2.402	0.016
3	0.000	0.000	.	.
4	0.533	0.179	2.983	0.003
5	-0.247	0.164	-1.505	0.132
6	0.689	0.217	3.170	0.002
Participant's Field Center - numeric				
1	-0.063	0.174	-0.363	0.717
2	-0.160	0.115	-1.398	0.162
3	0.483	0.163	2.953	0.003
4	0.000	0.000	.	.
Sex				
0	0.000	0.000	.	.
1	0.477	0.084	5.684	0.000
Born in mainland US (50 States + DC)				
0	0.000	0.000	.	.
1	-0.264	0.131	-2.022	0.043
Employment Status (includes retirees)				
1	0.000	0.000	.	.
2	-0.212	0.206	-1.030	0.303
3	-0.413	0.217	-1.905	0.057
4	-0.337	0.206	-1.636	0.102
Education Status (3 levels)				
1	0.000	0.000	.	.
2	-0.130	0.098	-1.329	0.184
3	-0.203	0.093	-2.190	0.029
TIME	0.087	0.009	9.441	0.000

5.3. Model-Based GEE

5.3.1. Visit 1 Sample, Visit-specific IPW

5.3.1.1. SAS

The steps for constructing the dataset sol_ipw_long are the same as those described in 5.2.1.1.

Model-Based GEE with IPW

```
proc genmod data=sol_ipw_long descending;
      class HH_ID AGEGROUP_C6(ref = '6') BKGRD1_C7NOMISS(ref = '3')
            CENTERNUM(ref = '4') SEX(ref = '0') US_BORN(ref = '0')
      EMPLOYED(ref = '1') EDUCATION_C3(ref = '1');
      weight WEIGHT_IPW_BY_VISIT;
      model HYPERTENSION2_AHA = AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX
            US_BORN EMPLOYED EDUCATION_C3 BMI TIME / dist=binomial;
      repeated subject = HH_ID / corr=ind;
run;
```

This **proc genmod** procedure fits GEE the long-format dataset sol_ipw_long. 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 WEIGHT_IPW_BY_VISIT which supplies the appropriate sampling weight for each visit. This variable equals the overall sampling weight for Visit 1 and the corresponding non-response-adjusted IPW weights for Visits 2 and 3. The model statement specifies HYPERTENSION2_AHA as the outcome and includes all covariates of interest, assuming a binomial distribution through dist=binomial. The repeated statement defines the clustering variable subject=HH_ID for household clusters. corr=ind specifies an independent working correlation structure. The descending option is specified so that the model targets the probability of HYPERTENSION2_AHA = 1, ensuring that regression coefficients are interpreted with respect to the odds of hypertension rather than non-hypertension. **proc genmod** automatically excludes records with missing weights, thereby ensuring consistency between the IPW-defined analysis and the GEE main analysis.

Estimates

Output 5.3-1 displays the estimates.

The estimated coefficient β_8 for BMI is 0.085 with a standard error of 0.007. This result indicates that 1 kg/m² increase in BMI within an individual is associated with a 9% increase in the odds of hypertension (odds ratio is exp(0.085)=1.09). This association is statistically significant ($p < .0001$).

The estimated coefficient β_9 for TIME is 0.096 with a standard error of 0.010. This result indicates that an additional year of aging is associated with a 10% increase in the odds of hypertension (odds ratio is exp(0.096)=1.10) given that their BMI value did not change during the additional year. This association is statistically significant ($p < .0001$).

Output 5.3-1

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-1.8325	0.2911	-2.4030	-1.2621	-6.30	<.0001
AGEGROUP_C6	1	-3.7875	0.2386	-4.2552	-3.3198	-15.87	<.0001
AGEGROUP_C6	2	-3.0122	0.2096	-3.4230	-2.6013	-14.37	<.0001
AGEGROUP_C6	3	-2.2396	0.1899	-2.6119	-1.8673	-11.79	<.0001
AGEGROUP_C6	4	-1.3032	0.1836	-1.6631	-0.9434	-7.10	<.0001
AGEGROUP_C6	5	-0.6320	0.1918	-1.0080	-0.2561	-3.30	0.0010
AGEGROUP_C6	6	0.0000	0.0000	0.0000	0.0000	.	.
BKGRD1_C7NOMISS	0	0.3829	0.2060	-0.0208	0.7867	1.86	0.0631
BKGRD1_C7NOMISS	1	-0.0541	0.1495	-0.3471	0.2389	-0.36	0.7174
BKGRD1_C7NOMISS	2	0.3333	0.1718	-0.0034	0.6700	1.94	0.0523
BKGRD1_C7NOMISS	4	0.4309	0.1877	0.0629	0.7988	2.30	0.0217
BKGRD1_C7NOMISS	5	-0.3616	0.1666	-0.6882	-0.0351	-2.17	0.0299
BKGRD1_C7NOMISS	6	0.9147	0.3603	0.2084	1.6209	2.54	0.0111
BKGRD1_C7NOMISS	3	0.0000	0.0000	0.0000	0.0000	.	.
CENTERNUM	1	0.1302	0.2017	-0.2651	0.5255	0.65	0.5187
CENTERNUM	2	0.0356	0.1173	-0.1943	0.2655	0.30	0.7615
CENTERNUM	3	0.6093	0.1682	0.2797	0.9389	3.62	0.0003
CENTERNUM	4	0.0000	0.0000	0.0000	0.0000	.	.
SEX	1	0.4186	0.0823	0.2573	0.5799	5.09	<.0001
SEX	0	0.0000	0.0000	0.0000	0.0000	.	.
US_BORN	1	-0.2552	0.1418	-0.5331	0.0227	-1.80	0.0719
US_BORN	0	0.0000	0.0000	0.0000	0.0000	.	.
EMPLOYED	2	0.0244	0.1779	-0.3242	0.3730	0.14	0.8909
EMPLOYED	3	-0.0112	0.2017	-0.4065	0.3840	-0.06	0.9556
EMPLOYED	4	-0.0912	0.1830	-0.4498	0.2675	-0.50	0.6184
EMPLOYED	1	0.0000	0.0000	0.0000	0.0000	.	.
EDUCATION_C3	2	-0.1259	0.1025	-0.3267	0.0750	-1.23	0.2193
EDUCATION_C3	3	-0.2146	0.1040	-0.4184	-0.0108	-2.06	0.0391
EDUCATION_C3	1	0.0000	0.0000	0.0000	0.0000	.	.
BMI		0.0849	0.0069	0.0714	0.0984	12.34	<.0001
TIME		0.0960	0.0102	0.0761	0.1159	9.44	<.0001

5.3.2. Visit 3 Sample, Visit 3 IPW

5.3.2.1. SAS

The steps for constructing the dataset sol_ipw_long are the same as those described in 5.2.1.1.

Model-Based GEE with IPW

```
proc genmod data= sol_ipw_long descending;
      class HH_ID AGEGROUP_C6(ref = '6') BKGRD1_C7NOMISS(ref = '3')
            CENTERNUM(ref = '4') SEX(ref = '0') US_BORN(ref = '0')
      EMPLOYED(ref = '1') EDUCATION_C3(ref = '1');
      weight WEIGHT_EXAMONLY_IPW_V3;
      model HYPERTENSION2_AHA = AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX
            US_BORN EMPLOYED EDUCATION_C3 BMI TIME / dist=binomial;
      repeated subject = HH_ID / corr=ind;
run;
```

The analysis restricted to participants who attended Visit 3. In this setting, the Visit 3 non-response-adjusted weight WEIGHT_EXAMONLY_IPW_V3 is available in sol_ipw_long. **proc genmod** automatically excludes records with missing weights, ensuring that only participants who contributed Visit 3 exam data are included in the estimation and that the IPW-defined analysis is consistent with the GEE main analysis. All other aspects of the modeling framework are identical to those described for the Visit 1 sample IPW analysis.

Estimates

Output 5.3-2 displays the estimates.

The estimated coefficient β_8 for BMI is 0.089 with a standard error of 0.006. This result indicates that 1 kg/m² increase in BMI within an individual is associated with a 9% increase in the odds of hypertension (odds ratio is exp(0.085)=1.09). This association is statistically significant ($p < .0001$).

The estimated coefficient β_9 for TIME is 0.087 with a standard error of 0.009. This result indicates that an additional year of aging is associated with a 9% increase in the odds of hypertension (odds ratio is exp(0.087)=1.09) given that their BMI value did not change during the additional year. This association is statistically significant ($p < .0001$).

Output 5.3-2

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-1.5630	0.2763	-2.1044	-1.0215	-5.66	<.0001
AGEGROUP_C6	1	-3.6902	0.2697	-4.2187	-3.1617	-13.68	<.0001
AGEGROUP_C6	2	-2.9740	0.2310	-3.4267	-2.5213	-12.88	<.0001
AGEGROUP_C6	3	-2.2355	0.2160	-2.6590	-1.8121	-10.35	<.0001
AGEGROUP_C6	4	-1.2918	0.2092	-1.7018	-0.8818	-6.18	<.0001
AGEGROUP_C6	5	-0.6421	0.2136	-1.0608	-0.2234	-3.01	0.0027
AGEGROUP_C6	6	0.0000	0.0000	0.0000	0.0000	.	.
BKGRD1_C7NOMISS	0	0.5387	0.1861	0.1739	0.9035	2.89	0.0038
BKGRD1_C7NOMISS	1	0.0479	0.1510	-0.2480	0.3438	0.32	0.7511
BKGRD1_C7NOMISS	2	0.4036	0.1678	0.0747	0.7326	2.40	0.0162
BKGRD1_C7NOMISS	4	0.5331	0.1787	0.1829	0.8833	2.98	0.0028
BKGRD1_C7NOMISS	5	-0.2473	0.1642	-0.5691	0.0744	-1.51	0.1319
BKGRD1_C7NOMISS	6	0.6885	0.2169	0.2635	1.1136	3.17	0.0015
BKGRD1_C7NOMISS	3	0.0000	0.0000	0.0000	0.0000	.	.
CENTERNUM	1	-0.0633	0.1742	-0.4047	0.2782	-0.36	0.7165
CENTERNUM	2	-0.1601	0.1147	-0.3848	0.0647	-1.40	0.1628
CENTERNUM	3	0.4825	0.1635	0.1620	0.8030	2.95	0.0032
CENTERNUM	4	0.0000	0.0000	0.0000	0.0000	.	.
SEX	1	0.4769	0.0839	0.3126	0.6413	5.69	<.0001
SEX	0	0.0000	0.0000	0.0000	0.0000	.	.
US_BORN	1	-0.2640	0.1305	-0.5198	-0.0082	-2.02	0.0431
US_BORN	0	0.0000	0.0000	0.0000	0.0000	.	.
EMPLOYED	2	-0.2122	0.2058	-0.6156	0.1912	-1.03	0.3026
EMPLOYED	3	-0.4129	0.2164	-0.8371	0.0114	-1.91	0.0565
EMPLOYED	4	-0.3374	0.2065	-0.7421	0.0674	-1.63	0.1023
EMPLOYED	1	0.0000	0.0000	0.0000	0.0000	.	.
EDUCATION_C3	2	-0.1303	0.0983	-0.3229	0.0623	-1.33	0.1849
EDUCATION_C3	3	-0.2031	0.0928	-0.3849	-0.0213	-2.19	0.0286
EDUCATION_C3	1	0.0000	0.0000	0.0000	0.0000	.	.
BMI		0.0890	0.0062	0.0768	0.1012	14.29	<.0001
TIME		0.0872	0.0092	0.0691	0.1053	9.46	<.0001

5.3.2.2. Stata

Construction of IPW Indicators

```
import sas using "sol_wide.sas7bdat", clear
set seed 2024

capture confirm variable WEIGHT_NORM_OVERALL_EXAMONLY_V3
if _rc==0 rename WEIGHT_NORM_OVERALL_EXAMONLY_V3 WEIGHT_EXAMONLY_V3


local basecov AGEGROUP_C6 BKGRD1_C7NOMISS CENTERNUM SEX US_BORN EMPLOYED
EDUCATION_C3 BMI_V1

egen miss_v2 = rowmiss(`basecov' HYPERTENSION2_AHA_V2 BMI_V2
YRS_BTWN_V1V2)
gen byte PARTICIPANT_V2_NOMISS = (PARTICIPANT_V2 == 1 & miss_v2 == 0)
drop miss_v2

egen miss_v3 = rowmiss(`basecov' HYPERTENSION2_AHA_V3 BMI_V3
YRS_BTWN_V1V3)
gen byte PARTICIPANT_EXAMONLY_V3_NOMISS = (PARTICIPANT_EXAMONLY_V3 == 1 &
miss_v3 == 0)
drop miss_v3
```

This section imports the wide-format analytic dataset and constructs visit-specific analytic contribution indicators for the binary-outcome IPW response models.

First, import sas using "sol_wide.sas7bdat", clear loads the SAS dataset into Stata and set seed 2024 sets a fixed random seed for reproducibility. It then checks for a potentially truncated Visit 3 exam-only weight variable name using capture confirm variable WEIGHT_NORM_OVERALL_EXAMONLY_V3; if present, it standardizes the name via rename ... WEIGHT_EXAMONLY_V3 so downstream code references a consistent weight variable.

Next, the baseline covariate set is stored in a macro (local basecov ...). For Visit 2, egen miss_v2 = rowmiss(...) counts missing values across the baseline covariates plus the binary outcome at Visit 2 (HYPERTENSION2_AHA_V2), Visit 2 BMI (BMI_V2), and elapsed time (YRS_BTWN_V1V2). The indicator PARTICIPANT_V2_NOMISS is then created with gen byte to equal 1 only for participants who attended Visit 2 (PARTICIPANT_V2 == 1) and have complete data on all required variables (miss_v2 == 0); otherwise it is 0. The temporary missingness counter is removed with drop.

The same logic is applied for Visit 3 exam-only participation: rowmiss() is computed across baseline covariates plus HYPERTENSION2_AHA_V3, BMI_V3, and YRS_BTWN_V1V3, and PARTICIPANT_EXAMONLY_V3_NOMISS is generated to equal 1 only for participants who completed the Visit 3 exam-only component (PARTICIPANT_EXAMONLY_V3 == 1) and have complete data on all required variables, with the temporary counter dropped afterward.

MI for IPW Estimation

```
mi set flong

mi register imputed AGEGROUP_C6 SEX BMI_V1 US_BORN EMPLOYED EDUCATION_C3
mi register passive BKGRD1_C7NOMISS CENTERNUM WEIGHT_FINAL_NORM_OVERALL

mi impute chained (regress) BMI_V1 (logit) US_BORN SEX (ologit) EMPLOYED
AGEGROUP_C6 (mlogit) EDUCATION_C3 = i.BKGRD1_C7NOMISS i.CENTERNUM
WEIGHT_FINAL_NORM_OVERALL, add(`num') rseed(2024)
```

This section performs MI of baseline covariates only, to support estimation of the response models used for IPW.

First, the MI framework is initialized in long format using mi set flong. Baseline variables subject to missingness are then declared with mi register imputed, including BMI_V1 and selected baseline categorical covariates (AGEGROUP_C6, SEX, US_BORN, EMPLOYED, EDUCATION_C3). Variables that are not directly imputed but are required in the imputation models are declared as passive via mi register passive, including background (BKGRD1_C7NOMISS), center (CENTERNUM), and the overall baseline sampling weight (WEIGHT_FINAL_NORM_OVERALL).

Next, mi impute chained is used to perform FCS MI, with imputation models chosen according to variable type: linear regression for continuous BMI_V1, logistic regression for binary variables (US_BORN, SEX), ordered logistic regression for ordinal variables (EMPLOYED, AGEGROUP_C6), and multinomial logistic regression for nominal EDUCATION_C3. All imputation models condition on background and center indicators and the overall baseline weight. The number of imputations is specified by add(num), and a fixed random seed (rseed(2024)) ensures reproducibility.

Estimation of IPW Models by Visit

```
tempfile base xb_v2_all xb_v3_all
save `base', replace

clear
set obs 0
gen ID = ""
gen double xb = .
gen int m = .
save `xb_v2_all', replace
save `xb_v3_all', replace

forvalues m = 1/`num' {

    use `base', clear
    mi extract `m', clear

    quietly logit PARTICIPANT_V2_NOMISS i.AGEGROUP_C6 i.BKGRD1_C7NOMISS
i.CENTERNUM i.SEX i.US_BORN i.EMPLOYED i.EDUCATION_C3 BMI_V1
WEIGHT_FINAL_NORM_OVERALL, nolog
    predict double xb_v2, xb

    keep ID xb_v2
    gen int m = `m'
    rename xb_v2 xb
    append using `xb_v2_all'
    save `xb_v2_all', replace

    use `base', clear
    mi extract `m', clear

    quietly logit PARTICIPANT_EXAMONLY_V3_NOMISS i.AGEGROUP_C6
i.BKGRD1_C7NOMISS i.CENTERNUM i.SEX i.US_BORN i.EMPLOYED i.EDUCATION_C3
BMI_V1 WEIGHT_FINAL_NORM_OVERALL i.PARTICIPANT_V2, nolog
    predict double xb_v3, xb

    keep ID xb_v3
    gen int m = `m'
    rename xb_v3 xb
    append using `xb_v3_all'
    save `xb_v3_all', replace
}

use `xb_v2_all', clear
collapse (mean) lp_v2=xb, by(ID)
gen double RR_V2 = invlogit(lp_v2)
tempfile rr2
save `rr2', replace

use `xb_v3_all', clear
collapse (mean) lp_v3=xb, by(ID)
gen double RR_V3 = invlogit(lp_v3)
tempfile rr3
save `rr3', replace
```

This section fits the visit-specific response (RR) models within each imputed dataset, saves the fitted linear predictors by participant, and then averages predictions across imputations to obtain participant-level response probabilities.

First, the MI dataset is saved to a temporary file (tempfile base and save \base) to serve as a clean starting point for repeated extraction. Two empty accumulation files (xb_v2_all and xb_v3_all) are then initialized and saved; these act as containers to store the fitted linear predictors (xb) across all imputations for Visit 2 and Visit 3, along with the imputation index m.

Next, the loop forvalues m = 1/num iterates over imputations. Within each iteration, mi extract m loads the m-th completed dataset. A logistic regression is fit for the Visit 2 response indicator PARTICIPANT_V2_NOMISS using baseline covariates, baseline BMI (BMI_V1), and the overall baseline weight (WEIGHT_FINAL_NORM_OVERALL) via quietly logit. The fitted linear predictor is generated using predict, then reduced to ID and the predictor, tagged with the imputation index (gen int m = ...), renamed to a common column (rename xb_v2 xb), appended into the cumulative file using append, and saved back to xb_v2_all.

The same process is repeated for Visit 3: the same imputation is reloaded via mi extract m, a logistic regression is fit for PARTICIPANT_EXAMONLY_V3_NOMISS with the same baseline covariates and WEIGHT_FINAL_NORM_OVERALL, additionally conditioning on prior Visit 2 participation (i.PARTICIPANT_V2), and the fitted linear predictor xb_v3 is generated, standardized to the common xb column, appended to xb_v3_all, and saved.

After looping over all imputations, collapse (mean) ... , by(ID) averages the stored linear predictors across imputations by participant, producing lp_v2 and lp_v3. These are transformed to response probabilities using invlogit() to create RR_V2 and RR_V3, which are saved as temporary datasets (rr2 and rr3) for downstream weight construction.

Construction of IPW Weights by Visit

```
use `base', clear
mi extract 0, clear

merge 1:1 ID using `rr2', nogen
merge 1:1 ID using `rr3', nogen

gen double WEIGHT_IPW_V2 = .
replace WEIGHT_IPW_V2 = WEIGHT_NONRESP / RR_V2 if PARTICIPANT_V2_NOMISS ==
1

gen double WEIGHT_EXAMONLY_IPW_V3 = .
replace WEIGHT_EXAMONLY_IPW_V3 = WEIGHT_NONRESP / RR_V3 if
PARTICIPANT_EXAMONLY_V3_NOMISS == 1
```

This section merges the imputation-averaged response probabilities back into the main wide dataset and then computes the visit-specific IPW weights.

First, it reloads the saved dataset (use \base) and extracts the observed (“m = 0”) version of the MI data using mi extract 0. The participant-level response probabilities are then merged in by ID using two one-to-one merges (merge 1:1 ID using `rr2' and merge 1:1 ID using `rr3'), adding RR_V2 and RR_V3 to the working dataset.

Next, the Visit 2 IPW weight WEIGHT_IPW_V2 is initialized to missing (gen double ... = .) and then filled in using replace as WEIGHT_NONRESP / RR_V2 for participants who contribute Visit 2 data (PARTICIPANT_V2_NOMISS == 1). All other records retain missing values for this weight.

Similarly, the Visit 3 exam-only IPW weight WEIGHT_EXAMONLY_IPW_V3 is generated and set to WEIGHT_NONRESP / RR_V3 for participants who contribute Visit 3 exam-only data (PARTICIPANT_EXAMONLY_V3_NOMISS == 1), leaving the weight missing otherwise.

Creation of IPW Long-Format Dataset

```
gen double TIME_1 = 0
gen double TIME_2 = YRS_BTWN_V1V2
gen double TIME_3 = YRS_BTWN_V1V3

gen double WBY_1 = WEIGHT_FINAL_NORM_OVERALL
gen double WBY_2 = WEIGHT_IPW_V2
gen double WBY_3 = WEIGHT_EXAMONLY_IPW_V3

rename HYPERTENSION2_AHA_V1 HYPERTENSION2_AHA_1
rename HYPERTENSION2_AHA_V2 HYPERTENSION2_AHA_2
rename HYPERTENSION2_AHA_V3 HYPERTENSION2_AHA_3

rename BMI_V1 BMI_1
rename BMI_V2 BMI_2
rename BMI_V3 BMI_3

reshape long HYPERTENSION2_AHA_ BMI_ TIME_ WBY_, i(ID) j(VISIT)

rename HYPERTENSION2_AHA_ HYPERTENSION2_AHA
rename BMI_ BMI
rename TIME_ TIME
rename WBY_ WEIGHT_IPW_BY_VISIT

sort ID VISIT
```

This section converts the wide-format dataset into a person-visit long format and consolidates visit-specific outcomes, time variables, and weights into single columns.

First, it creates visit-indexed time variables (TIME_1, TIME_2, TIME_3), setting baseline time to 0 and using the elapsed-time variables for Visits 2 and 3. It also creates visit-indexed weight variables (WBY_1, WBY_2, WBY_3), assigning the overall baseline sampling weight to Visit 1, the Visit 2 IPW weight to Visit 2, and the Visit 3 exam-only IPW weight to Visit 3.

Next, it harmonizes the naming of the visit-specific outcome and covariate variables by renaming HYPERTENSION2_AHA_V1/V2/V3 to HYPERTENSION2_AHA_1/2/3 and BMI_V1/V2/V3 to BMI_1/2/3, which matches the naming convention required for reshaping.

Then, reshape long HYPERTENSION2_AHA_ BMI_ TIME_ WBY_, i(ID) j(VISIT) stacks the dataset into long format, producing one record per participant per visit and creating the visit indicator VISIT. After reshaping, the suffix-based variable names are cleaned by renaming HYPERTENSION2_AHA_ to HYPERTENSION2_AHA, BMI_ to BMI, TIME_ to TIME, and WBY_ to WEIGHT_IPW_BY_VISIT.

Finally, sort ID VISIT orders the long-format dataset by participant and visit.

Model-Based GEE with IPW

```
encode ID, gen(ID_NUM)
xtset ID_NUM

xtgee HYPERTENSION2_AHA ib6.AGEGROUP_C6 ib6.BKGRD1_C7NOMISS ib4.CENTERNUM
ib0.SEX ib0.US_BORN ib1.EMPLOYED ib1.EDUCATION_C3 BMI TIME
[pw=WEIGHT_EXAMONLY_IPW_V3], family(binomial) link(logit)
corr(independent) vce(robust)
```

This section fits a weighted GEE for the binary outcome using the Visit 3 exam-only analytic sample with IPW weights.

First, encode ID, gen(ID_NUM) converts the participant identifier ID into a numeric panel identifier (ID_NUM), and xtset ID_NUM declares the panel structure so repeated measurements are indexed within participant.

Next, xtgee fits a logistic marginal mean model for HYPERTENSION2_AHA with baseline covariates (AGEGROUP_C6, BKGRD1_C7NOMISS, CENTERNUM, SEX, US_BORN, EMPLOYED, EDUCATION_C3) and the time-varying covariates BMI and TIME, with reference categories explicitly set using the ib#. notation.

IPW is applied through [pw=WEIGHT_EXAMONLY_IPW_V3]. Observations with missing weights are automatically excluded from model fitting. The model is fit with a binomial distribution and logit link (family(binomial) link(logit)), uses an independence working correlation structure (corr(independent)), and reports robust (sandwich) standard errors (vce(robust)).

Estimates

Output 5.3-3 displays the estimates.

The estimated coefficient β_8 for BMI is 0.089 with a standard error of 0.006. This result indicates that 1 kg/m² increase in BMI within an individual is associated with a 9% increase in the odds of hypertension (odds ratio is exp(0.089)=1.09). This association is statistically significant ($p < .0001$).

The estimated coefficient β_9 for TIME is 0.087 with a standard error of 0.009. This result indicates that an additional year of aging is associated with a 9% increase in the odds of hypertension (odds ratio is exp(0.087)=1.09) given that their BMI value did not change during the additional year. This association is statistically significant ($p < .0001$).

Output 5.3-3

HYPERTENSIO~A	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
AGEGROUP_C6						
1	-3.690138	.2741733	-13.46	0.000	-4.227507	-3.152768
2	-2.974383	.2392456	-12.43	0.000	-3.443296	-2.50547
3	-2.23548	.2214061	-10.10	0.000	-2.669429	-1.801532
4	-1.29202	.21494	-6.01	0.000	-1.713295	-.8707454
5	-.6420419	.2158128	-2.97	0.003	-1.065027	-.2190567
BKGRD1_C7NO~S						
0	-.1490363	.2301649	-0.65	0.517	-.6001513	.3020787
1	-.639714	.2302474	-2.78	0.005	-1.090991	-.1884375
2	-.2844499	.2320078	-1.23	0.220	-.7391769	.1702771
3	-.6881511	.2156985	-3.19	0.001	-1.110912	-.2653898
4	-.1542704	.2320935	-0.66	0.506	-.6091652	.3006245
5	-.9350603	.2330002	-4.01	0.000	-1.391732	-.4783884
CENTERNUM						
1	-.0639605	.1743519	-0.37	0.714	-.4056839	.2777629
2	-.1602173	.112687	-1.42	0.155	-.3810798	.0606452
3	.4826563	.1628923	2.96	0.003	.1633934	.8019193
1.SEX	.4774915	.0829646	5.76	0.000	.3148839	.6400991
1.US_BORN	-.2637035	.1307168	-2.02	0.044	-.5199038	-.0075033
EMPLOYED						
2	-.2121058	.2018781	-1.05	0.293	-.6077797	.183568
3	-.4124736	.2128121	-1.94	0.053	-.8295777	.0046306
4	-.3372325	.2026785	-1.66	0.096	-.734475	.06001
EDUCATION_C3						
2	-.1306387	.0992781	-1.32	0.188	-.3252203	.0639429
3	-.2036009	.0932516	-2.18	0.029	-.3863707	-.0208311
BMI	.0890197	.0061797	14.41	0.000	.0769076	.1011317
TIME	.0871669	.0091814	9.49	0.000	.0691716	.1051622
_cons	-.8744172	.3345707	-2.61	0.009	-1.530164	-.2186707

5.4. Results Summary

In **Table 5.4-1**, we summarize the key results from the illustrative example in this chapter, organized by analysis procedure, analytic sample, approach for addressing missing visits, and software implementation. The estimates from the various methods are similar as expected.

Table 5.4-1

Analysis Procedure	Analytic Sample	Missing-Visit Strategy	Software	Section	BMI Estimate (SE)	TIME Estimate (SE)
Complex-survey GEE	Visit 1 Sample	Visit-specific IPW	SUDAAN	5.2.1.1	0.085 (0.007)	0.096 (0.010)
	Visit 3 Sample	Visit 3 IPW	SUDAAN	5.2.2.1	0.089 (0.006)	0.087 (0.009)
Model-based GEE	Visit 1 Sample	Visit-specific IPW	SAS	5.3.1.1	0.085 (0.007)	0.096 (0.010)
	Visit 3 Sample	Visit 3 IPW	SAS	5.3.2.1	0.089 (0.006)	0.087 (0.009)
			Stata	5.3.2.2	0.089 (0.006)	0.087 (0.009)

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

Lohr, Sharon L. Sampling: Design and Analysis, Third Edition. CRC Press, 2022.

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