UNIVERSITY OF CALIFORNIA SAN DIEGO

Parametric and Semiparametric Models for Causal Inference, Longitudinal Data Analysis and Brain Imaging Data with Measurement Errors, Missing Data and Small Samples

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Biostatistics

by

Chenyu Liu

Committee in charge:

Professor Xin Tu, Chair
Professor Lin Liu, Co-Chair
Professor Loki Natarajan
Professor Wesley K Thompson
Professor Yuqi Qiu

2025

The Dissertation of Chenyu Liu is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2025

DEDICATION

To my parents: for your genuine love and unwavering support through all these years.

EPIGRAPH

In the sciences, we are now uniquely privileged to sit side
by side with the giants on whose shoulders we stand.
– *Gerald Holton*

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to take this opportunity to extend my thanks to the University of California, San Diego (UCSD) for providing an exceptional environment that has fostered both my academic and personal growth. The resources, facilities, and support at UCSD have been instrumental in the completion of this dissertation. I am especially thankful for the division of Biostatistics, whose rigorous program and dedicated faculty have greatly contributed to my development as a researcher. I am sincerely grateful to everyone I had the chance to connect with, collaborate with, and learn from during my time at UCSD.

First and foremost, I want to convey my deepest gratitude and admiration to my advisor, my committee chair, Dr. Xin Tu, for his unwavering support, guidance, and encouragement throughout the entirety of this five years. Under Dr. Tu's guidance, I have engaged in numerous captivating collaborative projects spanning diverse domains. Through these ventures, I've learned the art of effective collaboration with individuals from different backgrounds. These experiences have granted me direct insight into the immense motivation and satisfaction derived from applying biostatistical methods to solve real-world challenges. As an advisor, he consistently steered me towards discovering my research interests and inspired me to explore innovative methods. His dedication was evident through the numerous hours spent carefully reviewing my research manuscripts, offering constructive feedback, and collaboratively creating strategies to enhance the quality of my research. Dr. Tu has consistently shown support and respect the decisions I have made regarding my career path. His expertise, patience, and insightful feedback have significantly contributed to the development of my five-year PhD study.

I want to express my sincere gratitude to Professor Loki Natarajan for her exceptional support. Throughout more than two years, she offered invaluable help during my GSR work and played a crucial role in guiding a project on missing data in the PEP4PA study, which forms a significant part of this dissertation as a major collaborative research effort. At the beginning of my PhD journey, I worked as Professor Natarajan's teaching assistant, learning directly from her. Her mentorship not only enhanced my academic journey but also motivated me to pursue

excellence in my work. Her feedback, patience, and encouragement have been immeasurable, and I am truly grateful for the time and effort she invested in me.

I am immensely grateful to Professor Wesley Thompson for his unwavering support and guidance throughout my academic journey. His introduction to the fascinating field of human brain imaging research not only sparked a topic in this dissertation but also serves as a starting point for future explorations and discoveries. Professor Thompson's consistent encouragement, insightful feedback, and mentorship have been instrumental in helping me navigate challenges and grow as a researcher. His guidance has empowered me to actively engage with complex concepts, think critically about research methodologies, and confidently pose meaningful questions in diverse research work.

Additionally, I want to give many thanks to my committee members, Dr. Lin Liu and Dr. Yuqi Qiu. Their insightful feedback, constructive criticism, and valuable suggestions have played a pivotal role in enhancing the quality and depth of this work. Their expertise and rigorous evaluation have contributed immensely to refining the research methodology, strengthening the analysis, and ensuring the accuracy of the conclusions. I am truly appreciative of their dedicated involvement and contributions, which have enriched this dissertation and my overall academic journey.

Moreover, I would like to express my thanks for the Biostatistics program at UCSD. My journey in this program has been extremely rewarding. The UCSD Ph.D. program in biostatistics not only equipped me with statistics knowledge and programming skills but also instilled in me the ability to apply statistical methods in collaborative scientific endeavors. These acquired skills will undoubtedly serve as the cornerstone of my future scientific pursuits. I extend my sincere gratitude to Professors Armin Schwartzman, Florin Vaida, Karen Messer, Sonia Jain, Jingjing Zou, and Lily Xu for their guidance and instruction, which have enhanced my statistical knowledge level and elevated my understanding of statistics. I also wish to extend my profound thanks to Stella Tripp and Sarah Dauchez, whose contributions have been critical in the rapid development of our program.

Finally, I am deeply grateful to my family, my mother Hongli Xu, my father Guoying Liu, my boyfriend Mingxuan Han, my grandmother Yingrong Gao, my grandfather Shengfeng Xu, and the entire big family of mine, for their boundless love, encouragement, and understanding during this challenging journey. Their strong support has been my anchor, sustaining me through the highs and lows of my five-year PhD journey.

Chapter 2, in full, has been submitted for publication as "Chenyu Liu, Bohan Xu, Chun Fan, Wesley Thompson. *Estimating the Total Fraction of Variance Explained from Noisy Brain Imaging Data*". The dissertation author was the primary author on this paper.

Chapter 3, in full, is currently being prepared for submission for publication of the material as it may appear in "Chenyu Liu, Jinyuan Liu, Tuo Lin, Yuqi Qiu, Lin Liu and Xin Tu. *Sign-flipped Doubly Robust Score-like Test for Average Treatment Effect in Causal Inference*". The dissertation author was the primary author on this paper.

Chapter 4, in full, is currently being prepared for submission for publication of the material as it may appear in "Chenyu Liu, Loki Natarajan, Lin Liu, Yuqi Qiu, Xin Tu. *A Multiple Imputation Approach for Longitudinal Regression with Missing Response*". The dissertation author was the primary author on this paper.

VITA

2016	Bachelor of Science in Applied Mathematics. Hebei University, China
2019	Master of Science in Statistics. University of Minnesota, Twin Cities, U.S.
2019-2024	Graduate Student Researcher. University of California San Diego, U.S
2024	Doctor of Philosophy in Biostatistics. University of California San Diego, U.S

PUBLICATIONS

Liu, C., Zhang, X., Nguyen, T. T., Liu, J., Wu, T., Lee, E., & Tu, X. M. (2022). Partial least squares regression and principal component analysis: similarity and differences between two popular variable reduction approaches. General psychiatry, 35(1).

Tam, R. M., Zablocki, R. W., **Liu, C.**, Narayan, H. K., Natarajan, L., LaCroix, A. Z., Dillon, L., Sakoulas, E., & Hartman, S. J. (2023). Feasibility of a Health Coach Intervention to Reduce Sitting Time and Improve Physical Functioning Among Breast Cancer Survivors: Pilot Intervention Study. JMIR cancer, 9, e49934.

Crist, K., Full, K. M., Linke, S., Tuz-Zahra, F., Bolling, K., Lewars, B., **Liu, C.**, Shi, Y., Rosenberg, D., Jankowska, M., Benmarhnia, T., & Natarajan, L. (2022). Health effects and cost-effectiveness of a multilevel physical activity intervention in low-income older adults; results from the PEP4PA cluster randomized controlled trial. The international journal of behavioral nutrition and physical activity, 19(1), 75.

Adamowicz, David H., Paul D. Shilling, Barton W. Palmer, Tanya T. Nguyen, Eric Wang, **Chenyu Liu**, Xin Tu, Dilip V. Jeste, Michael R. Irwin, and Ellen E. Lee. "Associations between inflammatory marker profiles and neurocognitive functioning in people with schizophrenia and non-psychiatric comparison subjects." Journal of Psychiatric Research 149 (2022): 106-113.

Liu, Jinyuan; Zhang, Xinlian; Lin, Tuo; Chen, Ruohui; Zhong, Yuan; Chen, Tian ; Wu, Tsungchin; **Liu, Chenyu**; Huang, Anna; Nguyen, Tanya; Lee, Ellen; Jeste, Dilip; Tu, Xin. "A New Paradigm for High-dimensional Data: Distance-Based Semiparametric Feature Aggregation Framework via Between-Subject Attributes." Scandinavian Journal of Statistics (2023).

Chen, Ruohui, Tuo Lin, Lin Liu, Jinyuan Liu, Ruifeng Chen, Jingjing Zou, **Chenyu Liu**, et al. "A Doubly Robust Estimator for the Mann Whitney Wilcoxon Rank Sum Test When

Applied for Causal Inference in Observational Studies.” Journal of Applied Statistics, (2024), 1–25.

ABSTRACT OF THE DISSERTATION

Parametric and Semiparametric Models for Causal Inference, Longitudinal Data Analysis and Brain Imaging Data with Measurement Errors, Missing Data and Small Samples

by

Chenyu Liu

Doctor of Philosophy in Biostatistics

University of California San Diego, 2025

Professor Xin Tu, Chair
Professor Lin Liu, Co-Chair

Accurate statistical inference with complex, noisy data and limited sample sizes requires methods capable of addressing measurement errors, missingness, and model misspecification. This dissertation develops methodological tools for such challenges, grounded in rigorous theory and simulation studies. It comprises three studies.

The first study focuses on estimating the total variance explained by imaging-derived phenotypes (IDPs) in brain imaging data affected by measurement error. We propose a Bayesian model using Hamiltonian Monte Carlo (HMC) to correct for errors in predictors and to estimate the variance explained as if IDPs were measured without noise. Simulation studies confirm

the method's validity. Using resting-state fMRI data from the Adolescent Brain Cognitive Development (ABCD) Study, we train a noisy Best Linear Unbiased Predictor (BLUP) model and show that accounting for measurement error yields substantially higher and more realistic estimates of variance explained than models that ignore noise in IDPs.

The second study addresses hypothesis testing for average treatment effects (ATE) in causal inference. While outcome regression models, marginal structural models, and doubly robust estimators (DRE) have appealing asymptotic properties, they may yield biased inference in small samples, with anticonservative type I error or poor coverage, even when incorporating recent nonparametric extensions. To overcome these limitations, we develop a sign-flipping score-like test. Although score and score-like tests already outperform Wald tests for modest sample sizes, they can still behave anticonservatively, especially with heteroscedastic continuous outcomes or over-dispersed counts. Our proposed sign-flipping test preserves favorable asymptotic behavior while substantially improving finite-sample performance, as demonstrated by extensive simulations.

Motivated by the Peer Empowerment Program 4 Physical Activity (PEP4PA) study, the third study investigates multiple imputation (MI) for longitudinal data with missing responses. Despite MI's widespread use, many implementations lack established asymptotic guarantees, and Rubin's variance estimator can fail under certain conditions. We distinguish two MI types: type B, based on sampling from a posited regression model with consistent parameter estimates, and type A, which samples from the predictive distribution by integrating over parameter uncertainty. Because type B is difficult to implement generally, we propose a type A MI estimator using standard software such as the MICE package in R. Our approach accommodates missingness depending on variables outside the posited model and yields asymptotically valid inference with Rubin's variance formula. Simulations and application to PEP4PA data demonstrate its practical validity and ease of implementation.

Chapter 1

Introduction

In modern statistical research, dealing with complex and imperfect data is a ubiquitous challenge. In this thesis, we propose novel statistical methods to address such challenges across various domains. Specifically, we focus statistical methodologies for brain imaging data with measurement errors in explanatory variables, robust hypothesis testing for average treatment effect in causal inference, and multiple imputation approaches for longitudinal regression analysis with missing responses (dependent variables). These methods are crucial for improving the accuracy and reliability of statistical analyses in their respective fields.

Causal inference is a fundamental aspect of statistical analysis and research, focused on determining the cause-and-effect relationships between variables. Unlike associations or correlations, causal inference aims to understand how changes in one variable (the cause) directly influence another variable (the effect). This is crucial for making decisions in various fields, including medicine, economics, social sciences, and policy-making. Several approaches exist for inferring causality from observational data. Outcome regression models (ORM) estimate causal effects by controlling for other influencing variables, or confounders, using various models, such as parametric, semiparametric generalized linear and non-linear models. Despite its wide application, it faces challenges such as incorrect specification of the confounders, or covariates, in the model. Marginal structural models (MSM) mitigate confounding using propensity scores. In addition to misspecification of covariates, the inverse-propensity-score weighted estimators

also do not perform well with anticonservative bias in type I errors, even moderate samples and under correctly specified covariates in the model. The doubly robust estimator (DRE) integrates the ORM and MSM to offer consistency if only one of the models has a correct specification of covariates in the model, thus doubling the chance of correct inference, though it still struggles with small and moderate samples. Hemerik et al. (2020) proposed a sign-flipping score test for parametric generalized linear models and demonstrated that it was quite effective to control anticonservative bias for small to moderate samples, especially in the presence of overdispersion and heteroscedasticity for the response. This approach seems promising and adaptable to the context of causal inference with DRE to address the anticonservative bias.

Issues like Functional magnetic resonance imaging (fMRI) is a non-invasive neuroimaging technique used to measure and map brain activity. By detecting changes in blood flow, fMRI provides insights into the functional anatomy of the brain, allowing researchers to see which areas are involved in various cognitive and motor tasks. Research initiatives like the multi-site Adolescent Brain Cognitive Development (ABCD) Study have substantially augmented sample sizes in contrast to earlier neurodevelopmental cohorts to study. There's a realization that the effect sizes of brain-behavior relationships utilizing individual imaging-derived phenotypes (IDPs) are generally quite small. Furthermore, while the effects individually seem modest, they are dispersed widely across multiple IDPs rather than being concentrated on a few specific ones in many brain-behavior relationships. Consequently, there is a need for analytical methods capable of comprehensively assessing the overall variance explained by sets of IDPs without assuming sparsity. Genome-Wide Complex Trait Analysis (GCTA) is a commonly used tool for evaluating the total variance explained by extensive sets of genetic loci in genetics. Although there are notable parallels between genome-wide and brain-wide analyses, a key distinction lies in the fact that IDPs derived from magnetic resonance imaging (MRI) data are subject to considerably more noise compared to genetic data. This disparity may offer insight into why brain-behavior effect sizes observed in large-scale studies tend to be small.

Missing data is a prevalent issue in clinical research, often arising from loss to follow-

up, intercurrent events, or data collection errors, which inevitably result in not all intended measurements being obtained for all participants. This can lead to several problems, as it necessitates untestable assumptions about the effects of the unobserved data on analysis results. Improper handling of missing data can introduce bias into estimates, potentially leading to incorrect conclusions. Additionally, missing data reduce the available sample size, decreasing the statistical power of the analysis and leading to less precise estimates. This complicates both the analysis and interpretation of results. There are three primary mechanisms by which data can be missing: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). Clinical trial analyses typically assume MAR, where the probability of missingness is related to observed data but not the unobserved data. MCAR, where missingness is entirely random, is less defendable, while MNAR is difficult to address, as missingness is related to unobserved data itself and external information is needed to model the impact of the missing data on analysis results. A widely used approach for addressing MAR and MCAR is multiple imputation (MI), which involves filling in missing values by generating plausible numbers based on the distributions and relationships among observed variables. MI explicitly incorporates the uncertainty about the true value through multiply imputed data for valid inference. We develop innovative and robust statistical methods to handle complex and imperfect datasets. The subsequent chapters address specific research goals inspired by the identified gaps and limitations in our study. First, we extend the existing GCTA model by incorporating a more general assumption that IDPs derived from MRI data are measured with significant noise. Second, to improve the performance of the DREfor small or moderate samples, we propose a novel sign-flipping score-like test to reduce anticonservative bias in inference about average treatment effects. Last but not least, we develop a new MI approach to utilize existing software such as the MICE package in R to facilitate longitudinal data analysis with missing data under MCAR and MAR.

1.1 Aims and Organization of this Dissertation

In Chapter two, we develop the best linear unbiased predictor for noisy brain data. We propose a Bayesian methodology with the application of Hamiltonian Monte Carlo Algorithm (HMC) to correct for measurement errors in predictors and estimate the total fraction of variance explained by sets of imaging-derived phenotypes. We illustrate our proposed method with real data from the Adolescent Brain Cognitive Development (ABCD) Study, using resting-state fMRI data from Release 4.0.

In Chapter three, we explore several hypothesis testing methods for average treatment fffects (ATE) in causal inference. We provide an overview of existing methods such as outcome regression model (ORM), marginal structural models (MSM), and doubly robust estimators (DRE) and propose a novel sign-flipping score-like test for DRE to imporve its performance for small or moderate samples when applied for causal inference about ATE. Simulation study is employed to validate the theoretical characteristics of the proposed sign-flipping score-like test and provide comparisons with existing methods.

Chapter four conducts a study investigating multiple imputation (MI) approaches for longitudinal regression analysis with missing responses (dependent variables) under MAR and MCAR. By leveraging the work of Tsiatis (2006), we propose a novel MI estimator by utilizing popular software for missing imputation such as the MICE package in R for the imputation step. The proposed estimator is shown to have consistent and asymptotically normal properties. In addition, we show that Rubin's variance estimator is consistent when used to estimate the asymptotic variance of the propose MI estimator. We illustrate the approach with both real and simulated data. The simulation stuidy evaluates the performance of the proposed estimator, while the real study allows us to investigate whether the intervention group in a study on improving health with physical activity has significantly increased the minutes of moderate-vigorous-physical activity during the period of visits after baseline visit. Chapter five finally summarizes the overall findings and discusses the future studies.

Chapter 2

Estimating the Total Fraction of Variance Explained from Noisy Brain Imaging Data

2.1 Abstract

Due to the advent of large neurodevelopmental datasets such as the Adolescent Brain Cognitive Development (ABCD) Study, it has become increasingly clear that brain-behavior effect sizes using individual imaging-derived phenotypes (IDPs) are generally quite small. It is also appears to be the case that, while individually small, effects are widely distributed rather than concentrated on just a few IDPs for many brain-behavior relationships. Thus, analytical methods are needed for assessing the total variance explained by sets of IDPs *en masse* using methods that do not assume sparsity. In the field of genetics, Genome-Wide Complex Trait Analysis (GCTA) is commonly used to assess the total variance explained by large sets of genetic loci. While there are clear analogies between genome-wide and brain-wide analyses, a major difference is that IDPs obtained from magnetic resonance imaging (MRI) data are measured with considerably more noise than are genetic data, thus perhaps partially explaining why brain-behavior effect sizes observed in large studies are generally small. Here, we propose a GCTA-like model for assessing the total variance explained by sets of IDPs applicable to noisy brain imaging data. We develop a Bayesian implementation of this model using Hamiltonian Monte Carlo (HMC) algorithm for posterior inferences, which correct for measurement error in the predictors and estimate the total fraction of variance explained if IDPs were measured without noise. We

validate the noisy BLUP model in simulation studies. In real data analyses, we train the noisy BLUP model on resting state functional magnetic resonance imaging (rsfMRI) data from the ABCD Study, showing that the trained model provides substantially higher estimates of the total fraction of variance explained than that obtained by incorrectly assuming IDPs are measured without noise.

2.2 Introduction

Functional Magnetic Resonance Imaging (fMRI) measures changes in blood oxygenation and blood flow related to neuronal activity, providing researchers with the means to study human brain function *in vivo* [1]. fMRI has become an essential and widely-used tool for assessing brain-behavior relationships and neurodevelopmental outcomes. However, the explosive growth in the number of published fMRI studies has led to something of a replication crisis [2], wherein initial published effects are greatly attenuated or completely absent in follow-up studies. This crisis has likely been caused by a combination factors, including high-dimensional data, incentives for finding "signficant" associations, publication bias and, importantly, small sample sizes [3].

Studies such as the UK Biobank (UKB, [4]), the Human Connectome Project (HCP) [5], the Adolescent Brain Cognitive Development (ABCD) Study [6], and the upcoming HEALthy Brain Child Development (HBCD) Study [7] have greatly increased sample sizes ($n \approx$ 1k to 12k) compared to previous neurodevelopmental cohorts. An important consequence of larger samples is the ability to estimate brain-behavior effect sizes with more power and precision, and hence less prone to shrinkage due to the "Winner's Curse" [8]. In particular, recent results from the ABCD Study demonstrate that effect sizes of brain-behavior relationships are considerably smaller than have been published in the past [3, 9] from smaller, likely under-powered datasets.

Another realization arising from the the greater power and precision of large-scale neurodevelopmental studies is that brain-behavior associations tend to be widely distributed across the brain. For example, Zhao et al. [10] found that whole-brain ridge regression of all cortical vertices outperformed methods (e.g., LASSO) that assumed sparsity for predicting cognitive ability in the Adolescent Brain Cognitive Development (ABCD) Study using the N-back task MRI data. The presence of individually small but widely-distributed effects argues for using statistical methods that integrate associations across the brain (e.g., all networks, vertices or voxels) to assess the total variance in behavioral outcomes explained by brain morphometry or activity.

This is precisely the scenario that Genome-Wide Association Studies (GWAS) encountered over the last two decades. Under-powered candidate gene studies failed to replicate, effect sizes of individual loci were much smaller than researchers had hoped, but they were widely distributed across the genome [11]. For most human complex diseases and traits, loci identified by GWAS explain only a tiny fraction of the overall heritability. In response to this, Yang et al. [12] proposed Genome-wide Complex Trait Analysis (GCTA) for GWAS data, a method for assessing the total fraction of variance explained (FVE) for all loci in a GWAS, which can be applied to data when the number of loci far exceeds the number of participants. Other FVE methods include LD-Score Regression [13] and GWASH [14]. Using these and other methods, genetics researchers have determined that common variants tagged by microarrays (so-called "SNP heritability") account for a substantial proportion of total heritability, e.g., as assessed by family and twin studies, for many complex traits [15, 16]. This is an important observation, because it impacts how researchers interpret GWAS results and how important additive effects of common variants are as drivers of complex traits and diseases [17], compared, e.g., to rare mutations or structural variation.

The gap between the proportion of variance attributable to genetics from family studies and the proportion from microarray data is sometimes called the "missing heritability" problem [18]. While not usually framed in the same way, there is a similar issue in neurobehavioral studies. It is widely accepted among neuroscientists that all cognitive behaviors are ultimately a product of brain activity. However, to date fMRI studies have only explained a small proportion of variation in most measured cognitive behavioral responses in healthy human beings. This could be because the neural activity captured by a particular set of fMRI features is not directly relevant for the behavioral phenotype of interest. It could also be because the BOLD response is an indirect measure that only weakly captures the underlying neurobiology relevant for behavioral responses. Finally, it could also be due to measurement error in both the fMRI BOLD signal and in the behavioral responses themselves: if these sources of measurement error are uncorrelated (i.e., *non-differential*) they will tend to attenuate association effect sizes [19]. Here, we focus on

the issue of attenuation of effect sizes due to non-differential measurement error. The GCTA model assumes that genetic features are measured without noise. In contrast, fMRI whole-brain data is measured with considerable noise, e.g. intra-class correlations (ICCs) rarely exceed 0.5 in the ABCD fMRI data [20].

In the low-dimensional setting, there are existing methods addressing attenuation due to measurement error. Cook and Stefanski [21] proposed a simulation-extrapolation estimation (SIMEX) in parametric measurement error models. SIMEX is a simulation-based method of inference for parametric models with measurement error in which the measurement error variance is known or at least well estimated. The method added additional measurement error in known increments to the data, computing estimates from the contaminated data, establishing a trend between these estimates and the variance of the added errors, and extrapolating this trend back to the case of no measurement error. While in our case, the measurement error is unknown and it is difficult to estimate the measurement error. Wang et al.[22] proposed a new class of models, generalized linear mixed measurement error models, which model the correlation and the measurement error simultaneously, considering generalized linear mixed models for clustered data when one of the predictor with measurement error. They explored the model from two directions: bias analysis and functional inference using the SIMEX method. However, this method assumed that the data are obtained from independent clusters and the paper was more interested in computing the biases that result in common GLMMs when one ignores measurement error. Dellaportas and Stephens [23] adopted a Bayesian approach, analysing errors-in-variables models in full generality under a Bayesian formulation that makes inferences about the coefficients in errors-in-variables regression models.

However, no existing methods address measurement error in the context of the fraction of variance explained by a large set of features *en masse*. In this paper, we propose a Bayesian model to correct for measurement error in predictors when estimating the total fraction of variance explained (FVE) by sets of imaging-derived phenotypes. The model is implemented using a Hamiltonian Monte Carlo (HMC) algorithm. In Section 2, we describe our Bayesian model and

our derivation of best linear unbiased predict (BLUP) for linear mixed models. In Section 3, we propose a method that deals with noisy BLUPs. In Section 4, we do simulations under several settings, and apply our proposed method to resting state functional magnetic resonance imaging data from ABCD Study. In Section 5, we give our concluding remarks.

2.3 Methods

2.3.1 Model

We posit a generative linear model for continuous responses. Let $y = (y_1, \ldots, y_n)^T$ be a vector of standardized (i.e., mean zero and unit standard deviation) neurobehavioral values on n independent subjects. There are also B imaging-derived phenotypes (IDPs) potentially involved in generating y. Let $\mathbf{W}$ denote the $n \times B$ matrix

$$\mathbf{W} = \begin{pmatrix} \mathbf{w_1^T} \\ \cdots \\ \mathbf{w_n^T} \end{pmatrix},$$

where the $\mathbf{w_i}$ are B-dimensional vectors of standardized "noise-free" IDPs for the ith subject (note, the $\mathbf{w_i}$ are not directly observable in our setting). We assume the linear generative model

$$\mathbf{y} = \mathbf{W}\beta + \varepsilon, \tag{2.1}$$

where $\beta \sim \mathrm{N}_B(0, \sigma_\beta^2 I)$ are independent of $\varepsilon \sim \mathrm{N}_n(0, \sigma_\varepsilon^2 I)$. This is essentially the GCTA model. Under the GCTA framework, the parameters σ_β^2 and σ_ε^2 of model (2.1) are estimated using restricted maximum likelihood (REML) implemented via the average information algorithm [12]. Once estimates $(\hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2)$ are obtained, β can be estimated using the the Best Linear Unbiased

Predictor (BLUP) given by

$$\begin{aligned}\hat{\beta}_{\text{BLUP}} = \hat{E}\{\beta|W,y) &= \left(W^TW/\hat{\sigma}_\varepsilon^2 + \hat{\sigma}_\beta^{-2}I\right)^{-1} W^Ty/\hat{\sigma}_\varepsilon^2 \\ &= \left(\frac{1}{n}W^TW + \frac{\hat{\sigma}_\varepsilon^2}{n\hat{\sigma}_\beta^2}I\right)^{-1}\hat{b}_W, \end{aligned} \tag{2.2}$$

where $\hat{b}_W$ is the B-dimensional vector of mass univariate linear regression coefficients of $\mathbf{y}$ on $\mathbf{W}$. $\hat{\beta}_{\text{BLUP}}$ could then be used to infer the neurobehavioral value of a new participant by combining this estimate with their IDPs $\mathbf{w}_{\text{new}}$

$$\hat{y}_{\text{pred}} = \mathbf{w}_{\text{new}}^{\mathbf{T}}\hat{\beta}_{\text{BLUP}}.$$

This prediction method is functionally similar to the *polygenic risk score* (PRS) approaches commonly-used in genetics research, wherein PRS weights are computed from GWAS association statistics and used to infer phenotypic values based on genotypes [24].

However, in our application W is not directly observable, since each IDP is in fact measured with a significant level of noise. What is actually observed is

$$\begin{aligned}\mathbf{X} &= \mathbf{W} + \Delta \\ \Delta &= \begin{pmatrix} \delta_1^T \\ \cdots \\ \delta_n^T \end{pmatrix},\end{aligned} \tag{2.3}$$

whre δ_i is a B-dimensional vector of noise for the ith participant. We assume $w_i \sim \mathrm{N}_B(0,\Sigma_w)$ and $\delta_i \sim \mathrm{N}_B(0,\Sigma_\delta)$ are independent of each other and of β and ε. These assumptions (normality and independence) are often reasonable approximations in practice, as we demonstrate in our real-data application below.

In the model with noisy IDPs, the conditional posterior of β is given by

$$\begin{aligned} f(\beta|X,y,\sigma^2_\beta,\sigma^2_\varepsilon,\Sigma_w,\Sigma_\delta) &\propto \prod_{i=1}^{n} f(y_i|x_i,\beta,\sigma^2_\varepsilon,\Sigma_w,\Sigma_\delta) f(\beta|\sigma^2_\beta) \\ f(y_i|x_i,\beta,\sigma^2_\varepsilon,\Sigma_w,\Sigma_\delta) &= \int f(y_i|w_i,x_i,\beta,\sigma^2_\varepsilon) f(w_i|x_i,\Sigma_w,\Sigma_\delta) dw_i \end{aligned}$$

where

$$\begin{aligned} w_i|x_i,\Sigma_w,\Sigma_\delta &\sim \mathrm{N}(\mu_{w_i|x_i},\Sigma_{w|x}) \\ \mu_{w_i|x_i} &= (\Sigma_\delta^{-1}+\Sigma_{\mathbf{w}}^{-1})^{-1}\Sigma_\delta^{-1}\mathbf{x}_i \\ \Sigma_{w|x} &= (\Sigma_\delta^{-1}+\Sigma_{\mathbf{w}}^{-1})^{-1}. \end{aligned}$$

Using Woodbury's formula and integrating out the $\mathbf{W}$, we get

$$\begin{aligned} y_i|x_i,\beta,\Sigma_w,\Sigma_\delta &\sim \mathrm{N}(\mu_{y_i|x,\beta},\sigma^2_{y|x,\beta}) \\ \mu_{y_i|x_i,\beta} &= \mu^T_{w_i|x_i}\beta \\ \sigma^2_{y|x,\beta} &= \sigma^2_\varepsilon+\beta^T\Sigma_{w|x}\beta. \end{aligned} \tag{2.4}$$

Thus, the conditional posterior distribution of β given the observed data and the other paramters is given by

$$\begin{aligned} f(\beta|X,y,\Sigma_w,\Sigma_\delta) &\propto \exp\left\{-\frac{1}{2}\left[(y-M_{w|x}\beta)^T(y-M_{w|x}\beta)/(\sigma^2_\varepsilon+\beta^T\Sigma_{w|x}\beta)+\beta^T\beta/\sigma^2_\beta\right]\right\} \\ M_{w|x} &= \begin{pmatrix} \mu^T_{w_1|x_1} \\ \cdots \\ \mu^T_{w_n|x_n} \end{pmatrix} \end{aligned} \tag{2.5}$$

Equation (2.5) is not a standard distribution, and there is no closed-form expression for the posterior expectation to estimate β like the BLUP in (2.2). Instead, we propose to approximate

it (and the other parameters of (2.1)) via a Monte Carlo Markov Chain (MCMC) algorithm, as described next.

2.3.2 Hamiltonian Monte Carlo Inference

Denote the parameters in model (2.1) by $\theta = (\beta, \sigma_\beta^2, \sigma_\varepsilon^2)$ Since $f(\theta|\mathbf{X}, \mathbf{y})$ is not a standard distribution, we employ a Hamiltonian Monte Carlo (HMC) algorithm [25] to estimate the posterior. An essential advantage of HMC over more traditional MCMC methods, such as the Metropolis-Hastings (MH) algorithm, is greatly improved computational efficiency, especially with higher-dimensional and more complex models. HMC algorithms improve the efficiency of MH by employing a guided proposal generation scheme, using the gradient of the log posterior to direct the Markov chain towards regions of higher posterior density, where most samples are taken.

In a Hamiltonian system, the horizontal and vertical positions are given by $(\theta, \mathbf{p})$. In MCMC, we are interested in θ, which follows the posterior density $f(\theta)$ of interest. The momentum $\mathbf{p}$ is generated from a parametric distribution. Let $U(\theta) := -\log f(\theta)$; for momentum, we typically assume $\mathbf{p} \sim N_k(0, \mathbf{M})$, where $\mathbf{M}$ is a user-specified covariance matrix. Under this formulation, we have

$$H(\theta, \mathbf{p}) = -\log f(\theta) + \frac{1}{2}\mathbf{p}^T \mathbf{M}^{-1} \mathbf{p}$$

Over time, HMC travels on trajectories that are governed by the following first-order differential equations, known as the Hamiltonian equations:

$$\frac{d\mathbf{p}}{dt} = -\frac{\partial H(\theta, \mathbf{p})}{\partial \theta} = -\frac{\partial U(\theta)}{\partial \theta} = \nabla_\theta \log f(\theta)$$
$$\frac{d\theta}{dt} = \frac{\partial H(\theta, \mathbf{p})}{\partial \mathbf{p}} = \frac{\partial K(\mathbf{p})}{\partial \mathbf{p}} = \mathbf{M}^{-1}\mathbf{p}$$

The $leapfrog$ method is a widely-used approach for approximating the solutions to Hamiltonian equations [26]. The leapfrog algorithm uses a discrete step size ε individually for $\mathbf{p}$ and θ, with

a full step ε in θ sandwiched between two half-steps $\varepsilon/2$ for p,

$$\begin{aligned}
\mathbf{p}(t+\varepsilon/2) &= \mathbf{p}(t) + (\varepsilon/2)\nabla_\theta \log f(\theta(t)), \\
\theta(t+\varepsilon) &= \theta(t) + \varepsilon \mathbf{M}^{-1}\mathbf{p}(t+\varepsilon/2), \\
\mathbf{p}(t+\varepsilon) &= \mathbf{p}(t+\varepsilon/2) + (\varepsilon/2)\nabla_\theta \log f(\theta(t+\varepsilon)).
\end{aligned}$$

For a given momentum vector $\mathbf{p}$ within an HMC iteration, the path defined by the Hamiltonian equations is deterministic. An exact solution, if achievable, should always be accepted. However since our solution from the leapfrog is an approximation, a Metropolis style accept/reject step is added to ensure the newly generated proposal does not deviate too far from the specified Hamiltonian $H(\theta, \mathbf{p})$.

The efficiency of an HMC algorithm can be improved through parameter tuning and reparameterization. Two parameters that need to be specified are the step size ε and the number of leapfrog steps L. A smaller ε results in closer approximations and thus higher acceptance rates. But a small ε must be coupled with a large L to ensure the trajectory length εL is large enough to move the simulated parameter to a distant point in the distribution. If εL is too large the trajectory is likely to circle back. To tune ε and L is to find the right combinations of these values, which are usually chosen via monitoring the acceptance rate (approximately 65%, [27]). While ε and L can be tuned jointly, most analysts choose to select the step size first, then under a given step size, they fine-tune the number of steps per leapfrog L.

2.3.3 Posterior Inferences

Based on (2.4), the likelihood for $\mathbf{y}$ is given by

$$\begin{aligned}
f\left(\mathbf{y} \mid X, \beta, \Sigma_w, \Sigma_\delta, \sigma_\beta^2, \sigma_\varepsilon^2\right) = (2\pi(\sigma_\varepsilon^2 + \beta^T \Sigma_{w|x}\beta))^{-\frac{n}{2}} \cdot \\
\exp\{-\frac{1}{2}(y_i - \mu_{w_i|x_i}^T \beta)^2/(\sigma_\varepsilon^2 + \beta^T \Sigma_{w|x}\beta)\}
\end{aligned}$$

Consider the following priors for β, σ_β^2 and σ_ε^2:

$$\beta \sim N_B\left(0, \sigma_\beta^2 I\right)$$
$$\sigma_\beta^2 \sim \text{Inv}-\text{Gamma}\,(a_0, b_0)$$
$$\sigma_\varepsilon^2 \sim \text{Inv}-\text{Gamma}\,(a_1, b_1)$$

We assume $\Sigma_w = \sigma_w^2 \mathbf{I}$ is a diagonal matrix (perhaps after an appropriate transformation of IDPs to de-correlate the features). The prior for $\sigma_\mathbf{w}^2$ is also Inverse Gamma

$$\sigma_\mathbf{w}^2 \sim \text{Inv}-\text{Gamma}\,(a_2, b_2)$$

The support of σ_β^2, σ_ε^2, and $\sigma_\mathbf{w}^2$ is $(0,\infty)$, so we take a logarithmic transformation to expand the support to $\mathbb{R}$, that is, $\gamma_0 = \log(\sigma_\beta^2)$, $\gamma_1 = \log(\sigma_\varepsilon^2)$, $\gamma_2 = \log(\sigma_\mathbf{w}^2)$. Thus, we have

$$\log(\pi(\gamma_0|a_0, b_0)) \propto -a_0\gamma_0 - \frac{b_0}{e^{\gamma_0}}$$

$$\log(\pi(\gamma_1|a_1, b_1)) \propto -a_1\gamma_1 - \frac{b_1}{e^{\gamma_1}}$$
$$\log(\pi(\gamma_2|a_2, b_2)) \propto -a_2\gamma_2 - \frac{b_2}{e^{\gamma_2}}$$

The log posterior is proportional to the log-likelihood plus the log prior:

$$\begin{aligned}
&\log(f(\beta, \gamma_0, \gamma_1, \gamma_2|\mathbf{y}, \mathbf{X}, a_0, b_0, a_1, b_1, a_2, b_2)) \\
&\propto -\frac{n}{2}\log(e^{\gamma_1} + \beta^T\Sigma_{w|x}\beta) - \frac{1}{2}[(\mathbf{y} - \mathbf{M}_{\mathbf{w}|\mathbf{x}}\beta)^T(\mathbf{y} - \mathbf{M}_{\mathbf{w}|\mathbf{x}}\beta)/(e^{\gamma_1} + \beta^T\Sigma_{w|x}\beta)] \\
&- \gamma_0 p/2 - \frac{\beta^T\beta}{2e^{\gamma_0}} - a_0\gamma_0 - \frac{b_0}{e^{\gamma_0}} - a_1\gamma_1 - \frac{b_1}{e^{\gamma_1}} - a_2\gamma_2 - \frac{b_2}{e^{\gamma_2}}
\end{aligned}$$

The noise covariance matrix Σ_δ needs to be estimated from external test-retest data. We assume these data exist and are used to produce a positive definite covariance matrix $\hat{\Sigma}_\delta$ which is plugged into the HMC algorithm.

The parameters of interest are defined as $\theta = (\beta_0, \cdots, \beta_B, \gamma_0, \gamma_1, \gamma_2)^T$. To fit the model using Hamiltonian Monte Carlo Algorithm, we also need a vector for the gradient of the log posterior since the steps of the leapfrog algorithm used in Hamiltonian Monte Carlo Algorithm:

$$\nabla_\beta f \propto -\frac{\beta}{e^{\gamma_0}} - \frac{n\Sigma_{w|x}\beta}{e^{\gamma_1} + \beta^T\Sigma_{w|x}\beta} + \frac{\mathbf{M}_{\mathbf{w}|\mathbf{x}}^T(\mathbf{y} - \mathbf{M}_{\mathbf{w}|\mathbf{x}}\beta)(e^{\gamma_1} + \beta^T\Sigma_{w|x}\beta) + (\mathbf{y} - \mathbf{M}_{\mathbf{w}|\mathbf{x}}\beta)^T(\mathbf{y} - \mathbf{M}_{\mathbf{w}|\mathbf{x}}\beta)\Sigma_{w|x}\beta}{(e^{\gamma_1} + \beta^T\Sigma_{w|x}\beta)^2}$$

$$\nabla_{\gamma_0} f \propto -p/2 + \frac{\beta^T\beta}{2e^{\gamma_0}} - \alpha_0 + \beta_0 e^{-\gamma_0}$$

$$\nabla_{\gamma_1} f \propto -\frac{n}{2}\frac{e^{\gamma_1}}{e^{\gamma_1} + \beta^T\Sigma_{w|x}\beta} + \frac{1}{2}\frac{(\mathbf{y} - \mathbf{M}_{\mathbf{w}|\mathbf{x}}\beta)^T(\mathbf{y} - \mathbf{M}_{\mathbf{w}|\mathbf{x}}\beta)e^{\gamma_1}}{(e^{\gamma_1} + \beta^T\Sigma_{w|x}\beta)^2} - \alpha_1 + \beta_1 e^{-\gamma_1}$$

$$\nabla_{\gamma_2} f \propto -\frac{n}{2}\frac{\frac{\partial\beta^T\Sigma_{w|x}\beta}{\partial\gamma_2}}{e^{\gamma_1} + \beta^T\Sigma_{w|x}\beta} - \frac{1}{2}\frac{\frac{\partial(\mathbf{y}-\mathbf{M}_{\mathbf{w}|\mathbf{x}}\beta)^T(\mathbf{y}-\mathbf{M}_{\mathbf{w}|\mathbf{x}}\beta)}{\partial\gamma_2}e^{\gamma_1} + \beta^T\Sigma_{w|x}\beta - (\mathbf{y} - \mathbf{M}_{\mathbf{w}|\mathbf{x}}\beta)^T(\mathbf{y} - \mathbf{M}_{\mathbf{w}|\mathbf{x}}\beta)\frac{\partial\beta^T\Sigma_{w|x}\beta}{\partial\gamma_2}}{(e^{\gamma_1} + \beta^T\Sigma_{w|x}\beta)^2} - \alpha_2 + \beta_2 e^{-\gamma_2}$$

where

$$\frac{\partial\beta^T\Sigma_{w|x}\beta}{\partial\gamma_2} = \beta^T\frac{\partial\Sigma_{w|x}}{\partial\gamma_2}\beta$$

$$\begin{aligned}
\frac{\partial \Sigma_{w|x}}{\partial \gamma_2} &= \frac{\partial (\Sigma_\delta^{-1} + \exp(-\gamma_2) I)^{-1}}{\partial \gamma_2} \\
&= -(\Sigma_\delta^{-1} + \exp(-\gamma_2) I)^{-1} \frac{\partial (\Sigma_\delta^{-1} + \exp(-\gamma_2) I)}{\partial \gamma_2} (\Sigma_\delta^{-1} + \exp(-\gamma_2) I)^{-1} \\
&= -\Sigma_{w|x} [-\exp(-\gamma_2) I] \Sigma_{w|x} \\
&= \exp(-\gamma_2) \Sigma_{w|x} \Sigma_{w|x}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial (\mathbf{y} - \mathbf{M}_{\mathbf{w}|\mathbf{x}}\beta)^T (\mathbf{y} - \mathbf{M}_{\mathbf{w}|\mathbf{x}}\beta)}{\partial \gamma_2} &= \frac{\partial (\mathbf{y} - \mathbf{X}\Sigma_\delta^{-1}\Sigma_{w|x}\beta)^T (\mathbf{y} - \mathbf{X}\Sigma_\delta^{-1}\Sigma_{w|x}\beta)}{\partial \gamma_2} \\
&= \frac{\partial \beta^T \Sigma_{w|x}^T \Sigma_\delta^{-1} \mathbf{X}^T \mathbf{X} \Sigma_\delta^{-1} \Sigma_{w|x} \beta}{\partial \gamma_2} - 2 \frac{\partial \mathbf{y}^T \mathbf{X} \Sigma_\delta^{-1} \Sigma_{w|x} \beta}{\partial \gamma_2} \\
&= \beta^T \frac{\partial \Sigma_{w|x}^T}{\partial \gamma_2} \Sigma_\delta^{-1} \mathbf{X}^T \mathbf{X} \Sigma_\delta^{-1} \Sigma_{w|x} \beta + \beta^T \Sigma_{w|x}^T \Sigma_\delta^{-1} \mathbf{X}^T \mathbf{X} \Sigma_\delta^{-1} \frac{\partial \Sigma_{w|x}}{\partial \gamma_2} \beta \\
&\quad - 2\mathbf{y}^T \mathbf{X} \Sigma_\delta^{-1} \frac{\partial \Sigma_{w|x}}{\partial \gamma_2} \beta
\end{aligned}$$

After obtaining the log posterior and the vector of gradients of the log posterior, we implement the HMC algorithm using the `hmclearn` package in `R`.

2.4 Results

2.4.1 Simulations

To test the performance of the model in simulations, we are setting $n = 500$ and $B = 34$. The FVE (total fraction of variance explained) by $\mathbf{W}$ for phenotyoe $\mathbf{y}$ is given by $h^2 = B\sigma_\beta^2$, h^2 is called the "SNP heritability" in the GWAS literature. In our simulations, we set the number of MCMC samples $N = 5000$, FVE is set to 0.25, 0.5 and 0.75; σ_δ^2 is set to 0, 0.5, 1 and 2. Σ_ε set to be the true Σ_ε obtained from our real dataset of the N-back task MRI data in ABCD study. We also compare with the BLUP that does not assume measurement noise in the predictors.

As we can see from the Table 2.1, in which $\hat{h}^2_{\mathbf{w},\mathbf{x}}$ denotes the noisy BLUP and $\hat{h}^2_{\mathbf{x}}$ denotes the BLUP. The noisy BLUP provides FVE estimates which are very close to the true FVE we are setting

Table 2.1. Simulation results for Noisy BLUP and "Noisy-Free" BLUP

	$\sigma_\delta^2 = 0$	$\sigma_\delta^2 = 0.5$		$\sigma_\delta^2 = 1$		$\sigma_\delta^2 = 2$	
	$\hat{h}^2_{\mathbf{w},\mathbf{x}}$	$\hat{h}^2_{\mathbf{x}}$	$\hat{h}^2_{\mathbf{w}\|\mathbf{x}}$	$\hat{h}^2_{\mathbf{x}}$	$\hat{h}^2_{\mathbf{w}\|\mathbf{x}}$	$\hat{h}^2_{\mathbf{x}}$	$\hat{h}^2_{\mathbf{w}\|\mathbf{x}}$
$h^2 = 0.75$	0.73	0.39	0.74	0.38	0.76	0.38	0.75
$h^2 = 0.5$	0.53	0.29	0.52	0.30	0.54	0.26	0.48
$h^2 = 0.25$	0.24	0.19	0.24	0.14	0.26	0.13	0.25

regardless of the level of noise; while the BLUP gives biased FVE estimates which becomes progressively more attenuated as σ_ε^2 becomes larger.

2.4.2 Real Data Analysis: Application to Resting-State fMRI Data

We have applied our proposed method to data taken from the Adolescent Brain Cognitive Development (ABCD) Study, using resting-state fMRI data from Release 4.0. The ABCD Study has primary data collections sites at 21 research institutes across the US. Participants include $n = 33818$ subjects aged 40-60, there are 33814 participants for visit 1 and 3977 for visit 2; 3973 subjects involved in both visits.

The resting-state MRI features of interest $\mathbf{X}$, are Fisher z-transformed Pearson correlations between pairs of cortical vertices. The cognitive phenotype $\mathbf{Y}$ is the FIQ measured by the NIH Toolbox. We used resting state MRI and NIH Toolbox data collected at both visits. The number of features was $B = 210$, then the dimension of Σ_w is 210×210. We residualized each feature of X on age, sex, income and the first 10 principal components of genetic ancestry (PC1 to PC10) and MRI scanner instance. The residualized X were then standardized to have zero (column) means and unit variances. The error covariance matrix Σ_δ is calculated using the residuals of each visit.

We first applied the Noisy BLUP on the dataset and ran the Bayesian paradigm using HMC with 10,000 iterations with 3,000 burn-ins. The initial values were all set to be 0; the tuning parameters are $\varepsilon = 0.001$ and $L = 20$. To compare our model's performance with the model without measurement error, we also applied the HMC algorithm on the "noise-free" model (2.1), i.e. $\Delta = \mathbf{0}$ in (2.3), using the same parameters as in the Noisy BLUP model.

Since the dimension of Σ_W is 210×210, there are $(1 + 210) \times 210/2 = 22155$ parameters to estimate for a random positive definite matrix. We considered using a regularization method to make

the Σ_W more sparse. [28] proposed the graphic lasso method through the use of L_1 (lasso) regularization. After trying 10-fold cross-validation for estimation of the parameter ρ (the parameter of L_1 norm), the results indicated that the unregularized model is the best.

The results showed that for the "noisy-free" model, we get $FVE = 11.66\%$. Then we apply our noisy BLUP method on the resting-state fMRI data, we get $FVE = 27.97\%$. As can be seen in this result, the model that assumes there is measurement error in variables has higher fraction of variance explaine than the model does not assume. In this way, using our proposed method for noisy predictors could explain a larger fraction of the variance, that is, higher SNP heritability.

2.5 Discussion

In this paper, we have developed the best linear unbiased predictor (BLUP) for noisy brain data. We proposed a Bayesian methodology with the application of Hamiltonian Monte Carlo Algorithm (HMC) to correct for measurement error in predictors and estimate the SNP heritability h^2 and the FVE. While we applied the noisy BLUP model to resting state fMRI data from the ABCD Study, the model could improve the FVE significantly. In future work, we plan on implementing the noisy BLUP model with a more complex covariance matrix Σ_W assumption, e.g. AR(1), exchangeable.

Chapter 2, in full, has been submitted for publication as "Chenyu Liu, Bohan Xu, Chun Fan, Wesley Thompson. *Estimating the Total Fraction of Variance Explained from Noisy Brain Imaging Data*". The dissertation author was the primary author on this paper.

Chapter 3

Sign-flipped Doubly Robust Score-like Test for Average Treatment Effect in Causal Inference

3.1 Abstract

The doubly robust estimator (DRE) has been increasingly used to provide robust inference about the average treatment effect (ATE) in causal inference. By integrating the semiparametric outcome regression model (ORM) and marginal structural models (MSM), the DRE doubles the chance of making correct inference about ATE, as it only requires one of the ORM and MSM to be correctly specified for valid inference. More recently, the DRE has been extended to allow the ORM and MSM to have non-parametric specification for covariates, further increasing inference validity. Although the nice asymptotic properties of DRE ensure valid asymptotic inference, its performance for small and moderate samples is not good, with anticonservative bias in type I error rates, or bias in coverage probability, especially when there is heteroscedasticity for continuous and overdispersion for count responses.

In this paper, we propose a sign-flipping score-like test to address such bias in type I error rates when applying DRE to studies with small or moderate samples. The proposed approach leverages the work by Hemerik et al. (2020) and extends their sign-flipping score tests for parametric generalized linear models to the semiparametric and non-parametric settings of DRE. We examine both asymptotic properties and finite-sample performances of the sign-flipping score-like test proposed.

KEY WORDS: Causal inference, ATE, Machine Learning Methods, Score test, Sign-flipping

3.2 Introduction

In this paper, we consider the problem of hypothesis testing about average treatment effects in causal inference with small or moderate samples. Understanding causality is crucial for making accurate predictions and taking effective actions in our daily lives [29]. The fundamental problem of causal inference is that we can see only one potential outcome for any given unit [30]. One popular approach for inferring causality from observational data is outcome regression model (ORM) [31]. This classic regression approach has been widely used to estimate causal effects of a treatment by controlling for the effects of other variables that might be influencing the outcome [32]. Different types of regression models with different assumptions can be used for causal inference, such as parametric and semiparametric generalized linear models (GLM) [33]. A major issue of applying a parametric ORM is the correct specification of the regression model [34]. When applying parametric GLM, misspecified distribution model generally leads to biased estimates. Another major problem is the correct specification of the covariates in the regression model. Although a semiparametric MSM alleviates this problem, it still requires a correct specification of the covariates in the model. An alternative is the semiparametric marginal structural model (MSM). Like the semiparametric ORM, it imposes no parametric model on the conditional distribution of the mean given the explanatory variables [35]. Unlike the semiparametric ORM, the regression model of MSM only contains the treatment variables as the predictor and controls for confounders by balancing such covariates through weighting each subject with the inverse of the propensity score (usually modeled using logistic regression) [36]. Correct inference by either the semiparametric ORM and MSM requires a correct specification of the covariates (in the regression model for ORM and propensity score model for MSM). Moreover, estimators from MSM obtained by the inverse probability weight (IPW) technique generally converge slower to their asymptotic distributions than the ORM estimators, yielding anticonservative type I errors, or coverage probabilities, even for moderate and fairly large samples [37].

A doubly robust estimator (DRE) integrates the semiparametric ORM and MSM to improve the performance of either approach when applied individually [38]. Specifically, it provides valid inference about causal effects, if the covariates are correctly modeled in only one of the two modules. In addition to doubling the chance of correct inference, DRE has more power than either ORM or MSM, if the covariates

are correctly modeled in both modules [39]. DRE estimators are less anticonservative than those from MSM. However, it still does not perform well for small or moderate samples.

Hemerik et al. (2020) [40] proposed a novel sign-flipping score test for parametric GLM by sign-flipping individual score contributions. The proposed score test is shown to perform very well for small and moderate samples when employed to address violations of parametric models such as overdispersion for count and heteroscedasticity for continuous responses. Moreover, although their paper is focused on parametric GLM, the sign-flipped score has an asymptotic that actually does not depend posited parametric models for the conditional distribution of the response given the explanatory variables, thereby enabling it to apply to semiparametric GLM.

In this paper, we propose a sign-flipping score-like test for DRE to improve its performance for small or moderate samples when applied for causal inference about average treatment effects. In Section 3.3, we give an overview of the three approaches (semiparametric OR, MSM and DRE) in causal inference. In Section 3.4 we develop a sign-flipping score-like test for DRE to provide inference about average treatment effect. Sectionm 3.5 contains simulation studies to examine performance of the proposed approach for small and moderate samples. In Section 3.6 we give our concluding remarks.

3.3 Causal Inference

There are three general assumptions needed to draw causal inference [41]: (1) Counterfactual Consistency posits that an individual's potential outcome under their observed treatment is the same as their actual observed outcome; (2) Ignorability, which ensures that the potential outcomes are independent of treatment assignment; (3) Positivity, states that there is positive probability of receiving every level of exposure for every combination of values of exposure and confounders that occur among individuals in the population [42].

For notational brevity and without loss of generality, we focus on two treatment groups and a continuous outcome. Our considerations are readily extended to multiple treatment conditions.

Consider a sample of n subjects and let $\mathbf{y}_i = (y_{i,0}, y_{i,1})^\top$ denote the potential continuous outcomes

of the ith subject for the two treatment conditions. Let

$$E(y_{i,k}) = \mu_k = \beta_0 + \beta_1 k, \quad 1 \leq i \leq n, \quad k = 0, 1. \tag{3.1}$$

The average treatment effect (ATE) is defined by:

$$ATE = E(\Delta_i) = E(y_{i,1} - y_{i,0}) = \beta_1 \tag{3.2}$$

where $\Delta_i = y_{i,1} - y_{i,0}$ is the treatment difference for the ith subject. Since only one of the $y_{i,0}$ and $y_{i,1}$ is observed, Δ_i cannot be computed, precluding the application of conventional statistical methods for inference about β_1 such as the paired t-test. However, the ATE in (3.2) is well defined and estimated.

Below we review the three popular approaches for inference about the ATE β_1. We start with the semiparametric OR.

3.3.1 Outcome Regression Model

We start with randomized control trials (RCTs). Let z_i denote the indicator variable for treatment group, i.e., $z_i = 0$ (1) for the control (treated) group. In this case, the potential outcomes $y_{i,k}$ are independent of z_i, i.e., $y_{i,k} \perp z_i$, for $1 \leq i \leq n$, which is known as the ignorability assumption. By the counterfactual consistency assumption, we observe $y_{i,k}$ if the ith subject receives the kth intervention. We denote the observed individual subject's outcome by (y_i, z_i). Thus $y_i = y_{i,k}$ if $z_i = k$.

$$y_i = \begin{cases} y_{i,0} & \text{if } z_i = 0 \\ y_{i,1} & \text{if } z_i = 1 \end{cases}$$

We can express the mean of the potential outcome as:

$$E(y_{i,k}) = E(y_i \mid z_i = k) \tag{3.3}$$

By $y_{i,k} \perp z_i$, we have:

$$E(y_{i,k}) = E(y_{i,k} \mid z_i = k) = E(y_i \mid z_i = k) \tag{3.4}$$

If we model (y_i, z_i) using a semiparametric linear regression, we have:

$$E(y_i \mid z_i) = \beta_0 + \beta_1 z_i + \varepsilon_i, \quad 1 \leq i \leq n \tag{3.5}$$

By (3.4), we have:

$$E(y_{i,k}) = E(y_i \mid z_i = k) = \begin{cases} \beta_0 & \text{if } z_i = 0 \\ \beta_0 + \beta_1 & \text{if } z_i = 1 \end{cases} \tag{3.6}$$

$$ATE = E(\Delta_i) = E(y_{i,1} - y_{i,0}) = E(y_i \mid z_i = 1) - E(y_i \mid z_i = 0) = \beta_1$$

Thus for RCTs, we can fit the linear model in (3.5) to the observed data (y_i, z_i) for inference about the ATE β_1.

For observational studies, the assumption of $y_{i,k} \perp z_i$ is no longer true and thus the counterfactual consistency assumption (3.4) no longer holds. We assume instead the ignorability condition, $y_{i,k} \perp z_i \mid x_i$, in which case the counterfactual consistency assumption is given by:

$$E(y_{i,k} \mid x_i) = E(y_{i,k} \mid x_i, z_i = k) = E(y_i \mid x_i, z_i = k) \tag{3.7}$$

The OR approach is posit a parametric linear regression to explicitly control for x_i.

Let y_i denote the observed data:

$$y_i = \begin{cases} y_{i,0} & \text{if } z_i = 0 \\ y_{i,1} & \text{if } z_i = 1 \end{cases} \qquad 1 \leq i \leq n$$

Then under linear regression, we have:

$$E(y_i \mid x_i, z_i) = \beta_0 + \beta_1 z_i + \beta_2 x_i, \quad 1 \leq i \leq n \tag{3.8}$$

It follows from the above

$$E(E(y_i \mid x_i, z_i) \mid x_i) = E(y_i \mid x_i) = \beta_0 + \beta_1 E(z_i \mid x_i) + \beta_2 x_i \tag{3.9}$$

By (3.7) and (3.9), we have:

$$E\left(y_{i,k} \mid x_i\right) = E\left(y_i \mid x_i, z_i = k\right) = \beta_0 + \beta_1 k + \beta_2 x_i$$

By the iterated conditional expectation, we have:

$$\begin{aligned} E\left(y_{i,k}\right) &= E\left[E\left(y_{i,k} \mid x_i\right)\right] \\ &= E\left(\beta_0 + \beta_1 k + \beta_2 x_i\right) \\ &= \beta_0 + \beta_1 k + \beta_2 E\left(x_i\right) \end{aligned}$$

Thus

$$\begin{aligned} ATE &= E\left(y_{i,1}\right) - E\left(y_{i,0}\right) \\ &= \beta_0 + \beta_1 + \beta_2 E\left(x_i\right) - \left[\beta_0 + \beta_2 E\left(x_i\right)\right] \\ &= \beta_1 \end{aligned}$$

and β_1 in (3.9) is the ATE of interest.

Unlike RCTs, $\beta_0 + \beta_2 E\left(x_i\right)$ is the mean of $y_{i,0}$ for observational studies, but β_1 still has the same interpretation as the ATE.

Now consider testing the hypothesis concerning β_1:

$$H_0 : \beta_1 = \beta_{10} \quad \text{vs.} \quad H_a : \beta_1 \neq \beta_{10} \tag{3.10}$$

We can test the null using Wald or score-like test. Under the Wald test, we obtain an estimator $\widehat{\beta}_1$ of β_1 by solving the estimating equations (EE) and make inference about β_1 using the asymptotic distribution of the EE estimator $\widehat{\beta}_1$. If we use the score-like test, we partition the parameters into the parameter of interest, β_1, and the nuisance parameters, $\gamma = \left(\beta_0, \beta_2\right)^\top$. We estimate γ with $\beta_1 = \beta_{10}$ and then construct the efficient score-like test based on the estimated $\widehat{\gamma}$, which has an asymptotic normal distribution. We discuss such efficient scores in Section 3.4.1.

Although both the Wald and score-like test have nice asymptotic properties, type I errors based on such asymptotic distributions often exhibit anticonservative behavior for small and moderate samples. Although the efficient score-like test performs better, it still has bias. The sign-flipping score test in Hemerik et al.(2020) [40] shows much better type I error levels than its asymptotic counterpart when applied to parametric GLM using simulated data. As the asymptotic distribution of the sign-flipping score test does not depend on the posited parametric model, it can be applied to the semiparametric GLM to address the anticonservatism of the asymptotic score-like test for small and moderate samples. Our own simulation study also confirms the superior performance of this randomization test for semiparametric GLM. We will discuss the extension of this sign-flipping score test to the semiparametric OR in Section 3.4.

3.3.2 Marginal Structural Model

The MSM is a popular semiparametric alternative to model the potential outcomes $\mathbf{y}_i = (y_{i,0}, y_{i,1})^\top$ using missing data methods for semiparametric regression.

Consider the following semiparametric model for $\mathbf{y}_i$:

$$E(\mathbf{y}_i) = \mu, \quad 1 \le i \le n \tag{3.11}$$
$$\mu = (\mu_0, \mu_1)^\top, \quad \mu_0 = \beta_0, \quad \mu_1 = \beta_0 + \beta_1$$

Since one of the $y_{i,k}$'s is always missing, we can view inference about μ, or about $\beta = (\beta_0, \beta_1)^\top$, as a missing data problem and apply missing data methods for inference about β.

For the semiparametric linear model in (3.11), let:

$$D_i = \frac{\partial \mu}{\partial \beta} = \begin{pmatrix} \frac{\partial \mu_0}{\partial \beta_0} & \frac{\partial \mu_1}{\partial \beta_0} \\ \frac{\partial \mu_0}{\partial \beta_1} & \frac{\partial \mu_1}{\partial \beta_1} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad V_i = \mathbf{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$
$$\Delta_i = \begin{pmatrix} \frac{1-z_i}{1-\pi_i} & 0 \\ 0 & \frac{z_i}{\pi_i} \end{pmatrix}$$

For RCTs, $\pi_i = \pi$ is a constant, while for observational studies, $\pi_i = \pi(x_i)$ is a function of x_i. In real data

analysis, π_i is the probability of subject i to be assigned in the group $z_i = 1$.

Consider the following weighted estimating equations (WEE):

$$\mathbf{U}^n = n^{-1/2} \sum_{i=1}^{n} \mathbf{U}^{ni} = n^{-1/2} \sum_{i=1}^{n} D_i V_i^{-1} \Delta_i \left(\mathbf{y}_i - \mu \right) = \mathbf{0} \tag{3.12}$$

It is readily checked that

$$E\left(\mathbf{U}^n\right) = n^{-1/2} \sum_{i=1}^{n} E\left(\mathbf{U}^{ni}\right) = \mathbf{0}$$

Since

$$\begin{aligned}
E\left(\mathbf{U}^{ni}\right) &= E\left[\Delta_i \left(\mathbf{y}_i - \mu\right)\right] = E\left[E\left[\Delta_i \left(\mathbf{y}_i - \mu\right) \mid x_i\right]\right] \\
&= E\left\{ \begin{pmatrix} \frac{1}{1-\pi_i} & 0 \\ 0 & \frac{1}{\pi_i} \end{pmatrix} E\left[\begin{pmatrix} 1-z_i & 0 \\ 0 & z_i \end{pmatrix} \left(\mathbf{y}_i - \mu\right) \mid x_i \right] \right\} \\
&= E\left\{ \begin{pmatrix} \frac{1}{1-\pi_i} & 0 \\ 0 & \frac{1}{\pi_i} \end{pmatrix} E\left[\begin{pmatrix} 1-z_i & 0 \\ 0 & z_i \end{pmatrix} \mid x_i \right] E\left[\left(\mathbf{y}_i - \mu\right) \mid x_i\right] \right\} \\
&= E\left\{ \begin{pmatrix} \frac{1}{1-\pi_i} & 0 \\ 0 & \frac{1}{\pi_i} \end{pmatrix} \begin{pmatrix} 1-\pi_i & 0 \\ 0 & \pi_i \end{pmatrix} E\left(\mathbf{y}_i - \mu\right) \right\} \\
&= E\left[E\left(\mathbf{y}_i - \mu\right)\right] \\
&= E\left(\mathbf{y}_i - \mu\right) \\
&= \mathbf{0}
\end{aligned}$$

the WEE in (3.12) is unbiased and thus yields consistent estimators $\widehat{\beta}$ of $\beta = (\beta_0, \beta_1)^\top$. We refer to $\widehat{\beta}$ as the inverse probability weighting (IPW) estimator below.

The WEE above in (3.12) simplifies to:

$$\mathbf{U}^n = n^{-1/2}\sum_{i=1}^{n}\mathbf{U}^{ni} = n^{-1/2}\sum_{i=1}^{n}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\Delta_i(\mathbf{y}_i-\mu) \tag{3.13}$$

$$= n^{-1/2}\sum_{i=1}^{n}\begin{pmatrix} \frac{1-z_i}{1-\pi_i}(y_{i,0}-\beta_0)+\frac{z_i}{\pi_i}(y_{i,1}-(\beta_0+\beta_1)) \\ \frac{z_i}{\pi_i}(y_{i,1}-(\beta_0+\beta_1)) \end{pmatrix}$$

$$= n^{-1/2}\sum_{i=1}^{n_0}\begin{pmatrix} \frac{1}{1-\pi_i}(y_{i,0}-\beta_0) \\ 0 \end{pmatrix} + n^{-1/2}\sum_{i=1}^{n_1}\begin{pmatrix} \frac{1}{\pi_i}(y_{i,1}-(\beta_0+\beta_1)) \\ \frac{1}{\pi_i}(y_{i,1}-(\beta_0+\beta_1)) \end{pmatrix}$$

$$= n^{-1/2}\sum_{i=1}^{n_0}\frac{1}{1-\pi_i}\begin{pmatrix} y_{i,0}-\beta_0 \\ 0 \end{pmatrix} + n^{-1/2}\sum_{i=1}^{n_1}\frac{1}{\pi_i}\begin{pmatrix} y_{i,1}-(\beta_0+\beta_1) \\ y_{i,1}-(\beta_0+\beta_1) \end{pmatrix}$$

As in OR, we can also express the $\mathbf{U}^{ni}$ in (3.13) using the observed outcomes (y_i, z_i) $(1 \le i \le n)$.

Consider the semiparametric linear regression for the observed outcomes (y_i, z_i):

$$E(y_i \mid z_i) = \mu_i = \beta_0 + \beta_1 z_i, \quad 1 \le i \le n \tag{3.14}$$

There is no missing data in the observed (y_i, z_i). Let

$$D_i = \frac{\partial \mu_i}{\partial \beta} = \begin{pmatrix} \frac{\partial \mu_i}{\partial \beta_0} \\ \frac{\partial \mu_i}{\partial \beta_1} \end{pmatrix} = \begin{pmatrix} 1 \\ z_i \end{pmatrix}, \quad V_i = 1, \quad S_i = y_i - \mu_i \tag{3.15}$$

We can estimate $\beta = (\beta_0, \beta_1)^\top$ by the EE:

$$\mathbf{H}^n = n^{-1/2}\sum_{i=1}^{n}\mathbf{H}^{ni} = n^{-1/2}\sum_{i=1}^{n}D_i V_i^{-1} S_i = n^{-1/2}\sum_{i=1}^{n}\begin{pmatrix} 1 \\ z_i \end{pmatrix}(y_i-\mu_i) \tag{3.16}$$

$$= n^{-1/2}\sum_{i=1}^{n}\begin{pmatrix} y_i-(\beta_0+\beta_1 z_i) \\ z_i[y_i-(\beta_0+\beta_1 z_i)] \end{pmatrix}$$

$$= n^{-1/2}\sum_{i=1}^{n_0}\begin{pmatrix} y_{i,0}-\beta_0 \\ 0 \end{pmatrix} + n^{-1/2}\sum_{i=1}^{n_1}\begin{pmatrix} y_{i,1}-(\beta_0+\beta_1) \\ y_{i,1}-(\beta_0+\beta_1) \end{pmatrix}$$

By weighting $\mathbf{H}^{ni}$ with $w_i = \frac{1}{1-\pi_i}$ for $z_i = 0$ and $w_i = \frac{1}{\pi_i}$ for $z_i = 1$, the weighted version of (3.16) is given by:

$$\begin{aligned}\mathbf{H}_w^n &= n^{-1/2}\sum_{i=1}^{n} w_i \mathbf{H}^{ni} = n^{-1/2}\sum_{i=1}^{n} w_i D_i V_i^{-1} S_i \\ &= n^{-1/2}\sum_{i=1}^{n_0} \frac{1}{1-\pi_i}\begin{pmatrix} y_{i,0}-\beta_0 \\ 0 \end{pmatrix} + n^{-1/2}\sum_{i=1}^{n_1} \frac{1}{\pi_i}\begin{pmatrix} y_{i,1}-(\beta_0+\beta_1) \\ y_{i,1}-(\beta_0+\beta_1) \end{pmatrix}\end{aligned} \tag{3.17}$$

which is identical to the WEE in (3.13), i.e., $\mathbf{H}_w^n = \mathbf{U}^n$.

Thus we can use WEE for the semiparametric linear regression in (3.13) for inference about β, which is implemented in most statistical software. For notational brevity, we will use $\mathbf{U}^n$ to refer to either WEE in (3.13) or (3.17) unless stated otherwise.

If π_i is known, we just estimate $\beta = (\beta_0, \beta_1)^\top$. In this case, β_1 is the parameter of interest and β_0 is the nuisance parameter. We partition D_i in (3.15) into:

$$D_i = \frac{\partial \mu_i}{\partial \beta} = \begin{pmatrix} \frac{\partial \mu_i}{\partial \beta_0} \\ \frac{\partial \mu_i}{\partial \beta_1} \end{pmatrix} = \begin{pmatrix} D_{i0} \\ D_{i1} \end{pmatrix} = \begin{pmatrix} 1 \\ z_i \end{pmatrix}, \quad V_i = 1$$

and partition $\mathbf{U}^{ni}$ and $\mathbf{U}^n$ into:

$$\mathbf{U}^{ni} = w_i D_i V_i^{-1} S_i = \begin{pmatrix} w_i D_{i0} V_i^{-1} S_i \\ w_i D_{i1} V_i^{-1} S_i \end{pmatrix} = \begin{pmatrix} U_0^{ni} \\ U_1^{ni} \end{pmatrix} \tag{3.18}$$

Under the null $H_0 : \beta_1 = \beta_{10}$, we obtain an asymptotically normal estimator $\widehat{\beta}_0$ of β_0 by solving the following reduced EE:

$$U_0^n(\beta_0, \beta_{10}) = n^{-1/2}\sum_{i=1}^{n} U_0^{ni}(\beta_0, \beta_{10}) = n^{-1/2}\sum_{i=1}^{n} w_i D_{i0} V_i^{-1} S_i = 0$$

With the estimated $\widehat{\beta}_0$, we obtain a score-like test for the null:

$$U_1^n\left(\widehat{\beta}_0, \beta_{10}\right) = n^{-1/2}\sum_{i=1}^{n} U_1^{ni}\left(\widehat{\beta}_0, \beta_{10}\right)$$

Although $U_1^n\left(\widehat{\beta}_0,\beta_{10}\right)$ has an asymptotic normal distribution, its asymptotic variance depends on the asymptotic variance of $\widehat{\beta}_0$. One popular approach to eliminating this dependence is to construct the efficient score-like statistic by orthogonalizing $U_1^n\left(\widehat{\beta}_0,\beta_{10}\right)$ to the tangent space of the nuisance parameter β_0. We discuss such efficient score statistics in Section 3.4.

3.3.3 Doubly Robust Estimator

A third approach is to combine the ORM and MSM to create an augmented weighted estimating equations to obtain a doubly robust estimator (DRE) for inference about β_1. The DRE approach ensures consistent estimation of β_1, if the covariates in the ORM or MSM (propensity score model) is correctly specified, thus doubling the chance of correct inference. Moreover, if the covariates are correctly specified in both ORM and MSM, the DRE estimator is more efficient than either ORM or MSM when applied individually.

Consider the combined ORM and MSM by the following joint model:

$$\begin{aligned} E(y_i \mid z_i, x_i) &= \beta_1 z_i + (\beta_0 + \beta_2 x_i) \\ E(z_i \mid \mathbf{x}_i) &= \pi(\mathbf{x}_i, \eta) \quad 1 \le i \le n \end{aligned} \tag{3.19}$$

In the above model, the first module is the ORM which is partitioned into one involving the parameter of interest, $\beta_1 z_i$, and one involving the nuisance parameters, $h(\mathbf{x}_i,\gamma) = \beta_0 + \beta_2 x_i$. The second propensity module specifies a semiparametric GLM for the probability of treatment assignment given the confounders $\mathbf{x}_i$.

Inference about β_1 is based on an augmented weighted estimating equations (AWEE).

The DRE estimator for the combined semiparametric $\pi(\mathbf{x}_i,\eta)$ and $h(\mathbf{x}_i,\gamma)$ has the following influence function [43]:

$$\psi(y_i,z_i,x_i;\beta_{10},\eta_0,\gamma_0) = \frac{z_i}{\pi_0(x_i,\eta_0)}\left[y_i - \beta_{10} - h(x_i,\gamma_0)\right] - \frac{1-z_i}{1-\pi(x_i,\eta_0)}\left[y_i - h(x_i,\gamma_0)\right] \tag{3.20}$$

The above IF, $\psi(y_i,z_i,x_i;\beta_{10},\eta_0,\gamma_0)$, is Neyman orghogonalized to both tangent spaces of the nuisance parameters η and $\gamma = (\beta_0,\beta_2)^\top$. For such an IF, the variance of the IF $\psi(y_i,z_i,x_i;\beta_{10},\eta_0,\gamma_0)$ is the

asymptotic variance of the DRE estimator. Thus with a consistent estimator $\widehat{\eta}$ and $\widehat{\gamma}$ of the respective nuisance parameters, we can readily estimate the asymptotic variance of the DRE estimator and use the asymptotic normal distribution of $\widehat{\beta}_1$ for inference about β_1.

In addition to the Wald test, we can also construct a score-like test based on the IF $\psi(y_i, z_i, x_i; \beta_1, \eta, \gamma)$ to test the null H_0 in (3.10). By substituting a consistent estimator of $\widehat{\eta}$ and $\widehat{\gamma}$ in place of η and γ in (3.20), we have:

$$U_{\beta_1} = \sum_{i=1}^{n} \psi(y_i, z_i, x_i; \beta_{10}, \widehat{\eta}, \widehat{\gamma}) \tag{3.21}$$

$$= \sum_{i=1}^{n} \frac{z_i}{\pi(x_i, \widehat{\eta})} [y_i - \beta_{10} - h(x_i, \widehat{\gamma})] - \frac{1 - z_i}{1 - \pi(x_i, \widehat{\eta})} [y_i - h(x_i, \widehat{\gamma})] \tag{3.22}$$

Under the null, U_{β_1} has an asymptotic normal distribution:

$$\sqrt{n} U_{\beta_1} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(y_i, z_i, x_i; \beta_{10}, \widehat{\eta}, \widehat{\gamma}) \rightarrow_d N\left(0, \sigma^2\right) \tag{3.23}$$

where $\sigma^2 = var(\psi(y_i, z_i, x_i; \beta_{10}, \eta_0, \gamma_0))$. A consistent estimator of σ^2 is readily constructed by the sample version of $var(\psi(y_i, z_i, x_i; \beta_{10}, \eta_0, \gamma_0))$ (see below).

The DRE approach has been recently extended to non-parametric $\pi(\mathbf{x}_i, \eta)$ and $h(\mathbf{x}_i, \gamma)$ to further improve the robustness of the estimator $\widehat{\beta}_1$. Under the non-parametric $\pi(\mathbf{x}_i, \eta)$ and $h(\mathbf{x}_i, \gamma)$, the conditional mean $E(y_i \mid z_i, x_i)$ becomes a partial linear model, where the parameter of interest β_1 retains the same simple linear form, whereas the nuisance $h(\mathbf{x}_i, \gamma)$, $\pi(\mathbf{x}_i, \eta)$ or both has a non-parametric form. We can estimate η and γ using machine learning methods such as random forest. Let $\pi_0(\mathbf{x}_i)$ and $h_0(\mathbf{x}_i)$ denote the true form of $\pi(\mathbf{x}_i)$ and $h(\mathbf{x}_i)$ and let $\widehat{\pi}(\mathbf{x}_i)$ and $\widehat{h}(\mathbf{x}_i)$ denote the respective estimators. A major difference between parametric/semiparametric and non-parametric estimators is the slower convergence rate of the asymptotic distribution. In particular, non-parametric $\widehat{\pi}(\mathbf{x}_i)$ ($\widehat{h}(\mathbf{x}_i)$) no longer has $\sqrt{n}$ convergence rate and as such the score-like statistic does not have the asymptotic normal distribution in (3.23).

A popular approach to addressing this issue is to construct a "debiased" machine learning estimator using sample splitting [44, 45]. Under this approach, we randomly equally split the sample into two: one with size $= n_I$, indexed by $i \in I$, and the other with size $= n_{I^c}$, indexed by $i \in I^c$. We use one subsample such as I^c to estimate the nuisance $\pi(\mathbf{x}_i)$ and $h(\mathbf{x}_i)$, and substitue the estimated $\widehat{\pi}_{I^c}(\mathbf{x}_i)$ and $\widehat{h}_{I^c}(\mathbf{x}_i)$ into

the IF in (3.22):

$$\psi_i\left(y_i,z_i,x_i;\beta_{10},\widehat{\pi}_{I^c},\widehat{h}_{I^c}\right)=\frac{z_i}{\widehat{\pi}_{I^c}(\mathbf{x}_i)}\left[y_i-\beta_{10}-\widehat{h}_{I^c}(\mathbf{x}_i)\right]-\frac{1-z_i}{1-\widehat{\pi}_{I^c}(\mathbf{x}_i)}\left[y_i-\widehat{h}_{I^c}(\mathbf{x}_i)\right] \tag{3.24}$$

By switching the roles of the subsamples I and I^c in estimating the nuisance parameters and constructing the estimator, we obtain:

$$\psi_i\left(y_i,z_i,x_i;\beta_{10},\widehat{\pi}_I,\widehat{h}_I\right)$$

which substitutes the nuisance $\pi(\mathbf{x}_i)$ and $h(\mathbf{x}_i)$ with the estimated $\widehat{\pi}_I(\mathbf{x}_i)$ and $\widehat{h}_I(\mathbf{x}_i)$. By using the IF above, we can construct a score-like test to test the null hypothesis in (3.10):

$$\begin{aligned}U_{\beta_1}&=\sum_{i\in I}\psi_i\left(y_i,z_i,x_i;\beta_{10},\widehat{\pi}_{I^c},\widehat{h}_{I^c}\right)+\sum_{i\in I^c}\psi_i\left(y_i,z_i,x_i;\beta_{10},\widehat{\pi}_I,\widehat{h}_I\right)\\&=\sum_{i=1}^{n}\psi_i\left(y_i,z_i,x_i;\beta_{10},\widehat{\pi}_i,\widehat{h}_i\right)\end{aligned} \tag{3.25}$$

where

$$\widehat{\pi}_i=\begin{cases}\widehat{\pi}_{I^c}(\mathbf{x}_i) & \text{if } i\in I\\ \widehat{\pi}_I(\mathbf{x}_i) & \text{if } i\in I^c\end{cases},\quad \widehat{h}_i=\begin{cases}\widehat{h}_{I^c}(\mathbf{x}_i) & \text{if } i\in I\\ \widehat{h}_I(\mathbf{x}_i) & \text{if } i\in I^c\end{cases} \tag{3.26}$$

Under the null in (3.10), U_{β_1} has an asymptotic normal distribution:

$$\sqrt{n}U_{\beta_1}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi\left(y_i,z_i,x_i;\beta_{10},\widehat{\pi}_i,\widehat{h}_i\right)\to_d N\left(0,\sigma^2_{\beta_1}\right) \tag{3.27}$$

We can estimate $\sigma^2_{\beta_1}$ by:

$$\begin{aligned}\widehat{\sigma}^2_{\beta_1}&=\frac{1}{n}\sum_{i\in I}\psi_i^2\left(y_i,z_i,x_i;\beta_{10},\widehat{\pi}_{I^c},\widehat{h}_{I^c}\right)+\frac{1}{n}\sum_{i\in I^c}\psi_i^2\left(y_i,z_i,x_i;\beta_{10},\widehat{\pi}_I,\widehat{h}_I\right)\\&=\frac{1}{n}\sum_{i=1}^{n}\psi_i^2\left(y_i,z_i,x_i;\beta_{10},\widehat{\pi}_i,\widehat{h}_i\right)\end{aligned} \tag{3.28}$$

We can use this score-like test statistic along with its asymptotic normal distribution for inference about the null in (3.10).

Although the score-like test performs better for small/moderate samples than its Wald counterpart,

it generally still shows anticonservatisim. By flipping signs of $\psi_i\left(y_i, z_i, x_i; \beta_{10}, \widehat{\pi}, \widehat{h}\right)$ in (3.25), the resulting sign-flipped score-like test may address this issue. We review the sign-flipping score test for parameteric GLM next.

3.4 Sign-flipping Score and Score-like Test

3.4.1 Sign-flipping Score Test for Parametric GLM

We start with a overview of the sign-flipping score test for parametric GLM. More details can be found in Hemerik et al. (2020) [40].

Consider an i.i.d. sample $\mathbf{z}_i = \{y_i, \mathbf{x}_i\}$, where y_i is a response, or dependent variable, and $\mathbf{x}_i$ is a $q \times 1$ vector of independent variables. We assume that $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{iq})^\top$ with the first component of $\mathbf{x}_i$ corresponding to the intercept. A parametric generalized linear model (GLM) has the form:

$$y_i \mid \mathbf{x}_i \sim f(\mu_i, \theta), \quad \mu_i = E(y_i \mid \mathbf{x}_i) = h\left(\beta^\top \mathbf{x}_i\right), \quad 1 \leq i \leq n,$$

where $f(\mu_i, \theta)$ denotes a parametric distribution function with mean μ_i and parameter vector θ, $h^{-1}(\cdot)$ is a link function and β is a $q \times 1$ parameter vector. The parameter vector $\theta = \left(\beta^\top, \eta^\top\right)^\top$ is partitioned into the parameters of interest, β, and the nuisance, η.

The maximum likelihood estimator (MLE) is generally used for Inference about θ. Let U_θ denote the score vector, i.e.,

$$U_\theta^n = \sum_{i=1}^{n} U_\theta^{ni}$$

where U_θ^{ni} is the derivative of the loglikelihood for $\mathbf{z}_i$ with respect to $\theta = \left(\beta^\top, \eta^\top\right)^\top$. If β and η are variationally independent, inference about β can be based on U_β, since the resulting influence function for the MLE of β is orthogonal to the nuisance tangent space of η. For the discussion below, we assume variationally independent β and η so inference about β is based on U_β.

As discussed earlier, inference based on the asymptotic distribution of U_β often exhibits anti-conservatism for small and moderate samples. The sign-flipping score test in Hemerik et al. (2020) [40] provides an effective solution to reduce such bias. We now describe this score test for testing the

null hypothsis involving a subset of independent variables. For notational brevity and without loss of generality, we consider testing the null hypothesis involving one independent scalar variable x_{ip} with $2 \leq p \leq q$:

$$H_0 : \beta_p = \beta_{0p}, \quad H_a : \beta_p \neq \beta_{0p} \tag{3.29}$$

Let $\mathbf{x}_i = \left(\mathbf{x}_{ic}^\top, x_{ip}\right)^\top$ denote a partition of $\mathbf{x}_i$ and $\beta = \left(\beta_c^\top, \beta_p^\top\right)^\top$ denote the corresponding partition of β, where $\mathbf{x}_{ic}$ (β_c) is a $(q-1)\times 1$ column vector. We are interested in testing the null concerning the predictor, x_{ip} of $\mathbf{x}_i$, $H_0 : \beta_p = \beta_{0p}$. If β_c is known, The score statistic is a scalar, $U^n_{\beta_p}$. Under the null H_0, the random sequence, $n^{-1/2}U^n_{\beta_p}$, has an asymptotic normal distribution, which can be used for inference about H_0. For notational brevity, we denote the score test $U^n_{\beta_p}$ as $U^n(\mathbf{z}, \beta_{0p})$ or simply as U^n when no confusion arises.

Let $0 < m \leq 2^n$ be a positive integer. For $1 \leq j \leq m$, let $\mathbf{g}_j = \left(g_{j1}, \ldots, g_{jn}\right)^\top$ be a sequence of $(n \times 1)$ independently distributed random vectors, where $\mathbf{g}_1 = (1, \ldots, 1)^\top$ and $\mathbf{g}_j$ follows the n-dimensional discrete uniform $\{-1, 1\}^n$ for $2 \leq j \leq m$. For each $1 \leq j \leq m$, consider a transformation of the score statistic U by $\mathbf{g}_j$:

$$U^n_j(\mathbf{z}, \beta_{0p}) = n^{-1/2} \sum_{i=1}^{n} g_{ji} U^{ni}(\mathbf{z}, \beta_{0p}), \quad 1 \leq j \leq m \tag{3.30}$$

Thus $U^n_1 = U^n$, while for $2 \leq j \leq m$, U^n_j is a transformed version of U^{ni} with the sign of U^{ni} flipped for at least one i $(1 \leq i \leq n)$.

For any $c \in \mathbb{R}$, let $\lceil c \rceil$ $(\lfloor c \rfloor)$ be the smallest (largest) integer which is at least (most) c. For inference using the finite number of sign-flipped scores U^n_j, let α $(0 < \alpha < 1)$ denote the type I error such as $\alpha = 0.05$. Let $U^n_{(1)} \leq \ldots \leq U^n_{(m)}$ be the ordered values of U^n_j. Let $r = m - \lfloor m\alpha \rfloor$. Let m^+ and m^0 be the number of values $U^n_{(r)}$ greater than and equal to $U^n_{(r)}$, respectively. Let

$$b(\mathbf{z}, \beta_p) = \frac{m\alpha - m^+(\mathbf{z}, \beta_{0p})}{m^0(\mathbf{z}, \beta_{0p})}.$$

Define the rejection rule as:

$$\phi(\mathbf{z}, \beta_{0p}) = \begin{cases} 1 & \text{if } U^n > U^n_{(r)} \\ b(\mathbf{z}, \beta_{0p}) & \text{if } U^n = U^n_{(r)} \\ 0 & \text{if } U^n < U^n_{(r)} \end{cases} . \tag{3.31}$$

Then the test $\phi(\mathbf{z}_p, \beta_{0p})$ in (3.31) is unbiased for randomization hypothesis, i.e., if U^n_j has the same distribution as U^n_1 $(2 \leq j \leq m)$, then $E[\phi(\mathbf{z}_p, \beta_{0p})] = \alpha$ [46]. Within the current context, the U^n_j's do not have the same distribution as U^n_1 and as such the randomization test $\phi(\mathbf{z}_p, \beta_{0p})$ in (3.31) may be biased. However, as the following theorem shows, the U^n_j's all have the same asymptotic distribution and the test based on rejecting the null if $U^n_1 > U^n_{(r)}(\mathbf{z}_p, \beta_{0p})$ is asymptotically conservative.

Theorem 1 *Let* $s_n^2 = \frac{1}{n}\sum_{i=1}^n Var\left(U^{ni}\right)$ *and* $e_n = \frac{1}{n}\sum_{i=1}^n E\left[\left(U^{ni}\right)^2 I\left(n^{\frac{1}{2}}U^{ni} > \varepsilon\right)\right]$. *If*

a) $s_n^2 = \frac{1}{n}\sum_{i=1}^n var(U^{ni}) \to s^2$, *as* $n \to \infty$;

b) $e_n = \frac{1}{n}\sum_{i=1}^n E\left[\left(U^{ni}\right)^2 I\left(n^{\frac{1}{2}}U^{ni} > \varepsilon\right)\right] \to 0$ *for any* $\varepsilon > 0$,

then under H_0, *we have:*

a) U^n_j *are asymptotically normal and independent with mean 0 and common variance* $s^2 = \lim_{n\to\infty} s_n^2$ *under* H_0;

b) $\Pr\left(U^n_1 > U^n_{(r)}\right) \to \lfloor \alpha m \rfloor / m \leq \alpha$.

Thus, although the randomization-like test in (3.31) is not unbiased, it addresses the key issue of anticonservatism for asymptotic tests when applied to relatively small sample sizes. Next we consider the practical case when the coefficients for the covariates are unknown, but estimated.

Consider again the null hypothesis $H_0 : \beta_p = \beta_{0p}$. As the nuisance parameter vector β_c is unknown, it needs to be estimated. Within the context of semiparametric GLM, we can estimate β_c by partitioning the score statistic $\mathbf{U}^{ni}$ $(\mathbf{U}^n)$ into one for the nuisance β_c and one for the predictor of interest β_p. To this end, let

$$\mathbf{U}^{ni} = \begin{pmatrix} \mathbf{U}^{ni}_c \\ U^{ni} \end{pmatrix}, \quad \mathbf{U}^n_c = n^{-1/2}\sum_{i=1}^n \mathbf{U}^{ni}_c, \quad U^n = n^{-1/2}\sum_{i=1}^n U^{ni} \tag{3.32}$$

where $\mathbf{U}^{ni}_c$ (U^n) denotes the derivative of the loglikelihood with respect to β_c (β_p). By setting $\beta_p = \beta_{0p}$

under H_0, we obtain an asymptotically normal estimator $\widehat{\beta}_c$ of β_c by solving the following score equations:

$$\mathbf{U}_c^n(\mathbf{z}, \beta_c, \beta_{0p}) = \mathbf{0}. \tag{3.33}$$

By substituting $\widehat{\beta}_c$ in place of β_c, we denote the score vectors in (3.32) by:

$$\begin{aligned} \widehat{U}^{ni} &= U^{ni}\left(\mathbf{z}_i, \widehat{\beta}_c, \beta_{0p}\right), \quad \widehat{\mathbf{U}}_c^{ni} = \mathbf{U}_c^{ni}\left(\mathbf{z}_i, \widehat{\beta}_c, \beta_{0p}\right) \\ \widehat{U}^n &= U_p^n\left(\mathbf{z}, \widehat{\beta}_c, \beta_{0p}\right) = n^{-1/2}\sum_{i=1}^n \widehat{U}^{ni} \\ \widehat{\mathbf{U}}_c^n &= \mathbf{U}_c^n\left(\mathbf{z}, \widehat{\beta}_c, \beta_{0p}\right) = n^{-1/2}\sum_{i=1}^n \widehat{\mathbf{U}}_c^{ni}. \end{aligned} \tag{3.34}$$

For $1 \le j \le m$, the sign-flipped score U_j^n and $\mathbf{U}_{cj}^n$ are defined by:

$$U_j^n = n^{-1/2}\sum_{i=1}^n g_{ji}U^{ni}, \quad \mathbf{U}_{cj}^n = n^{-1/2}\sum_{i=1}^n g_{ji}\mathbf{U}_c^{ni}. \tag{3.35}$$

The sign-flipping score test is based on the efficient, or effective, score, which is obtained by orthorganlizing U^n with respect to the nuisance tangent space.

Let Σ_U denote the asymptotic variance of the score statistic $\mathbf{U}^n(\beta)$ in (3.32): and a consistent estimator given by:

$$\Sigma_U = E\left[U^{ni}\left(U^{ni}\right)^\top\right]$$

Let $\mathcal{I}_U$ denote the inverse of Σ_U, which is partitioned according to U^n and $\mathbf{U}_c^n$:

$$\mathcal{I}_U = \begin{pmatrix} \mathcal{I}_{pp} & \mathcal{I}_{pc}^\top \\ \mathcal{I}_{pc} & \mathcal{I}_{cc} \end{pmatrix}$$

The efficient version of U^n as:

$$\begin{aligned} U^{n*} &= U^n - \mathcal{I}_{pc}^\top \mathcal{I}_{cc}^{-1}\mathbf{U}_c^n = n^{-1/2}\sum_{i=1}^n U^{ni*} \\ U^{ni*} &= U^{ni} - \mathcal{I}_{pc}^\top \mathcal{I}_{cc}^{-1}\mathbf{U}_c^{ni} \end{aligned} \tag{3.36}$$

Unlike U^n, the asymptotic variance of the efficient U^{n*} does not depend on how β_c is estimated by:

$$\widehat{\Sigma}_{U^*} = \frac{1}{n}\sum_{i=1}^{n} \widehat{U}^{ni*}\left(\widehat{U}^{ni*}\right)^\top = \frac{1}{n}\sum_{i=1}^{n} U^{ni*}\left(\mathbf{z}_i, \widehat{\beta}_c, \beta_{0p}\right)\left(U^{ni*}\left(\mathbf{z}_i, \widehat{\beta}_c, \beta_{0p}\right)\right)^\top$$

For $1 \leq j \leq m$, the sign-flipped efficient score statistics $\widehat{U}_j^{n*}$ with an estimated β_c are defined by:

$$\widehat{U}_j^{n*} = n^{-1/2}\sum_{i=1}^{n} g_{ji}\widehat{U}^{ni*}, \quad \widehat{U}^{ni*} = \widehat{U}^{ni} - \widehat{\mathcal{I}}_{pc}^\top \widehat{\mathcal{I}}_{cc}^{-1}\widehat{\mathbf{U}}_c^{ni}. \tag{3.37}$$

As in the case with known β_c, let $\widehat{U}_{(1)}^{n*} \leq \ldots \leq \widehat{U}_{(m)}^{n*}$ be the ordered values of $\widehat{U}_j^{n*}$. Then the $\widehat{U}_j^{n*}$'s all have the same asymptotic distribution and the test based on rejecting the null if $\widehat{U}_1^{n*} > \widehat{U}_{(r)}^{n*}(\mathbf{z}_p, \beta_{0p})$ is asymptotically conservative.

Theorem 2 *Consider the efficient sign-flipped score statistics $\widehat{U}_j^{n*}$ above $(1 \leq j \leq m)$. Under H_0, the rejection probability* $\Pr\left(\widehat{U}_1^{n*} > \widehat{U}_{(r)}^{n*}\right) \to \lfloor \alpha w \rfloor / m \leq \alpha$, *where $\widehat{U}_{(r)}^{n*}$ denotes the ordered version of $\widehat{U}_j^{n*}$ $(1 \leq j \leq m)$.*

3.4.2 Sign-flipping Score-like Test for Average Treatment Effect

We now consider extending the sign-flipping score test for parametric models in the proceeding section to semi- and non-parametric DREs discussed in Section 3.3.3.

Let $\widehat{\pi}$ and $\widehat{h}$ be an estimated nuisance parameter for semi- or non-parametric π and h. Note that for non-parametric π and h, $\widehat{\pi}$ and $\widehat{h}$ are defined in (3.26) based on splitted sample. Then the score-like test statistic $\widehat{U}_{\beta_1}$ in (3.25) has an asymptotic normal distribution in (3.27) for both semi- and non-parametricaly specified π and h, with the asymptotic variance given in (3.27) and consistently estimated as in (3.28).

To construct a sign-flipping score-like test based on the score-like test in (3.25), let $0 < m \leq 2^n$ be a positive integer and for $1 \leq j \leq m$ let $\mathbf{g}_j = (g_{j1}, \ldots, g_{jn})^\top$ be defined the same as in Section 3.4.1. The sign-flipped efficient score-like test statistics $\widehat{U}_{\beta_1,j}^{n*}$ are defined by:

$$\widehat{U}_{\beta_1,j}^{n*} = n^{-1/2}\sum_{i}^{n} g_{ji}\boldsymbol{\psi}\left(y_i, z_i, x_i; \beta_{10}, \widehat{\pi}, \widehat{h}\right) \tag{3.38}$$

Given $\widehat{U}_{\beta_1}$'s asymptotic properties and orthogonality with respect to the nuisance parameters π and h, we can readily extend Theorem 2 to the semi- or non-parametrically estimated π and h.

Corollary 3 *Consider the orthogonalized sign-flipped score-like statistics* $\widehat{U}^{n*}_{\beta_1,j}$*. Under the null hypothesis* H_0*, the rejection probability* $\Pr\left(\widehat{U}^{n*}_{\beta_1,1} > \widehat{U}^{n*}_{\beta_1,(r)}\right) \to \lfloor \alpha w \rfloor / m \leq \alpha$*, where* $\widehat{U}^{n*}_{\beta_1,(r)}$ *denotes the ordered version of* $\widehat{U}^{n*}_{\beta_1,j}$ $(1 \leq j \leq m)$.

3.5 Applications

We illustrate the proposed approach with both real and simulated data. We start with simulation study to help assess the performance of the proposed approach in addressing anticonservative behavior of type I errors for small and moderate samples. In particular we considered simulating the datasets for learning causal effects.

3.5.1 Simulation Study

The covariate $X \in \mathbb{R}$ was drawn from a normal distribution with zero mean and $var(X) = 1$. The treatment condition Z was generated from Bernoulli distribution with probability π_i, where π_i is calculated from a logistic regression based on X:

$$Z_i \sim Bernoulli\left(\pi_i\right), \quad \text{logit}\left(\pi_i\right) = \gamma_0 + \gamma_1 X_i$$

The potential continuous outcome Y of the ith subject satisfied

$$y_{i,0} = \beta_0 + \beta_2 x_i + \varepsilon_i, \quad y_{i,1} = \beta_0 + \beta_1 + \beta_2 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0,1) \quad 1 \leq i \leq n$$

$$y_i = \begin{cases} y_{i,0} & \text{if } z_i = 0 \\ y_{i,1} & \text{if } z_i = 1 \end{cases} \qquad 1 \leq i \leq n$$

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 x_i + \varepsilon_i, \quad 1 \leq i \leq n$$

As we discussed in Section 3.3, β_1 is the ATE that we are interested in. We use the simulated datasets to test the hypothesis $H_0 : \beta_1 = 0$ v.s. $H_1 : \beta_1 \neq 0$. For the simulation study, we set the nuisance parameters $\beta_0 = 1$, $\beta_2 = 0$, $\gamma_0 = 1$, $\gamma_1 = 1$, and generate $z_1, \cdots, z_n, y_1, \cdots, y_n$ according to the model given above.

We consider six tests to help assess the performance of the proposed approach: (a) outcome regression; (b) inverse probability weighting score-like test; (c) doubly robust score test; (d) sign-flipping doubly robust score test (use linear regression to solve nuisance parameters); (e) sign-flipping doubly robust score test (use random forest to solve nuisance parameters); (f) sign-flipping inverse probability weighting score-like test.

The sample sizes are set to $n = 50$, 100, 500, and 1000 to study performance of the proposed approach for small, medium, and large samples. The true type I error α is set to 0.05. For the sign-flipping tests, i.e., tests (d), (e), (f) above, we set the Monte Carlo sampling number $m = 1000$.

3.5.2 Simulation Results

Shown in Table 3.1 are the empirical type I error. For the sign-flipping doubly robust score test, the empirical type I errors were very close to $\alpha = 0.05$ even for small sample size (e.g., $n = 50$), either using linear regression or random forest to estimate the nuisance parameters; and it performed better than the doubly robust score test. For the outcome regression, since we included the covariate X when fitting the model, the empirical type I errors are also very close to $\alpha = 0.05$. The tests using inverse probability weighted score-like statistics gave biased empirical type I errors; when sample sizes are large, applying sign-flipping to inverse probability weighted score-like statistics resulted in unbiased empirical type I error.

Table 3.1. Simulation results for comparing (a) outcome regression; (b) inverse probability weighting score-like test; (c) doubly robust score test; (d) sign-flipping doubly robust score test (use linear regression to solve nuisance parameters); (e) sign-flipping doubly robust score test (use random forest to solve nuisance parameters); (f) sign-flipping inverse probability weighting score-like test.

	$n = 50$	$n = 100$	$n = 500$	$n = 1000$
Doubly robust score test	0.041	0.064	0.059	0.06
Outcome regression	0.044	0.046	0.049	0.051
IPW score-like test	0.157	0.118	0.079	0.072
sign-flip doubly robust score test(lm)	0.047	0.05	0.049	0.051
sign-flip doubly robust score test(random forest)	0.045	0.044	0.049	0.055
sign-flip IPW score-like test	0.143	0.105	0.051	0.047

3.6 Discussion

We proposed a new sign-flipping score-like test for inference about the average treatment effect (ATE) based on the doubly robust estimator approach (DRE) to improve anticonservatism with small and moderate samples. Like its components of semiparametric ORM and MSM, DRE has nice asymptotic properties, but its performance for small and moderate samples is still not good with anticonservative type I error, or poor coverage probability, especially when there is heteroscedasticity for continuous and overdispersion for count responses. Although score-like tests generally perform better than Wald tests, they still exhibits anticonservative bias in type I error, or coverage probability. The proposed sign-flipping score-like test provides a more effective approach to address such bias in small and moderate samples.

In addition to the semiparametric DRE, the sign-flipping score-like test also applies to the DRE with non-parametric models for covariates. In the ORM, the linear predictor has a partial linear form with a parametric specification for the predictors of interest (ATE) and non-parametric specification for the nuisance covariates. In MSM, the nuisance covariates are modeled non-parametrically for the propensity score. The non-parametric models for the covariates are generally estimated using machine learning methods such as random forest. By reconstructing a sign-flipping score-like test using sample splitting, which has been used in addressing sub-$\sqrt{n}$ convergence rates of non-parametric nuisance estimators in recent studies, the reconstructed sign-flipping score-like test statistic again has nice asymptotic properties. Moreover, it performed very well in our limited simulation studies.

Chapter 3, in full, is currently being prepared for submission for publication of the material as it may appear in "Chenyu Liu, Jinyuan Liu, Tuo Lin, Yuqi Qiu, Lin Liu and Xin Tu. *Sign-flipped Doubly Robust Score-like Test for Average Treatment Effect in Causal Inference*". The dissertation author was the primary author on this paper.

Chapter 4

A Multiple Imputation Approach for Longitudinal Regression with Missing Response

4.1 Abstract

Missing data is a common issue in clinical research, often resulting from loss to follow-up, data collection errors, or unforeseen events. This issue complicates statistical analysis since assumptions about the distribution of unobserved data are necessary. Missing data mechanisms include Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR), with MAR being the most commonly assumed in clinical trials. Multiple imputation (MI) is a widely used approach to handle missing data, generating plausible values based on observed data distributions and relationships. These imputed datasets are analyzed separately, and results are pooled to account for uncertainty. The properties of MI estimators, such as consistency and asymptotic normality, are crucial for reliable inference. However, many imputation methods lack well-established large sample properties. Tsiatis (2006) examined these properties for two popular imputation schemes, type A and B, finding both to be consistent and asymptotically normal, though their practical application to real-world data remains challenging. We propose a type A MI estimator using predictive distribution sampling via software like the MICE package in R. This approach allows the response's missingness to depend on variables outside the regression model, enhancing the MAR assumption's validity. We show that the MI estimator follows an asymptotic normal distribution with variance estimated by Rubin's formula. We

demonstrate the method with real and simulated data, showing easy implementation using MICE in R. Applying this to the PEP4PA study, we assess whether the intervention group significantly increased their moderate-to-vigorous physical activity minutes after the baseline visit.

KEY WORDS: Missing data, Multiple imputation, MICE, Longitudinal data, asymptotic properties, Rubin's variance formula

4.2 Introduction

Missing data is a common occurrence in clinical research. Loss to follow-up, data collection errors, or intercurrent events due to various reasons in clinical trials is unintended but inevitable so that we often cannot measure what we intended to for all participants [47, 48]. The outcomes could be unattainable due to the type of deviation (e.g., missed patient visit) [49] and, depending on the nature of the estimand and analysis, even the values that were recorded post deviation may be better considered as missing [50].

The presence of missing data creates complexity, since any statistical analysis necessarily makes an untestable assumption about the distribution of the unobserved [51]. There are three types of missing data mechanisms in terms of the impact of missing data on statistical inference [52]: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). When data are MCAR, whether the data are missing is independent of the observed and unobserved data [53]. When data are MAR, missing data is systematically related to the observed but not the unobserved data. When data are MNAR, whether the data is missing is systematically related to the unobserved data [51], that is, the missingness is related to events or factors which are not measured in the study. Typically, the primary analysis of a clinical trial is conducted under the assumption of MAR [54]. MCAR assumption is less common in real-life datasets, which occurs when the missingness is truly random and unrelated to any other variables in the study [55]. MNAR requires external information in addition to the observed data and as such is not addressable unless such information is collected [56], in which case MNAR becomes MAR.

A widely used approach for addressing missing data for parametric models is multiple imputation (MI) [57]. Multiple imputation fills in missing values by generating plausible numbers derived from distributions of and relationships among observed variables [58, 59]. Each imputed complete data is then analyzed and analysis results are pooled to derive inference about parameters of interest [60]. In this way,

MI allows the user to explicitly incorporate the uncertainty about the true value of the parameters based on the analysis of the imputed variables [61, 62].

When using MI to impute missing data, we are more interested in the properties of MI estimators, such as consistency and asymptotic normality. Different imputation methods have been proposed in the statistical and applied literatures [63, 64, 65]. However, many of these procedures have unknown large sample properties. Tsiatis (2006) [43] investigated such large sample properties for two popular imputation schemes, known as type A and B in the statistical literature [66]. Although both types of MI estimators have been shown to yield consistent and asymptotically normal estimators [67, 68, 69], neither has straight-forward application to real study data. The difficulty lies in obtaining imputed data that satisfy the requirement for the large sample properties of the two estimators [70].

In this paper, we will propose an approach to facilitate applications of type A estimator by leveraging imputation methods available in popular statistical packages such as the `mice` package in `R` [63]. In Section 2, we discuss the multiple imputation for parametric regression models for longitudinal data. In Section 3, we use simulation studies and the real dataset to illustrate the approach. In Section 4, we give our concluding remarks.

4.3 Multiple Imputation for Parametric Regression Models for Longitudinal Data

4.3.1 Data Generating Process and Missing Data

Consider an i.i.d. sample of size n with observed-data value $\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i); 1 \leq i \leq n\}$, where Z_i is a vector of variables, $\mathcal{C}_i$ is a categorical variable for missing patterns and $G_{\mathcal{C}_i}(Z_i)$ denotes the subvector of variables with observed data. We let $\mathcal{C}_i = \infty$ indicate fully observed Z_i, i.e., $G_\infty(Z_i) = Z_i$, and $\mathcal{C}_i \neq \infty$ for missing data in Z_i with $G_{\mathcal{C}_i}(Z_i)$ denoting the variables with observed data. Let Z_i be distributed as:

$$Z_i \sim p_Z(\mathbf{z}, \theta), \quad i = 1, \ldots, n \tag{4.1}$$

where $p_Z(\mathbf{z}, \theta)$ denotes a parametric model parameterized by θ.

Under MAR, the observed $G_{\mathcal{C}_i}(Z_i)$ has the distribution:

$$Z_i \sim p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}\left(\mathbf{z} \mid \mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \theta\right), \quad i = 1, \ldots, n$$

The observed-data likelihood can be used to obtain the MLE of θ by maximizing

$$\prod_{i=1}^{n} p_{G_{r_i}(Z_i)}(g_{r_i}; \theta), \quad p_{G_r(Z)}(g_r; \theta) = \int_{\{\mathbf{z}: G_r(\mathbf{z}) = g_r\}} p_Z(\mathbf{z}; \theta) d\mathbf{z}.$$

Denoting the full-data score vector by:

$$S^F(Z_i, \theta) = \frac{\partial \log p_Z(Z_i, \theta)}{\partial \theta}$$

Then the observed data score $S(\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i); \theta)$ is related to $S^F(Z_i, \theta)$ by:

$$S(\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i); \theta) = E\left[S^F(Z_i; \theta) \mid \mathcal{C}_i, G_{\mathcal{C}_i}(Z_i); \theta\right]$$

By the law of conditional variance,

$$\begin{aligned} \operatorname{var}\left(S^F(Z_i)\right) &= \operatorname{var}\left[E\left(S^F(Z_i) \mid \mathcal{C}_i, G_{\mathcal{C}}(Z_i)\right)\right] + E\left[\operatorname{var}\left(S^F(Z_i) \mid \mathcal{C}_i, G_{\mathcal{C}}(Z_i)\right)\right] \\ &= \operatorname{var}\left(S(\mathcal{C}_i, G_{\mathcal{C}}(Z_i))\right) + E\left[\operatorname{var}\left(S^F(Z_i) \mid \mathcal{C}_i, G_{\mathcal{C}}(Z_i)\right)\right] \end{aligned}$$

it follows that

$$E\left[\operatorname{var}\left(S^F(Z_i) \mid \mathcal{C}_i, G_{\mathcal{C}}(Z_i)\right)\right] = \operatorname{var}\left(S^F(Z_i)\right) - \operatorname{var}\left(S(\mathcal{C}_i, G_{\mathcal{C}}(Z_i))\right) = I^F(\theta) - I(\theta) \tag{4.2}$$

where $I^F(\theta)$ and $I(\theta)$ are the Fisher information for the full and observed data.

4.3.2 Multiple Imputation for Missing Data

Consider a longitudinal study with n subjects and m assessment times. For regression analysis for longitudinal data, let y_{it} and $\mathbf{x}_{it}$ denote the response and vector of explanatory (independent) variables from the ith subject at time t ($1 \le i \le n$, $1 \le t \le m$). Let $Z_i = (y_{i1}, \ldots, y_{im})^\top$ and $X_i = \left(\mathbf{x}_{i1}^\top, \ldots, \mathbf{x}_{im}^\top\right)^\top$. In

regression analysis for longitudinal data, we model the conditional distribution of Z_i given X_i. Further, we assume that missing data only occurs to the responses y_{it}. Thus we can view the parametric model in (4.1) as the regression model for the conditional distribution of Z_i given X_i. However, the parameter θ is partitioned into β and η with β being of primary interest such as regression coefficients and η being nuisance parameters such as variances of random effects and error terms in linear mixed-effects models.

Multiple imputation is a general approach to the problem of missing data, which allows for the uncertainty about the missing data by creating several different and plausible imputed data sets and appropriately combining analysis results obtained from each of them [60]. Generally, the multiple imputation procedure can be broken down into three stages.

1. Imputation. Missing values are imputed following some imputation approach. This is done m times independently to create m completed datasets.

2. Analysis. Each of the m datasets is analyzed, using statistical methods for complete data to fit the model of interest with imputed data.

3. Pooling. Analysis results from m imputed datasets are combined to give a multiple imputation estimate for inference.

Different approaches have been proposed in the literature for the imputation step.

Under MAR, multivariate imputation by chained equations (MICE) is a popular multiple imputation technique and readily implemented with an `R` package called "`mice`" [63]. Let $R_{\mathcal{C}_i}(Z_i)$ denote the subvector Z_i with missing data for each subject. Then Z_i consists of the observed $G_{\mathcal{C}_i}(Z_i)$ and missing $R_{\mathcal{C}_i}(Z_i)$. For notational brevity, let $Z_{i,obs} = G_{\mathcal{C}_i}(Z_i)$ and $Z_{i,mis} = R_{\mathcal{C}_i}(Z_i)$. Then $Z_i = \left(Z_{i,obs}^\top, Z_{i,mis}^\top\right)^\top$. Under MICE, the parameters θ are partitioned into multiple $\theta_1, \cdots, \theta_p$ that are specific to the respective conditional densities of the parameters or $Z_{i,mis}$ given the observed $Z_{i,obs}$. Starting from a simple draw from observed marginal distributions, the lth iteration of chained equations is a Gibbs sampler that successively draws from the following conditional distributions:

$$\theta_k^{(l)} \sim p\left(\theta_1 \mid Z_{i,obs}, Z_{i,mis}^{(l-1)}\right), \quad Z_{i,mis}^{(l)} \sim p\left(Z_{i,mis}^{(l)} \mid Z_{i,obs}, Z_{i,mis}^{(l-1)}, \theta_k^{(l)}\right),$$
$$1 \leq k \leq p$$

where $Z_i^{(l)} = \left(Z_{i,obs}^{\top}, \left(Z_{i,mis}^{(l-1)} \right)^{\top} \right)^{\top}$ denotes the complete imputed data for Z_i at the lth iteration. Previous imputations $Z_{i,mis}^{(l-1)}$ only enter $Z_{i,mis}^{(l)}$ through its relation with other variables, and not directly. The sampling process continues until convergence. The completed $Z_i^{(l)}$'s through MICE imputation at convergence are sampled from their predictive distributions based on the observed data $Z_{i,obs}$. We can repeat this process m times to generate m imputed full-data sets.

If θ_0 is known (as in simulation study), we can view the imputed data $Z_i(\theta_0)$ from a single imputation as being generated from the following steps:

a) Generate $\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$ from the distribution with density $p_{\mathcal{C},G_{\mathcal{C}}(\mathbf{Z})}(r, g_r; \theta_0)$;

b) Generate $Z_i(\theta_0)$ from the conditional distribution of Z given $\mathcal{C}, G_{\mathcal{C}}(Z)$ with the conditional density:

$$p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)} \{\mathbf{z} \mid \mathcal{C}_i, G_{\mathcal{C}_i}(Z_i); \theta_0\}.$$

We then have a full data for all subjects and obtain the MLE $\widehat{\theta}_n$ by solving the full-data likelihood equation:

$$\sum_{i=1}^{n} S^F \left(\mathbf{Z}_i(\theta_0); \widehat{\theta}_n \right) = \mathbf{0} \tag{4.3}$$

Under mild regularity conditions, the MLE $\widehat{\theta}_n$ is consistent and asymptotically normal:

$$\sqrt{n} \left(\widehat{\theta}_n - \theta_0 \right) \to_d N(0, \Sigma)$$

where the asymptotic variance Σ and a consistent estimator are given by:

$$\Sigma = I_1^{-1}, \quad I_1 = E \left(-\frac{\partial S^F(Z_i, \theta_0)}{\partial \beta^{\top}} \right)$$

$$\widehat{\Sigma} = \widehat{I}_1 = \frac{1}{n} \widehat{I}_n, \quad \widehat{I}_n = -\sum_{i=1}^{n} \frac{\partial S^F \left(Z_i, \widehat{\theta}_n \right)}{\partial \beta^{\top}}$$

where I_1 is the Fisher information.

In practice, θ_0 is unknown. Large sample properties of MI estimators depend critically on how the missing data is imputed. To ensure consistency and asymptotic normality, imputed values for the missing $R_{\mathcal{C}_i}(Z_i)$ must satisfy certain requirements. Following [43], we consider two imputation schemes

and the resulting MI estimators that are consistent and asymptotically normal.

4.3.3 Multiple Imputation Estimators

Frequentist Type B Estimator

If θ_0 is unknown and a consistent estimator, $\widehat{\theta}_n^I$, of θ_0 is available, we can substitute $\widehat{\theta}_n^I$ in place of θ_0 and generate $\mathbf{Z}_i\left(\widehat{\theta}_n^I\right)$ for each missing data $\{\mathcal{C}_i \neq \infty, G_{\mathcal{C}_i}(\mathbf{Z}_i)\}$ m times to obtain m imputed full-data sets:

$$Z_{ij} = \begin{cases} Z_{ij}(\theta_0) & \text{if } \mathcal{C}_i = \infty \\ Z_{ij}\left(\widehat{\theta}_n^I\right) & \text{if } \mathcal{C}_i \neq \infty \end{cases}, \quad 1 \leq i \leq n, \quad 1 \leq j \leq m \tag{4.4}$$

For each j-th set of imputed full data, $\left\{Z_{ij}, 1 \leq i \leq n, 1 \leq j \leq m\right\}$, we obtain the j-th estimator $\widehat{\theta}_{nj}^*$ by solving the full-data score equations:

$$\sum_{i=1}^{n} S^F\left(Z_{ij}\left(\widehat{\theta}_n^I\right); \widehat{\theta}_{nj}^*\right) = \mathbf{0} \tag{4.5}$$

The multiple-imputation estimator is for the m imputed full-data sets is:

$$\widehat{\theta}_n^* = m^{-1} \sum_{j=1}^{m} \widehat{\theta}_{nj}^* \tag{4.6}$$

Under mild regularity conditions, the MI estimator $\widehat{\theta}_n^*$ is consistent and asymptotically normal [71]:

$$\sqrt{n}\left(\widehat{\theta}_n^* - \theta_0\right) \rightarrow_d N(0, \Sigma^*)$$

where the asymptotic variance Σ^* is given by:

$$\Sigma^* = \left(I^F\right)^{-1}\left[I + m^{-1}\left(I^F - I\right) + \left(I^F - I\right)\text{var}\left(q\left(\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right)\right)\left(I^F - I\right) + 2\left(I^F - I\right)\right]\left(I^F\right)^{-1} \tag{4.7}$$

where I^F and I denote the full-data and observed-data Fisher information and $q\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)$ denotes the influence function of the estimator $\widehat{\theta}_n^I$:

$$n^{1/2}\left(\widehat{\theta}_n^I - \theta_0\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} q\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right) + \mathbf{o}_p\left(1\right)$$

The asymptotic variance Σ^* can be consistently estimated [43]. However, the popular variance estimator proposed by [66] below is not a consistent estimator of Σ^*:

$$\widehat{\Sigma}^* = m^{-1} \sum_{j=1}^{m} \left\{ \frac{1}{n} \sum_{i=1}^{n} -\frac{\partial S^F\left(Z_{ij}\left(\widehat{\theta}_n^I\right), \widehat{\theta}_{nj}^*\right)}{\partial \theta^\top} \right\}^{-1} + \left(\frac{m+1}{m}\right) n \frac{\sum_{j=1}^{m}\left(\widehat{\theta}_{nj}^* - \widehat{\theta}_n^*\right)\left(\widehat{\theta}_{nj}^* - \widehat{\theta}_n^*\right)^\top}{m-1} \tag{4.8}$$

Furthermore, this type B MI estimator requires a consistent estimator $\widehat{\theta}_n^I$ to ensure consistency and asymptotic normality of the MI estimator. In practice, it is difficult to obtain such a consistent estimator.

For example, if we model longitudinal data under MAR using the linear mixed-effects model, we can obtain a consistent estimator of β of the regression coefficients using semiparametric linear regression such as weighted generalized estimating equations (WGEE). However, WGEE does not provide any estimator of the nuisance parameters η. Although we may model the second moments of Z_i using semiparametric WGEE II [72], popular statistical packages such as R do not support WGEE II.

Bayesian Type A Estimator

A popular alternative to sampling Z_{ij} from the statistical model with an consistent estimator $\widehat{\theta}_n^I$ is to generate such values from the predictive distribution ($1 \leq i \leq n$, $1 \leq j$) [66]:

$$p_{Z|\mathcal{C},G_{\mathcal{C}}(\mathbf{Z})}\left(\mathbf{z} \mid \mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right) \tag{4.9}$$

In this case, we can view the generated Z_{ij} as through the following sampling process:

a) Treat θ as a random quantity and specify a prior distribution for θ, from which we obtain the *posterior* distribution:

$$p_{\theta|\mathcal{C},G_{\mathcal{C}}(Z)}\left(\theta \mid \mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right) = \int p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}\left(\mathbf{z} \mid \mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right); \theta\right) p_{\beta|\mathcal{C},G_{\mathcal{C}}(Z)}\left(\theta \mid \mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right) d\theta \tag{4.10}$$

b) Randomly select $\theta^{(j)}$ from the posterior distribution in (4.10);

c) Conditional on $\theta^{(j)}$, we sample from the conditional predictive distribution given $\theta^{(j)}$:

$$p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}\left(\mathbf{z} \mid \mathcal{C}_i, G_{\mathcal{C}_i}(Z_i); \theta^{(j)}\right) \tag{4.11}$$

to obtain $Z_{ij}\left(\theta^{(j)}\right)$.

By repeating (b) and (c) m times, we obtain m imputed full-data sets $\left\{Z_{ij}, 1 \le i \le n, 1 \le j \le m\right\}$

For each jth set of imputed full-data set, $\left\{Z_{ij}\left(\theta^{(j)}\right), 1 \le i \le n\right\}$, we solve the full-data score equations as in the case of Type B estimator, but with a differently imputed full-data set $\left\{Z_{ij}\left(\theta^{(j)}\right), 1 \le i \le n\right\}$:

$$\sum_{i=1}^{n} S^F\left(Z_{ij}\left(\theta^{(j)}\right); \widehat{\theta}^*_{nj}\right) = \mathbf{0} \tag{4.12}$$

The multiple-imputation estimator m imputed full-data sets is again given by averaging the m estimators $\widehat{\theta}^*_{nj}$ in (4.12) obtained from the m imputed full-data sets.

Rubin refers to imputation with the values for imputing missing Z_{ij} drawn from the predictive distribution in (4.9) as type A estimator. The imputed data from MICE are generated this way. Like the Type B estimator, Type A estimator is also consistent and asymptotically normal. However, unlike Type B estimator, this Type A estimator is readily implemented by leveraging available software such as the MICE package in R [63]. Thus we focus on the large sample properties for Type A estimator.

4.3.4 Large Sample Properties of Type A Estimator

In Bayesian imputation, the jth imputation is obtained by sampling from the predictive distribution in (4.9), which can be viewed as first drawing $\theta^{(j)}$ from the posterior distribution in (4.10) and then sampling Z_{ij} from the conditional predictive distribution in (4.11). For lage sample size n, the posterior distribution of θ and the sampling distribution of the estimator are closely approximated by each other. The initial estimator $\widehat{\theta}^I_n$ was assumed to be asymptotically normal:

$$n^{1/2}\left(\widehat{\theta}^I_n - \theta_0\right) \sim N\left(\mathbf{0}, \mathrm{var}\left(q\left(\mathcal{C}, G_{\mathcal{C}}(Z)\right)\right)\right)$$

Therefore, mimicking the idea of Bayesian imputation, instead of fixing the values $\widehat{\theta}_n^I$ for each of the m imputations, at the jth imputation, we sample $\theta^{(j)}$ from

$$N\left(\widehat{\theta}_n^I, \frac{\widehat{\text{var}}\left(q\left(\mathcal{C}, G_{\mathcal{C}}(Z)\right)\right)}{n}\right),$$

where $q\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)$ denotes the influence function of $\widehat{\theta}_n^I$ and $\widehat{\text{var}}\left(q\left(\mathcal{C}, G_{\mathcal{C}}(Z)\right)\right)$ a consistent estimator for the asymptotic variance of $\widehat{\theta}_n^I$.

By a Taylor series expansion, we have:

$$n^{1/2}\left(\widehat{\theta}_n^* - \theta_0\right) = n^{-1/2}\sum_{i=1}^{n}\left(I^F\left(\theta_0\right)\right)^{-1}\left[m^{-1}\sum_{j=1}^{m} S^F\left(Z_{ij}\left(\theta^{(j)}\right), \theta_0\right)\right] + \mathbf{o}_p(1) \tag{4.13}$$

Since $m^{-1}\sum_{j=1}^{m} S^F\left(Z_{ij}\left(\theta^{(j)}\right), \theta_0\right)$ is a dependent sequence with $\theta^{(j)}$,
$\left(I^F\left(\theta_0\right)\right)^{-1}\left[m^{-1}\sum_{j=1}^{m} S^F\left(Z_{ij}\left(\theta^{(j)}\right), \theta_0\right)\right]$ is not the influence function. To find the asymptotic variance, we need to find the influence function.

To this end, let

$$\begin{aligned}
U_n &= n^{-1/2}\sum_{i=1}^{n}\left[m^{-1}\sum_{j=1}^{m} S^F\left(Z_{ij}\left(\theta_0\right), \theta_0\right)\right] \\
V_n &= n^{-1/2}\left(I^F\left(\theta_0\right)\right)^{-1}\sum_{i=1}^{n}\left[m^{-1}\sum_{j=1}^{m} S^F\left(Z_{ij}\left(Z_{ij}\left(\theta^{(j)}\right)\right), \theta_0\right) - m^{-1}\sum_{j=1}^{m} S^F\left(Z_{ij}\left(\theta_0\right), \theta_0\right)\right]
\end{aligned}$$

We can express (4.13) as:

$$n^{1/2}\left(\widehat{\theta}_n^* - \theta_0\right) = \left(I^F\left(\theta_0\right)\right)^{-1}\left(U_n + V_n\right) + \mathbf{o}_p(1) + \mathbf{o}_p(1) \tag{4.14}$$

The first term U_n is made up of a sum of n i.i.d. elements and thus here the ith element of the sum is equal to

$$G_i\left(m, \theta_0\right) = m^{-1}\sum_{j=1}^{m} S^F\left(Z_{ij}\left(\theta_0\right), \theta_0\right), \quad i = 1, \ldots, n. \tag{4.15}$$

For the second term V_n, we show in the Appendix that

$$V_n = \left(I^F - I\right) m^{-1} \sum_{j=1}^{m} n^{1/2} \left(\theta^{(j)} - \widehat{\theta}_n^I\right) + \left(I^F - I\right) n^{-1/2} \sum_{i=1}^{n} q\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right), \theta_0\right) + \mathbf{o}_p\left(1\right) \tag{4.16}$$

Since the m random terms $n^{1/2}\left(\theta^{(j)} - \widehat{\theta}_n^I\right)$ $(1 \leq j \leq m)$ are asymptotically equivalent to m i.i.d. normal random variables with mean zero and variance $\operatorname{var}\left(q\left(\mathcal{C}, G_{\mathcal{C}}(Z)\right)\right)$, $V_1, \ldots, V_m$, we can express (4.16) as:

$$V_n = \left(I^F - I\right) m^{-1} \sum_{j=1}^{m} V_j + \left(I^F - I\right) n^{-1/2} \sum_{i=1}^{n} q\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right), \theta_0\right) + \mathbf{o}_p\left(1\right) \tag{4.17}$$

By (4.14), (4.15) and (4.17), we can express (4.13) as:

$$\begin{aligned} n^{1/2}\left(\widehat{\theta}_n^* - \theta_0\right) &= n^{-1/2} \sum_{i=1}^{n} \left(I^F\right)^{-1} \left[\left(m^{-1} \sum_{j=1}^{m} S^F\left(Z_{ij}\left(\theta_0\right), \theta_0\right)\right) + \left(I^F - I\right) q\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right), \theta_0\right)\right] + \\ &+ \left(I^F\right)^{-1} m^{-1} \sum_{j=1}^{m} \left(I^F - I\right) V_j + \mathbf{o}_p\left(1\right) \\ &= n^{-1/2} \sum_{i=1}^{n} \left(I^F\right)^{-1} \left[\left(m^{-1} \sum_{j=1}^{m} S^F\left(Z_{ij}\left(\theta_0\right), \theta_0\right)\right) + \left(I^F - I\right) q\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right), \theta_0\right)\right] + \\ &+ n^{-1/2} \sum_{i=1}^{n} \left(n^{1/2} \left(I^F\right)^{-1} m^{-1} \sum_{j=1}^{m} \left(I^F - I\right) V_j\right) + \mathbf{o}_p\left(1\right) \end{aligned} \tag{4.18}$$

Thus the influence function $\psi_i\left(\theta_0\right)$ for $n^{1/2}\left(\widehat{\theta}_n^* - \theta_0\right)$ is given by:

$$\begin{aligned} \psi_i\left(Z_i^m\left(\theta_0\right); \theta_0\right) = \left(I^F\right)^{-1} &\left[\left(m^{-1} \sum_{j=1}^{m} S^F\left(Z_{ij}\left(\theta_0\right), \theta_0\right)\right) + \left(I^F - I\right) q\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right] + \\ &+ \left(n^{1/2} \left(I^F\right)^{-1} m^{-1} \sum_{j=1}^{m} \left(I^F - I\right) V_j\right) \end{aligned}$$

where $Z_i^m\left(\theta_0\right) = \left\{Z_{ij}\left(\theta_0\right); 1 \leq j \leq m\right\}$ denotes the observed and pseudo-imputed data $Z_{ij}\left(\theta_0\right)$, i.e., $Z_{ij}\left(\theta_0\right)$ is drawn from the true data-generating process.

The asymptotic variance Σ_θ of the multiple imputation estimator $n^{1/2}\left(\widehat{\theta}_n^* - \theta_0\right)$ is the variance of the influence function $\psi_i\left(Z_i^m\left(\theta_0\right); \theta_0\right)$, i.e., $\Sigma = \operatorname{var}\left(\psi_i\left(Z_i^m\left(\theta_0\right); \theta_0\right)\right)$. In the Appendix, we show that

$\text{var}\left(\psi_i\left(Z_i^m\left(\theta_0\right);\theta_0\right)\right)$ is given by:

$$\begin{aligned}
\Sigma_\theta &= \Sigma_1 + \Sigma_2 \qquad (4.19)\\
\Sigma_1 &= \left(I^F\left(\theta_0\right)\right)^{-1} \text{var}\left[S^F\left(Z_i\left(\theta_0\right),\theta_0\right) + \left(I^F\left(\theta_0\right) - I\left(\theta_0\right)\right) q\left(C_i, G_{C_i}\left(Z_i\right),\theta_0\right)\right] \left(I^F\left(\theta_0\right)\right)^{-1}\\
\Sigma_2 &= m^{-1}\left(I^F\left(\theta_0\right)\right)^{-1}\left(I^F\left(\theta_0\right) - I\left(\theta_0\right)\right) \text{var}\left(q\left(C_i, G_{C_i}\left(Z_i\right),\theta_0\right)\right)\left(I^F\left(\theta_0\right) - I\left(\theta_0\right)\right)\left(I^F\left(\theta_0\right)\right)^{-1}
\end{aligned}$$

We can estimate Σ_θ directly by substituting the sample moments for the respective expectations and $\widehat{\theta}_n^*$ in place of θ_0. Alternatively, we can use the variance estimator suggested by Rubin:

$$\begin{aligned}
\widehat{\Sigma}_\theta &= m^{-1}\sum_{j=1}^{m}\left[n^{-1}\sum_{i=1}^{n}\frac{-\partial S^F\left(Z_{ij}\left(\theta^{(j)}\right),\widehat{\theta}_{nj}^*\right)}{\partial\theta^\top}\right]^{-1} + \left(\frac{m+1}{m}\right) n\sum_{j=1}^{m}\frac{\left(\widehat{\theta}_{nj}^* - \widehat{\theta}_n^*\right)\left(\widehat{\theta}_{nj}^* - \widehat{\theta}_n^*\right)^\top}{m-1} \qquad (4.20)\\
&= m^{-1}\sum_{j=1}^{m} n^{-1}\widehat{I}_n\left(Z_{ij}\left(\theta^{(j)}\right),\widehat{\theta}_{nj}^*\right) + \left(\frac{m+1}{m}\right) n\sum_{j=1}^{m}\frac{\left(\widehat{\theta}_{nj}^* - \widehat{\theta}_n^*\right)\left(\widehat{\theta}_{nj}^* - \widehat{\theta}_n^*\right)^\top}{m-1}
\end{aligned}$$

In (4.20), $\widehat{I}_n\left(Z_{ij}\left(\theta^{(j)}\right),\widehat{\theta}_{nj}^*\right)$ in the first sum is the observed information from the jth imputed full-data set and this is ready obtained. The second term in (4.20) is readily computed from the multiple imputation estimates of θ from the m imputed full-data sets. Thus the Rubin's estimator is readily computed. In the Appendix, we show that $\widehat{\Sigma}_\theta$ is a consistent estimator of Σ^*.

4.4 Application

We use both real and simulated study data to illustrate the approach. We start with simulated data to investigate the properties of the multiple imputation estimator. In both real and simulated studies, we set type I error level at two-sided $\alpha = 0.05$.

4.4.1 Simulation Study

We consider a longitudinal study with two groups and three assessments $T = 3$. We consider a continuous response, y_{it}, and one continuous baseline confounder, x_i. We model the treatment effect, z_i, on the response using linear mixed-effects models.

Generating Complete Data

We assume that $x_i \sim N(0,1)$, $z_i \sim BN(p=\pi_i)$ and y_{it} given x_i and z_i follows a linear mixed-effects model:

$$
\begin{aligned}
& y_{it} = \beta_0 + \beta_1 z_i + \beta_2 x_i + b_i + \beta_3 I(t=2) + \beta_4 I(t=3) + \varepsilon_{it}, \\
& \varepsilon_{it} \sim N(0,\sigma^2), \quad x_i \sim N(0,\sigma_x^2), \quad z_i \sim BN(\pi_i), \quad b_i \sim N(0,\sigma_b^2) \\
& 1 \le t \le T \quad i = 1,2,\cdots,n
\end{aligned}
\tag{4.21}
$$

where $BN(\pi)$ denotes a Bernoulli with mean π. To model the confounder effect of x_i, we assume a logistic regression for z_i given x_i:

$$\text{logit}(\pi_i) = \xi_0 + \xi_1 x_i, \quad i = 1,2,\cdots,n$$

For the simulation study, we set the parameter values as follows:

$$
\begin{aligned}
& \beta_0 = 10, \quad \beta_1 = -2, \quad \beta_2 = 2, \quad \beta_3 = 2.5, \quad \beta_4 = 6 \\
& \sigma^2 = 2, \quad \sigma_x^2 = 1, \quad \sigma_b^2 = 1, \quad \xi_0 = 1, \quad \xi_1 = 1
\end{aligned}
$$

Generating Missing Data

We generate missing data for both the missing completely at random (MCAR) and missing at random (MAR) mechanism.

For creating missing y_{it} under MCAR, we randomly remove y_{i2} and y_{i3}. To do so, we generate two observed data indicators, r_{i2} and r_{i3}, from two Bernoulli distribution:

$$r_{i2} \sim BN(0.8), \quad r_{i3} \sim BN(0.8), \quad i = 1,2,\cdots,n$$

We set y_{i2} (y_{i3}) as missing when r_{i2} $(r_{i3}) = 0$.

For creating missing y_{it} under MAR, we model r_{i2} and r_{i3} using two logistic regression models:

$$r_{i2} \sim BN(p_{i3}), \quad r_{i3} \sim BN(p_{i3}), \quad i = 1,2,\cdots,n.$$
$$p_{i2} = \text{expit}\left(\gamma_0 + \gamma_1 y_{i1}\right), \quad p_{i3} = \text{expit}\left(\zeta_0 + \zeta_1 y_{i1} + \zeta_2 y_{i2}\right)$$

For our simulation, we set

$$\gamma_1 = \zeta_1 = \zeta_2 = 0.8$$

To ensure 80% missing for y_{i2} and 70% missing for y_{i3}, we set $\gamma_0 = -4.5$ and $\zeta_0 = -13.5$ to create 80% missing in y_{i2} and 70% missing in y_{i3}.

MAR Missing x_i. To do so, we generate one observed data indicator, v_i, from a Bernoulli distribution:

$$v_i \sim BN\left(p_i\right), \quad i = 1,2,\cdots,n$$

where

$$E\left(v_i \mid z_i\right) = p_i, \quad \text{logit}\left(p_i\right) = \eta_0 + \eta_1 z_i, \quad i = 1,2,\cdots,n$$

To ensure 80% observed data for v_i, we set $\eta_1 = 1$ and $\eta_0 = 0.9$.

Monte Carlo Simulation to Evaluate Performance of Estimator

We use Monte Carlo (MC) simulation to evaluate the performance of the multiple imputation estimator. We set the MC sample size $M = 500$. For multiple imputation, we set the number of imputations $m = 50$. We set $n = 300$ for the sample size.

We first create missing data and imputed full-data sets. For $k = 1,\cdots,M$, we simulate the kth observed data with missing y_{i2} and y_{i3}:

$$Z^{(k)} = \left\{y_{i1}, y_{i2}, y_{i3}, z_i; 1 \le i \le n\right\}.$$

For the kth observed data $Z^{(k)}$, we create m imputed datasets:

$$Z_m^{(k)} = \left\{Z^{(k,j)}; 1 \le i \le n, 1 \le j \le m\right\}, \quad Z^{(k,j)} = \left\{y_{i1}, y_{i2}^{(j)}, y_{i3}^{(j)}, z_i; 1 \le i \le n\right\}$$

where

$$y_{it}^{(j)} = \begin{cases} y_{it}(\theta_0) & \text{if } y_{it} \text{ is observed} \\ y_{it}\left(\theta^{(j)}\right) & \text{if } y_{it} \text{ is missing} \end{cases}, \quad 1 \le i \le n, 1 \le j \le m, 1 \le t \le T$$

For each imputed full-data set $Z^{(k,j)}$, we fit a linear mixed-effect model (LMM) and obtain an MLE $\widehat{\theta}_{nj}^{*(k)} = \left(\widehat{\beta}_{nj}^{*(k)}, \widehat{\eta}_{nj}^{*(k)}\right)$, where $\widehat{\beta}_{nj}^{*(k)}$ denotes the estimated parameters of interest β and $\widehat{\eta}_{nj}^{*(k)}$ denotes the estimated nuisance parameters η. We repeat this model fitting step for all the m imputed full-data sets $Z_m^{(k)}$, we have the multiple-imputation estimate for the k th observed dataset $Z^{(k,j)}$ as:

$$\widehat{\theta}_n^{*(k)} = m^{-1} \sum_{j=1}^{m} \widehat{\theta}_{nj}^{*(k)} \tag{4.22}$$

We also obtain a consistent estimate of the asymptotic variance $\widehat{\Sigma}^{*(k)}$ using (4.20):

$$\widehat{\Sigma}_\theta^{*(k)} = m^{-1} \sum_{j=1}^{m} n^{-1} \widehat{I}_n\left(Z_{ij}\left(\theta^{(j)}\right), \widehat{\theta}_{nj}^{*}\right) + \left(\frac{m+1}{m}\right) n \sum_{j=1}^{m} \frac{\left(\widehat{\theta}_{nj}^{*} - \widehat{\theta}_n^{*}\right)\left(\widehat{\theta}_{nj}^{*} - \widehat{\theta}_n^{*}\right)^\top}{m-1} \tag{4.23}$$

Since most packages such as R only provide variance estimates for the estimated parameters of interest β, we can apply (4.20) to β to obtain a consistent estimator of the corresponding $\widehat{\beta}_n^{*(k)}$. This allows us to compare the asymptotic with empirical variance estimates for the simulation study.

By repeating the MC iterations M times, we obtain the MC estimate of θ, along with the asymptotic and empirical variance estimates:

$$\begin{aligned} \overline{\widehat{\theta}}_n^{*} &= \frac{1}{M} \sum_{k=1}^{M} \widehat{\theta}_n^{*(k)}, \quad \widehat{\Sigma}_\theta^{asymp} = \frac{1}{M} \sum_{k=1}^{M} \widehat{\Sigma}_\theta^{*(k)} \\ \widehat{\Sigma}_\theta^{emp} &= \frac{1}{M-1} \sum_{k=1}^{M} \left(\widehat{\theta}_n^{*(k)} - \overline{\widehat{\theta}}_n^{*}\right)\left(\widehat{\theta}_n^{*(k)} - \overline{\widehat{\theta}}_n^{*}\right)^\top \end{aligned} \tag{4.24}$$

As asymptotic variance estimates are generally not available from most packages, we can apply (4.24) to β to evaluate the performance of the proposed estimator.

In addition to comparing estimates with true parameter values and asymptotic with empirical variance estimates, we also compare type I error rates. For example, for the parameter of interest β_1, we

compute the Wald statistic $W^{(k)}$ for testing the null $H_0 : \beta_1 = \beta_{10}$ at each MC iteration:

$$W_{\beta_1}^{(k)} = \left(\widehat{\sigma}_{\beta_1}^{2(k)}\right)^{-1} \left(\widehat{\beta}_1^{(k)} - \beta_{10}\right)^2, \quad 1 \leq k \leq M,$$

where $\widehat{\sigma}_{\beta_1}^{2(k)}$ denotes the element of the estimated asymptotic variance $\frac{1}{n}\widehat{\Sigma}_{\theta}^{*(k)}$ corresponding to $\widehat{\beta}_1^{(k)}$. The empirical type I error rate for testing H_0 by:

$$\widehat{\alpha} = \frac{1}{M} \sum_{k=1}^{M} I_{\left\{W_{\beta_1}^{(k)} \geq q_{1,0.95}\right\}}$$

where $q_{1,0.95}$ is the 95th percentile of the central χ_1^2 with one degree of freedom.

Results

We follow the steps generating the dataset and doing the Monte Carlo simulation as described in Section 4.4.1. When there is no missing values in covariates x, shown in Table 4.1 are the results for MCAR and in Table 4.2 are the results for MAR. When there are missing values in covariates x with assumption MAR, shown in Table 4.3 are the results for MCAR and in Table 4.4 are the results for MAR. As we can see from Table 4.1, Table 4.2, 4.3 and Table 4.4 the multiple imputation estimates $\widehat{\beta}_n^*$ were close to the true parameter value β_0, the empirical variance estimates $\widehat{\Sigma}_{\beta}^{emp}$ were also close to the MC asymptotic variance estimates $\widehat{\Sigma}_{\beta}^{asymp}$, and the type I error rates were close to the nominal$\alpha = 0.05$ for both MCAR and MAR mechanisms.

Table 4.1. Monte Carlo Simulation for MCAR assumption

	True Param.	Est. Param.	Mean Asymp. Var	Emp. Var	Emp. Type I error
Intercept	10	9.97	0.0408	0.0383	0.052
z	-2	-1.98	0.0490	0.0502	0.054
x	2	1.99	0.0104	0.0106	0.044
$I(t=2)$	2.5	2.50	0.0310	0.0294	0.046
$I(t=3)$	6	6.02	0.0311	0.0292	0.048

4.4.2 Real Study Data

We apply the proposed approach to a real study data. The Peer Empowerment Program 4 Physical Activity (PEP4PA) study [73] is a cluster randomized control trial with participants randomized to either

Table 4.2. Monte Carlo Simulation for MAR assumption

	True Param.	Est. Param.	Mean Asymp. Var	Emp. Var	Emp. Type I error
Intercept	10	9.99	0.0426	0.0399	0.042
z	-2	-1.99	0.0531	0.0500	0.044
x	2	1.99	0.0126	0.0119	0.041
$I(t=2)$	2.5	2.50	0.0338	0.0323	0.046
$I(t=3)$	6	6.04	0.0423	0.0416	0.048

Table 4.3. Monte Carlo Simulation for MCAR assumption on y_{it}, MAR on x_i

	True Param.	Est. Param.	Mean Asymp. Var	Emp. Var	Emp. Type I error
Intercept	10	10.01	0.0481	0.0483	0.056
z	-2	-2.01	0.0581	0.0596	0.054
x	2	1.99	0.0117	0.0112	0.042
$I(t=2)$	2.5	2.50	0.0312	0.0296	0.044
$I(t=3)$	6	6.00	0.0313	0.0292	0.048

a peer-led PA intervention or usual center programming [74]. The initial phase is a cluster-randomized control trial with subjects randomized to either a peer-led PA intervention or usual center programming. The study includes 476 participants, with 209 assigned the Control group and 267 assigned the Intervention group. For each group, the participants were recruited from 6 senior center sites. The patients are measured at baseline (A1), 6 month (A2), 12 month (A3), 18 month (A4) and 24 month (A5). We are interested in whether the Intervention group has significantly increase the minutes of f_mvpa during the period of visits after baseline visit. We consider applying the approach we proposed for PEP4PA dataset, the covariates are: age, gender, race, education, income, awake wear time, SPPB, 6MWT. We set $m = 50$ for the number of imputations when applying the method.

For the dataset, there are also some missing values in the covariates, we apply the proposed proper multiple imputation approach to the dataset with imputed covariates. We fit the linear mixed-effect model on the interaction of condition (Control or Intervention group) and different visits with all the covariates:

$$\begin{aligned} Y_{it} = {} & b_{i1} + b_{i2} + \beta_0 + \beta_1 \text{ Gender} + \beta_2 \text{ Age} + \beta_3 \text{ Race(Black)} + \beta_4 \text{ Race(Asian)} \\ & + \beta_5 \text{ Race(Other)} + \beta_6 \text{baseline\_edu\_bin} + \beta_7 \text{ baseline\_income\_bin} \\ & + \beta_8 \text{ awake wear time} + \beta_9 \text{ SPPB} + \beta_{10} \text{ MWT} + \beta_{11} \text{ Condition} \\ & + \beta_{12} I(tp = t) + \beta_{13} \text{ Condition} * I(tp = t) + \varepsilon_{it} \end{aligned}$$

Table 4.4. Monte Carlo Simulation for MAR assumption on y_{it}, MAR on x_i

	True Param.	Est. Param.	Mean Asymp. Var	Emp. Var	Emp. Type I error
Intercept	10	9.99	0.0504	0.0503	0.048
z	-2	-2.00	0.0621	0.0630	0.052
x	2	1.99	0.0139	0.0138	0.046
$I(t=2)$	2.5	2.51	0.0344	0.0373	0.048
$I(t=3)$	6	6.04	0.0430	0.0435	0.048

where longitudinal outcome Y_{it} is corresponding to f_mvpa at visit t for subject i, b_{i1} is the random intercept for each individual and b_{i2} is the random intercept for each individual within each site.

To make statistical inference about the multiple imputation estimators, we use the Rubin's variance formula in (4.20) as the variance. Results shown in below Table 4.5 and Table 4.6. After applying the proposed approach, the results showed that there was sufficient evidence (with p-value $< .001$) such that the Intervention group has significantly increased the minutes of moderate-vigorous-physical activity (f_mvpa) during the five visits after baseline visit, 6 month (A2), 12 month (A3), 18 month (A4) and 24 month (A5).

4.5 Discussion

In this paper, we proposed an approach that utilizes existing software such as the `R` package `mice` to implement multiple imputation. We focused on Type A estimators, as the popular Rubin's variance estimators are consistent only for this class of estimators. Although the type A estimator has been shown to be consistent and asymptotically normal, it is not easy to apply this estimator in real studies because of the difficulty to samples from the posited regression model for imputing missing responses. The problem lies in the difficulty of deriving the predictive model from the posited model. We addressed this issue by viewing the posited model as derived from a more complex parametric model for the joint distribution of the responses and explanatory variables used in imputing missing values in software such as MICE in R. Under this viewpoint, we can readily leverage existing software such as MICE to sample from the predictive distribution (from the parametric model for the joint distribution, rather than the conditional distribution of the response given the explanatory variables in the posited regression model) to impute missing data. Moreover, if the missingness also depends on observed data from other variables (not

Table 4.5. Type A multiple Imputation estimator on PEP4PA Dataset Results (With Observed Covariates)

	Est.	**Std. Error**	**z value**	***p*-value**
(Intercept)	-6.9505	14.5876	-0.4765	0.634
gender	-0.8954	2.0759	-0.4313	0.666
age	-0.22242	0.1072	-2.0921	0.036
Race				
Black	1.3886	3.0488	-0.4555	0.649
Asian	-4.7028	4.2800	-1.0988	0.272
Other	-1.7112	3.4148	-0.5011	0.616
education	2.5962	1.8709	1.3877	0.165
income	-2.3471	1.8484	-1.2698	0.204
awake wear time	0.0164	0.0109	1.4989	0.134
SPPB	0.3220	0.5226	0.6162	0.538
6MWT	0.0674	0.0135	4.9965	$< .001$
condition	-5.7007	2.4644	-2.3132	0.021
tpA2	-2.3890	1.2156	-1.9653	0.049
tpA3	-3.2251	1.2625	-2.5545	0.011
tpA4	-4.3520	1.2259	-3.5501	$< .001$
tpA5	-4.8442	1.3637	-3.5521	$< .001$
Intervention:tpA2	8.1383	1.8579	4.3804	$< .001$
Intervention:tpA3	10.1961	1.8520	5.5053	$< .001$
Intervention:tpA4	9.1850	1.7453	5.2628	$< .001$
Intervention:tpA5	8.4691	2.0745	4.0825	$< .001$

the observed responses and explanatory variables), we simply include these additional variables when computing the predictive distribution.

The proposed approach performed very well with missing under both MCAR and MAR, as demonstrated by the simulation study for varying sample sizes. In the future, we will consider additional missing data due to reasons other than lost-to-followup such as left- or right-censoring for both responses and explanatory variables. For example, missingness due to detection limit is a common issue in analyses of biomarker data. Multiple imputation can be readily used to address this type of missing data. Work is underway to extend the proposed approach to this and other types of missing data.

Chapter 4, in full, is currently being prepared for submission for publication of the material as it may appear in "Chenyu Liu, Loki Natarajan, Lin Liu, Yuqi Qiu, Xin Tu. *A Multiple Imputation Approach for Longitudinal Regression with Missing Response*". The dissertation author was the primary author on

Table 4.6. Type A multiple Imputation estimator on PEP4PA Dataset Results (With Imputed Missing Covariates)

	Est.	**Std. Error**	**z value**	***p*-value**
(Intercept)	-7.3077	13.3514	-0.5473	0.584
gender	-1.5267	1.8833	-0.8107	0.418
age	-0.2128	0.0966	-2.2037	0.028
Race				
Black	-0.7422	2.5901	-0.2865	0.774
Asian	-4.3087	4.1048	-1.0497	0.294
Other	-1.5657	3.0125	-0.5197	0.603
education	2.2384	1.6492	1.3573	0.175
income	-1.6982	1.6856	-1.0075	0.314
awake wear time	0.0159	0.0101	1.5841	0.113
SPPB	0.3052	0.4856	0.6285	0.530
6MWT	0.0674	0.01119	5.6697	$< .001$
condition	-5.5262	2.2626	-2.4425	0.015
tpA2	-2.6018	1.0978	-2.3701	0.018
tpA3	-3.4479	1.1619	-2.9675	0.003
tpA4	-4.4046	1.1925	-3.6938	$< .001$
tpA5	-4.9271	1.2606	-3.9086	$< .001$
Intervention:tpA2	7.9310	1.6145	4.9125	$< .001$
Intervention:tpA3	9.6879	1.6258	5.9589	$< .001$
Intervention:tpA4	8.2524	1.6912	4.8796	$< .001$
Intervention:tpA5	8.1120	1.7712	4.5798	$< .001$

this paper.

Chapter 5

Conclusion and Future Work

In the thesis, we developed innovative statistical methods aimed at tackling complex statistical issues arising from imperfect data problems across different fields. We focused on statistical methodologies for brain imaging data involving measurement errors in explanatory variables, robust hypothesis testing for average treatment effects in causal inference, and multiple imputation strategies for longitudinal regression analysis when dealing with missing responses (dependent variables). Simulation studies were used to investigate the performance of these proposed methods, while real studies were used to illustrate their applications.

In particular, in chapter two we investigated the performance of the best linear unbiased predictor (BLUP) for noisy brain data. We presented a Bayesian methodology that utilizes the Hamiltonian Monte Carlo (HMC) algorithm to correct for measurement errors in predictors and estimate the SNP heritability h2 as well as the fraction of variance explained (FVE). Our simulation studies demonstrated that the noisy BLUP provides FVE estimates very close to the true FVE, regardless of the noise level. In contrast, the "noise-free" BLUP yields biased FVE estimates that become progressively more attenuated as σ_ε^2 increases. By applying the noisy Best Linear Unbiased Predictor (BLUP) model to resting-state fMRI data from the ABCD Study, we observed a significant improvement in the FVE. The noisy BLUP model effectively addresses the inaccuracies in the predictor variables, enhancing the reliability of heritability estimates and the explained variance. In the future work, we plan to extend this model to incorporate more complex assumptions for the covariance matrix $\Sigma_{\mathbf{W}}$. For instance, we will explore using an autoregressive model of order 1 (AR(1)) or an exchangeable covariance structure to better capture the underlying dependencies in the data. These advanced covariance structures are expected to further refine the model's performance

and accuracy, providing more robust estimates of SNP heritability and the fraction of variance explained. By addressing these enhancements, we aim to expand the applicability and precision of the noisy BLUP model in various genetic and neuroimaging studies.

In chapter three, we introduced a new sign-flipping score-like test for inferring the average treatment effect (ATE) based on the doubly robust estimator (DRE) approach to improve anti-conservatism in small and moderate samples. While the DRE, along with its semi-parametric Outcome Regression Model (ORM) and Marginal Structural Model (MSM) components, exhibits desirable asymptotic properties, it struggles with anti-conservative type I error and bias in coverage probability in small and moderate samples, particularly for responses with heteroscedasticity and overdispersion . Although score-like tests generally outperform Wald tests, they still exhibit anticonservative bias in type I error and bias in coverage probability. The proposed sign-flipping score-like test addresses this bias more effectively in small and moderate samples. Additionally, this test applies to DRE with non-parametric models for covariates. In the ORM, the linear predictor includes a parametric specification for ATE and a non-parametric specification for the nuisance covariates, while in the MSM, the propensity score's nuisance covariates are modeled non-parametrically. The non-parametric models for the covariates typically use machine learning methods such as random forests for estimating model parameters. By reconstructing the sign-flipping score-like test using sample splitting, which has been effective in addressing sub-$\sqrt{n}$ convergence rates of non-parametric nuisance estimators, we achieved a test statistic not only with robust asymptotic properties, but also remarkably strong performance for small and moderate samples, as demonstrated by the simulation studies.

In chapter four, we proposed an approach that leverages existing software such as the `R` package `MICE` to implement multiple imputation. Our focus was on Type A estimators, as Rubin's variance estimators are consistent only for this class of estimators. Although Type A estimators are consistent and asymptotically normal, they are difficult to apply in real studies due to challenges in sampling from the posited regression model for imputing missing responses. This difficulty arises from the complexity of deriving the predictive model from the posited model. We addressed this challenge by interpreting the posited model as being derived from a more complex parametric model for the joint distribution of the responses and explanatory variables implemented in popular statistical software for imputing

missing data such as the MICE package in R. When missingness depends on observed data from other variables (not the ones for the posited regression model), we can include these variables in computing the predictive distribution, increasing inference validity under the missing at random (MAR) mechanism. Our method performed well under both MAR and the missing completely at random (MCAR) assumptions, as demonstrated in our simulation study with varying simulation settings. Future work will address additional types of missing data, such as left-censoring due to detection limit , which is a common issue in analyses of biomarker data and missing data in explanatory variables. By extending the proposed approach to handle these scenarios, we aim to enhance its applicability and robustness across a wider range of real-world data sets and missing data mechanisms, ultimately improving the reliability and validity of imputed data in various research contexts.

Appendix

A Additional materials for Chapter 3

A.1 Proof of Theorem 1

Proof. Suppose H_0 holds. We will show that $U^n = (U_1^n, \ldots, U_m^n)$ converges in distribution to a multivariate normal distribution with mean $\mathbf{0}$ and variance $\lim_{n\to\infty} s_n^2 I$, where I is the $m \times m$ identity matrix. It then follows from Lemma 11 in [40] that $\mathbb{P}\left(U_1^n > U_{[1-\alpha]}^n\right) \to \lfloor \alpha m \rfloor / m$.

Under H_0, for each $1 \leq j \leq m, \mathbb{E}\left(U_j^n\right) = 0$. For every $1 \leq j \leq m, \text{var}\left(U_j^n\right) = n^{-1} \sum_{i=1}^n \text{var}\left(U^{ni}\right) = s_n^2$. Let Q_n be the covariance matrix of $U^n \cdot Q_n$ has zeroes off the diagonal. Indeed, for $1 \leq j < k \leq m$

$$\text{cov}\left(U_j^n, U_k^n\right) = \text{cov}\left(n^{-1/2} \sum_{i=1}^n g_{ji} U^{ni}, n^{-1/2} \sum_{i=1}^n g_{ki} U^{ni}\right) = 0$$

since the $g_{ki}, 2 \leq k \leq m$, are independent with mean 0 . Hence Q_n converges to $\lim_{n\to\infty} s_n I$. Note that U^n is a sum of n vectors. By the multivariate LindebergFeller central limit theorem [75], U^n converges in distribution to a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\lim_{n\to\infty} s_n^2 I$. We have shown that U^n converges in distribution to a vector U, say, of i.i.d. normal random variables. It now follows from Lemma 11 that $\mathbb{P}\left(U_1^n > U_{[1-\alpha]}^n\right) \to \lfloor \alpha m \rfloor / m$. ■

A.2 Proof of Theorem 2

Proof. Suppose that H_0 holds. Note that

$$\widehat{U}_p^{n*} = \widehat{U}_p^n - \hat{\mathcal{I}}'_{pc}\hat{\mathcal{I}}_{cc}^{-1}\widehat{\mathbf{U}}_c^n = \widehat{U}_p^n - \mathcal{I}'_{pc}\mathcal{I}_{cc}^{-1}\widehat{\mathbf{U}}_c^n + o_{\mathbb{P}_{\beta_0,\gamma_0}}(1) =$$
$$U_p^n - \mathcal{I}'_{pc}\sqrt{n}\,(\hat{\gamma}-\gamma_0) - \mathcal{I}'_{pc}\mathcal{I}_{cc}^{-1}\left\{\mathbf{U}_c^n - \mathcal{I}_{cc}\sqrt{n}\,(\hat{\gamma}-\gamma_0)\right\} + o_{\mathbb{P}_{\beta_0,\gamma_0}}(1) =$$
$$U_p^{n*} + o_{\mathbb{P}_{\beta_0,\gamma_0}}(1).$$

Let $2 \leq j \leq w$ and

$$U_{pj}^{n+} = n^{-1/2}\sum_{i=1}^{n}\mathbf{1}_{\{g_{ji=1}\}}U_p^{ni} = n^{-1/2}\sum_{i=1}^{n}\mathbf{1}_{\{g_{ji=1}\}}D_{ip}V_i^{-1}S_i$$
$$U_{pj}^{n-} = n^{-1/2}\sum_{i=1}^{n}\mathbf{1}_{\{g_{ji=-1}\}}U_p^{ni} = n^{-1/2}\sum_{i=1}^{n}\mathbf{1}_{\{g_{ji=-1}\}}D_{ip}V_i^{-1}S_i.$$

Note that

$$\widehat{U}_{pj}^n = \widehat{U}_{pj}^{n+} - \widehat{U}_{pj}^{n-} = \left\{U_{pj}^{n+} - \frac{1}{2}\sqrt{n}\mathcal{I}'_{pc}\,(\hat{\gamma}-\gamma_0)\right\} - \left\{U_{pj}^{n-} - \frac{1}{2}\sqrt{n}\mathcal{I}'_{pc}\,(\hat{\gamma}-\gamma_0)\right\} + o_{\mathbb{P}_{\beta_0,\gamma_0}}(1) =$$
$$U_{pj}^{n+} - U_{pj}^{n-} + o_{\mathbb{P}_{\beta_0,\gamma_0}}(1) = U_{pj}^n + o_{\mathbb{P}_{\beta_0,\gamma_0}}(1).$$

The intuitive reason why $\widehat{U}_{pj}^n = U_{pj}^n + o_{\mathbb{P}_{\beta_0,\gamma_0}}(1)$, is that the estimation of $\hat{\gamma}$ does not cause the summands underlying $\widehat{U}_{pj}^n$ to be correlated. Similarly we find that $\widehat{\mathbf{U}}_{cj}^n = \mathbf{U}_{cj}^n + o_{\mathbb{P}_{\beta_0,\gamma_0}}(1)$ and conclude that $\widehat{U}_{pj}^{n*} = U_{pj}^{n*} + o_{\mathbb{P}_{\beta_0,\gamma_0}}(1)$.

Let U^n be as in the proof of Theorem 1, with U^{ni} replaced by $U_p^{ni*} = U_p^{ni} - \mathcal{I}_{cp}^{\top}\mathcal{I}_{cc}^{-1}\mathbf{U}_c^{ni}$. Suppose H_0 holds and $\hat{\mathcal{I}} = \mathcal{I}$, so that the summands underlying U_j^n are independent. For every $1 \leq i \leq n, \mathbb{E}\left(U_p^{ni*}\right) = 0$. The elements of U^n are uncorrelated and have common variance $\mathrm{var}\left(U_p^{n1*}\right)$. By the multivariate central limit theorem [75, 76], U^n converges in distribution to $N\left(\mathbf{0}, \mathrm{var}\left(U_p^{ni*}\right)I\right)$. We supposed that $\hat{\mathcal{I}} = \mathcal{I}$ to use the central limit theorem, but the asymptotic distribution of U^n is the same if $\hat{\mathcal{I}}$ is any consistent estimator of $\mathcal{I}$.

Let $\hat{U}^n$ be as in the proof of Theorem 1, with ν_i replaced by $\widehat{U}_p^{ni*} = \widehat{U}_p^{ni} - \widehat{\mathcal{I}}_{cp}^{\top}\widehat{\mathcal{I}}_{cc}^{-1}\widehat{\mathbf{U}}_c^{ni}$. For every $1 \leq j \leq w$, $\widehat{U}_{pj}^{n*} = U_{pj}^{n*} + o_{\mathbb{P}_{\beta_0,\gamma_0}}(1)$. Thus $\hat{U}^n$ and U^n are asymptotically equivalent. The result now follows

from Lemma 11 [40]. ■

B Additional materials for Chapter 4

B.1 Show (4.16)

For the term V_n, consider the stochastic process parameterized by β:

$$J_n(\beta) = n^{-1/2} \sum_{i=1}^{n} m^{-1} \sum_{j=1}^{m} S^F(Z_{ij}(\beta), \beta_0)$$

Then we can express V_n as:

$$V_n = J_n\left(\widehat{\beta}_n^I\right) - J_n(\beta_0) \tag{.1}$$

If $J_n(\beta)$ converges weakly to a tight Gaussian process, then $J_n(\beta)$ is stochastically equicontinuous in the sense that $J_n(\beta')$ and $J_n(\beta'')$become arbitrarily close to each other, if β' and β'' are sufficiently close to each other. We assume that $J(\beta)$ converges weakly to a tight Gaussian process.

Let

$$\lambda(\beta, \beta_0) = E\left[S^F(Z_{ij}(\beta), \beta_0)\right] \tag{.2}$$

Then the mean $E[J_n(\beta)]$ of $J_n(\beta)$ is given by:

$$W_n(\beta) = J_n(\beta) - E[J_n(\beta)] = J_n(\beta) - n^{-1/2} \sum_{i=1}^{n} E\left[S^F(Z_{ij}(\beta), \beta_0)\right] = J_n(\beta) - n^{1/2}\lambda(\beta, \beta_0)$$

Consider the centered version $W_n(\beta)$ of $J_n(\beta)$:

$$W_n(\beta) = J_n(\beta) - E[J_n(\beta)] = J_n(\beta) - n^{1/2}\lambda(\beta, \beta_0)$$

If the stochastic equicontinuity holds, then

$$W_n\left(\widehat{\beta}_n^I\right) - W_n(\beta_0) = \mathbf{o}_p(1) \tag{.3}$$

whenever $\widehat{\beta}_n^I$ is a consistent estimator of β_0.

To prove (.3), choose arbitrary $\varepsilon, \eta > 0$. Then, by stochastic equicontinuity, there exists a $\delta(\varepsilon, \eta)$ and $n(\varepsilon, \eta)$ such that

$$P\left\{\sup_{\beta', \beta'' : \|\beta' - \beta''\| < \delta(\varepsilon,\eta)} \left\|W_n(\beta') - W_n(\beta'')\right\| \geqslant \varepsilon\right\} \leq \eta/2$$

for all $n > n(\varepsilon, \eta)$. By consistency, there exists for every δ and η an $n^*(\delta, \eta)$ such that

$$P\left(\left\|\widehat{\beta}_n^I - \beta_0\right\| \geqslant \delta\right) \leq \eta/2 \text{ for all } n > n^*(\delta, \eta).$$

Note that the event

$$\begin{aligned}
&\left\{\left\|\widehat{\beta}_n^I - \beta_0\right\| < \delta(\varepsilon, \eta) \text{ and } \sup_{\beta', \beta'' : \|\beta' - \beta''\| < \delta(\varepsilon,\eta)} \left\|W_n(\beta') - W_n(\beta'')\right\| < \varepsilon\right\} \\
&\subseteq \left\{\left\|W_n\left(\widehat{\beta}_n^I\right) - W_n(\beta_0)\right\| < \varepsilon\right\}
\end{aligned}$$

By taking complements, we obtain

$$\begin{aligned}
&P\left\{\left\|W_n\left(\widehat{\beta}_n^I\right) - W_n(\beta_0)\right\| \geqslant \varepsilon\right\} \\
&\leq P\left[\left\{\left\|\widehat{\beta}_n^I - \beta_0\right\| \geqslant \delta(\varepsilon, \eta)\right\}\right. \\
&\quad \left. \cup \left\{\sup_{\beta', \beta'' : \|\beta' - \beta''\| < \delta(\varepsilon,\eta)} \left\|W_n(\beta') - W_n(\beta'')\right\| \geqslant \varepsilon\right\}\right] \\
&\leq P\left\{\left\|\widehat{\beta}_n^I - \beta_0\right\| \geqslant \delta(\varepsilon, \eta)\right\} \\
&\quad + P\left\{\sup_{\beta', \beta'' : \|\beta' - \beta''\| < \delta(\varepsilon,\eta)} \left\|W_n(\beta') - W_n(\beta'')\right\| \geqslant \varepsilon\right\}.
\end{aligned}$$

By choosing $n > \max\left[n(\varepsilon, \eta), n^*\{\delta(\varepsilon, \eta), \eta\}\right]$, we obtain

$$P\left\{\left\|W_n\left(\widehat{\beta}_n^I\right) - W_n(\beta_0)\right\| \geqslant \varepsilon\right\} \leq \eta/2 + \eta/2 = \eta.$$

By (.3), we have

$$\begin{aligned} V_n &= J_n\left(\widehat{\beta}_n^I\right) - J_n\left(\beta_0\right) = W_n\left(\widehat{\beta}_n^I\right) + n^{1/2}\lambda\left(\widehat{\beta}_n^I, \beta_0\right) - W_n\left(\beta_0\right) + n^{1/2}\lambda\left(\beta_0, \beta_0\right) \\ &= W_n\left(\widehat{\beta}_n^I\right) - W_n\left(\beta_0\right) + n^{1/2}\left[\lambda\left(\widehat{\beta}_n^I, \beta_0\right) - \lambda\left(\beta_0, \beta_0\right)\right] \\ &= n^{1/2}\left[\lambda\left(\widehat{\beta}_n^I, \beta_0\right) - \lambda\left(\beta_0, \beta_0\right)\right] + \mathbf{o}_p\left(1\right) \end{aligned} \tag{.4}$$

By a Taylor series expansion, we have:

$$\begin{aligned} V_n &= n^{1/2}\left[\lambda\left(\widehat{\beta}_n^I, \beta_0\right) - \lambda\left(\beta_0, \beta_0\right)\right] + \mathbf{o}_p\left(1\right) \\ &= \left.\frac{\partial\lambda\left(\beta, \beta_0\right)}{\partial\beta^\top}\right|_{\beta=\beta_0} n^{1/2}\left(\widehat{\beta}_n^I - \beta_0\right) + \mathbf{o}_p\left(1\right) \end{aligned} \tag{.5}$$

Now we show that the derivative $\left.\frac{\partial\lambda(\beta,\beta_0)}{\partial\beta^\top}\right|_{\beta=\beta_0}$ is given by:

$$\left.\frac{\partial\lambda\left(\beta, \beta_0\right)}{\partial\beta^\top}\right|_{\beta=\beta_0} = I^F\left(\beta_0\right) - I\left(\beta_0\right) \tag{.6}$$

By the law of conditional expectations, (.2) equals

$$E\left[E\left(S^F\left(Z_{ij}\left(\beta\right), \beta_0\right) \mid \mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right), \beta\right)\right] \tag{.7}$$

Because $Z_{ij}\left(\beta\right)$ is a random draw from the conditional distribution with conditional density

$$p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}\left(z \mid \mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right), \beta\right)$$

this means that the inner expectation of (.2) is equal to

$$\int S^F\left(z, \beta_0\right) p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}\left(z \mid \mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right), \beta\right) d\nu_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}\left(z \mid \mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right).$$

The outer expectation of (.7), however, involves the distribution of the observed data $\{\mathcal{C}_i, G_{\mathcal{C}}\left(Z_i\right)\}$, which

are generated from the truth, β_0. Therefore, the overall expectation is given as:

$$\lambda(\beta,\beta_0) = \int \left[\int S^F(z,\beta_0)\, p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}(z \mid r,g_r,\beta)\, d\nu_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}(z \mid r,g_r) \right] \times p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta_0)\, d\nu_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r)$$

Because

$$p_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}(z \mid r,g_r,\beta)\, p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta_0)\, d\nu_{Z|\mathcal{C},G_{\mathcal{C}}(Z)}(z \mid r,g_r)\, d\nu_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r) = p_{\mathcal{C},Z}(r,z,\beta)\, d\nu_{\mathcal{C},Z}(r,z)$$

we can write the above as

$$\int S^F(z,\beta_0) \left(\frac{p_{\mathcal{C},Z}(r,z,\beta)}{p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta)} \right) p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta_0)\, d\nu_{\mathcal{C},Z}(r,z). \tag{.8}$$

Taking the derivative of (.8) with respect to β and evaluating at β_0, we obtain

$$\begin{aligned} \left.\frac{\partial \lambda(\beta,\beta_0)}{\partial \beta^\top}\right|_{\beta=\beta_0} &= \int S^F(z,\beta_0) \left.\frac{\partial}{\partial \beta^\top}\left(\frac{p_{\mathcal{C},Z}(r,z,\beta)}{p_{\mathcal{C},G_{\mathcal{C}}}(r)} \right)\right|_{\beta=\beta_0} \\ &\quad \times p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta_0)\, d\nu_{\mathcal{C},Z}(r,z) \\ &= \int S^F(z,\beta_0) \left(\frac{p_{\mathcal{C},Z}(r,z,\beta_0)}{p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta_0)} \right) \left.\frac{\partial}{\partial \beta^\top} \log\left(\frac{p_{\mathcal{C},Z}(r,z,\beta)}{p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta)} \right)\right|_{\beta=\beta_0} \\ &\quad \times p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta_0)\, d\nu_{\mathcal{C},Z}(r,z) \end{aligned} \tag{.9}$$

Because the data are coarsened at random,

$$\frac{p_{\mathcal{C},Z}(r,z,\beta)}{p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta)} = \frac{p_Z(z,\beta)}{p_{G_r(Z)}(g_r,\beta)}.$$

Therefore

$$\left.\frac{\partial}{\partial \beta^\top} \log\left(\frac{p_{\mathcal{C},Z}(r,z,\beta)}{p_{\mathcal{C},G_{\mathcal{C}}(Z)}(r,g_r,\beta)} \right)\right|_{\beta=\beta_0} = \frac{\partial}{\partial \beta^\top} \log p_Z(z,\beta_0) - \frac{\partial}{\partial \beta^\top} \log p_{G_r(Z)}(g_r,\beta_0). \tag{.10}$$

When data are coarsened at random, we showed in Lemma 7.2 [43] that

$$\frac{\partial}{\partial \beta^{\top}} \log p_{G_r(Z)}\left(g_r, \beta_0\right) = E\left(S^F(Z) \mid \mathcal{C} = r, G_{\mathcal{C}}(Z) = g_r\right) = S\left(r, g_r\right)$$

Therefore (.10) is equal to

$$\left(S^F\left(z, \beta_0\right)\right)^{\top} - S^{\top}\left(z, \beta_0\right)$$

Substituting this last result into (.9) and rearranging some terms yields

$$\begin{aligned}
&\int S^F\left(z, \beta_0\right)\left[\left(S^F\left(z, \beta_0\right)\right)^{\top} - S^{\top}\left(r, g_r, \beta_0\right)\right] p_{\mathcal{C},Z}\left(r, z, \beta_0\right) d\nu_{\mathcal{C},Z}(r, z) \\
&= E\left[S^F\left(Z, \beta_0\right)\left(S^F\left(Z, \beta_0\right)\right)^{\top}\right] - E\left[S^F\left(Z, \beta_0\right) S^{\top}\left(\mathcal{C}, G_{\mathcal{C}}(Z), \beta_0\right)\right] \\
&= I^F\left(\beta_0\right) - E\left\{E\left[S^F\left(Z, \beta_0\right) S^{\top}\left(\mathcal{C}, G_{\mathcal{C}}(Z), \beta_0\right)\right] \mid \mathcal{C}, G_{\mathcal{C}}(Z), \beta_0\right\} \\
&= I^F\left(\beta_0\right) - E\left[S^{\top}\left(\mathcal{C}, G_{\mathcal{C}}(Z), \beta_0\right) S^{\top}\left(\mathcal{C}, G_{\mathcal{C}}(Z), \beta_0\right)\right] \\
&= I^F\left(\beta_0\right) - I\left(\beta_0\right)
\end{aligned}$$

By combining the above results, we can expresses V_n in terms of the IF of $n^{1/2}\left(\widehat{\beta}_n^I - \beta_0\right)$:

$$V_n = n^{-1/2} \sum_{i=1}^{n} \left(I^F\left(\beta_0\right) - I\left(\beta_0\right)\right) q\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right) + \mathbf{o}_p(1) \tag{.11}$$

By (4.8) and (.11), we have:

$$\begin{aligned}
n^{1/2}\left(\widehat{\beta}_n^* - \beta_0\right) &= \left(I^F\left(\beta_0\right)\right)^{-1}\left(U_n + V_n\right) + \mathbf{o}_p(1) \\
&= n^{-1/2} \sum_{i=1}^{n} \left(I^F\right)^{-1}\left[m^{-1} \sum_{j=1}^{m} S^F\left(Z_{ij}\left(\beta_0\right), \beta_0\right) + \left(I^F - I\right) q\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right] + \mathbf{o}_p(1)
\end{aligned}$$

B.2 Show $\operatorname{var}\left(\psi_i\left(Z_i^m\left(\theta_0\right); \theta_0\right)\right)$ is given by (4.19)

Let

$$A_m = m^{-1} \sum_{j=1}^{m} S^F\left(Z_{ij}\left(\beta_0\right), \beta_0\right), \quad B = \left(I^F - I\right) q\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)$$

Then

$$\operatorname{var}\left(\psi_i\left(Z_i^m\left(\theta_0\right);\theta_0\right)\right)=\operatorname{var}\left(A_m\right)+E\left(A_mB^{\top}\right)+E\left(BA_m^{\top}\right)+\operatorname{var}\left(B\right)$$

By the law of conditional variance, we have:

$$\operatorname{var}\left(A_m\right)=E\left[\left(\operatorname{var}\left(A_m\mid\mathcal{C}_i,G_{\mathcal{C}_i}\left(Z_i\right)\right)\right)\right]+\operatorname{var}\left(E\left(A_m\mid\mathcal{C}_i,G_{\mathcal{C}_i}\left(Z_i\right)\right)\right)=J_{m1}+J_{m2}$$

By (4.1), we have:

$$\begin{aligned}J_{m1}&=E\left[\operatorname{var}\left(m^{-1}\sum_{j=1}^{m}S^F\left(Z_{ij}\left(\beta_0\right),\beta_0\right)\mid\mathcal{C}_i,G_{\mathcal{C}_i}\left(Z_i\right)\right)\right]\\&=E\left[m^{-1}\operatorname{var}\left(S^F\left(Z_i,\beta_0\right)\mid\mathcal{C}_i,G_{\mathcal{C}_i}\left(Z_i\right)\right)\right]\\&=m^{-1}E\left[\operatorname{var}\left(S^F\left(Z_i,\beta_0\right)\mid\mathcal{C}_i,G_{\mathcal{C}_i}\left(Z_i\right)\right)\right]\\&=m^{-1}\left(I^F\left(\beta_0\right)-I\left(\beta_0\right)\right)\end{aligned}$$

Similarly, we have:

$$\begin{aligned}J_{m2}&=\operatorname{var}\left[E\left(m^{-1}\sum_{j=1}^{m}S^F\left(Z_{ij}\left(\beta_0\right),\beta_0\right)\mid\mathcal{C}_i,G_{\mathcal{C}_i}\left(Z_i\right)\right)\right]=\operatorname{var}(E\left(S^F\left(Z_i,\beta_0\right)\mid\mathcal{C}_i,G_{\mathcal{C}_i}\left(Z_i\right)\right)\\&=\operatorname{var}(S\left(\mathcal{C}_i,G_{\mathcal{C}_i}\left(Z_i\right)\right)\\&=I\left(\beta_0\right)\end{aligned}$$

Thus

$$\operatorname{var}\left(A_m\right)=m^{-1}\left(I^F-I\right)+I.$$

Next we compute the covariance matrix

$$\begin{aligned}
E\left(A_m B^\top\right) &= E\left[\left(m^{-1}\sum_{j=1}^{m} S^F\left(Z_{ij}\left(\beta_0\right),\beta_0\right)\right)\left[\left(I^F - I\right)q\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right]^\top\right] \\
&= m^{-1}\sum_{j=1}^{m} E\left[S^F\left(Z_{ij}\left(\beta_0\right),\beta_0\right)q^\top\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right]\left(I^F - I\right) \\
&= E\left[S^F\left(Z_{ij}\left(\beta_0\right),\beta_0\right)q^\top\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right]\left(I^F - I\right) \\
&= A\left(I^F - I\right)
\end{aligned}$$

Since

$$\begin{aligned}
A &= E\left[S^F\left(Z_{ij}\left(\beta_0\right),\beta_0\right)q^\top\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right] \\
&= E\left\{E\left[S^F\left(Z_{ij}\left(\beta_0\right),\beta_0\right)q^\top\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right) \mid \mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right]\right\} \\
&= E\left[E\left(S^F\left(Z_{ij}\left(\beta_0\right),\beta_0\right) \mid \mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right]q^\top\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right) \\
&= E\left[S\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)q^\top\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right] \\
&= \mathbf{I}_q
\end{aligned}$$

we have:

$$E\left(A_m B^\top\right) = I^F\left(\beta_0\right) - I\left(\beta_0\right)$$

Since $E\left(A_m B^\top\right)$ is symmetric,

$$E\left(A_m^\top B\right) = E\left(A_m B^\top\right) = I^F\left(\beta_0\right) - I\left(\beta_0\right)$$

Finally,

$$\begin{aligned}
E\left(BB^\top\right) &= \operatorname{var}\left[\left(I^F - I_0\right)q\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right] \\
&= \left(I^F - I\right)\operatorname{var}\left(q\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right)\left(I^F - I\right)
\end{aligned}$$

Thus

If we use $q\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right) = \varphi_{\text{eff}}\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right) + h\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)$ with $\text{var}\left(\varphi_{\text{eff}}\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right) = I^{-1}$ and $\varphi_{\text{eff}}\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right) \perp h\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)$, the above can be expressed as:

$$\begin{aligned}
E\left(BB^{\top}\right) &= \left(I^F - I\right)\text{var}\left(q\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right)\left(I^F - I\right) \\
&= \left(I^F - I\right)\left[I^{-1} + \text{var}\left(h\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right)\right]\left(I^F - I\right) \\
&= \left(I^F - I\right)I^{-1}\left(I^F - I\right) + \left(I^F - I\right)\text{var}\left(h\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right)\left(I^F - I\right) \\
&= \left(I^F I^{-1} - II^{-1}\right)\left(I^F - I\right) + \left(I^F - I\right)\text{var}\left(h\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right)\left(I^F - I\right) \\
&= I^F I^{-1} I^F - II^{-1} I^F - I^F I^{-1} I + II^{-1} I + \left(I^F - I\right)\text{var}\left(h\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right)\left(I^F - I\right) \\
&= I^F I^{-1} I^F - 2I^F + I + \left(I^F - I\right)\text{var}\left(h\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right)\left(I^F - I\right)
\end{aligned}$$

Thus we have:

$$\begin{aligned}
\Sigma^* &= \left(I^F\right)^{-1}\left[E\left(A_m A_m^{\top}\right) + E\left(BB^{\top}\right) + E\left(A_m B^{\top}\right) + E\left(BA_m^{\top}\right)\right]\left(I^F\right)^{-1} \\
&= \left(I^F\right)^{-1}\left[I + m^{-1}\left(I^F - I\right) + \left(I^F - I\right)\text{var}\left(q\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right)\left(I^F - I\right) + 2\left(I^F - I\right)\right]\left(I^F\right)^{-1}
\end{aligned}$$

If use $q\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right) = \varphi_{\text{eff}}\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right) + h\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)$, we obtain:

$$\begin{aligned}
J_4 &= E\left(A_m A_m^{\top}\right) + E\left(BB^{\top}\right) + E\left(A_m B^{\top}\right) + E\left(BA_m^{\top}\right) \\
&= I + m^{-1}\left(I^F - I\right) + I^F I^{-1} I^F - 2I^F + I + \left(I^F - I\right)\text{var}\left(h\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right)\left(I^F - I\right) + \\
&+ 2\left(I^F - I\right) \\
&= I^F I^{-1} I^F + m^{-1}\left(I^F - I\right) + \left(I^F - I\right)\text{var}\left(h\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right)\left(I^F - I\right)
\end{aligned}$$

Thus we have:

$$\begin{aligned}
\Sigma^* &= \left(I^F\right)^{-1} J_4 \left(I^F\right)^{-1} \\
&= I^{-1} + m^{-1}\left(I^F\right)^{-1}\left(I^F - I\right)\left(I^F\right)^{-1} + \left(I^F\right)^{-1}\left(I^F - I\right)\text{var}\left(h\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right)\left(I^F - I\right)\left(I^F\right)^{-1}
\end{aligned}$$

B.3 show $\widehat{\Sigma}_\theta$ is a consistent estimator of Σ^*

First, we have

$$m^{-1}\sum_{j=1}^{m}\left[n^{-1}\sum_{i=1}^{n}\frac{-\partial S^F\left(Z_{ij}\left(\theta^{(j)}\right),\widehat{\theta}^*_{nj}\right)}{\partial\theta^\top}\right]^{-1}\to_p\left(\frac{-\partial S^F\left(Z_i\left(\theta_0\right),\theta_0\right)}{\partial\theta^\top}\right)^{-1}=\left(I^F\right)^{-1}\tag{.12}$$

By (4.18) and (4.20), we have:

$$\begin{aligned}
n^{1/2}\left(\widehat{\beta}^*_{nj}-\widehat{\beta}^*_n\right)&=n^{1/2}\left(\widehat{\beta}^*_{nj}-\beta_0\right)-n^{1/2}\left(\widehat{\beta}^*_n-\beta_0\right)\\
&=\left(I^F\right)^{-1}n^{-1/2}\sum_{i=1}^{n}\left[S^F\left(Z_{ij}\left(\beta_0\right),\beta_0\right)-\overline{S}^F_i\left(\beta_0\right)\right]+\\
&+\left(I^F\right)^{-1}\left(I^F-I\right)\left(V_j-\overline{V}\right)+\mathbf{o}_p(1)\\
&=\left(I^F\right)^{-1}n^{-1/2}\sum_{i=1}^{n}G_{ij}+\left(I^F\right)^{-1}\left(I^F-I\right)\left(V_j-\overline{V}\right)+\mathbf{o}_p(1)
\end{aligned}$$

where

$$\begin{aligned}
G_{ij}&=S^F\left(Z_{ij}\left(\beta_0\right),\beta_0\right)-\overline{S}^F_i\left(\beta_0\right)\\
\overline{S}^F_i\left(\beta_0\right)&=m^{-1}\sum_{j=1}^{m}S^F\left(Z_{ij}\left(\beta_0\right),\beta_0\right),\quad\overline{V}=m^{-1}\sum_{j=1}^{m}V_j
\end{aligned}$$

Since V_j and $S^F\left(Z_{ij}\left(\beta_0\right),\beta_0\right)$ are independent, $G_{ij}=S^F\left(Z_{ij}\left(\beta_0\right),\beta_0\right)-\overline{S}^F_i\left(\beta_0\right)$ and $V_j-\overline{V}$ are independent. Thus we have:

$$\begin{aligned}
E\left[n\left(\widehat{\beta}^*_{nj}-\widehat{\beta}^*_n\right)\left(\widehat{\beta}^*_{nj}-\widehat{\beta}^*_n\right)^\top\right]&\to\left(I^F\right)^{-1}\left[n^{-1}nE\left(G_{ij}G_{ij}^\top\right)\right]\left(I^F\right)^{-1}+\\
&+\left(I^F\right)^{-1}\left(I^F-I\right)\mathrm{var}\left(V_j-\overline{V}\right)\left(I^F-I\right)\left(I^F\right)^{-1}\\
&=\left(I^F\right)^{-1}E\left(G_{ij}G_{ij}^\top\right)\left(I^F\right)^{-1}+\\
&+\left(I^F\right)^{-1}\left(I^F-I\right)\mathrm{var}\left(V_j-\overline{V}\right)\left(I^F-I\right)\left(I^F\right)^{-1}\\
&=\left(I^F\right)^{-1}\left(A+B\right)\left(I^F\right)^{-1}
\end{aligned}$$

and

$$
\begin{aligned}
E\left[n\sum_{j=1}^{m}\left(\widehat{\beta}_{nj}^{*}-\widehat{\beta}_{n}^{*}\right)\left(\widehat{\beta}_{nj}^{*}-\widehat{\beta}_{n}^{*}\right)^{\top}\right] &\to \left(I^{F}\right)^{-1}E\left[\sum_{j=1}^{m}\left(\sum_{i=1}^{n}\sum_{i'=1}^{n}G_{ij}G_{i'j}^{\top}\right)\right]\left(I^{F}\right)^{-1}+ \\
&+\left(I^{F}\right)^{-1}\left(I^{F}-I\right)\left(\sum_{j=1}^{m}\operatorname{var}\left(V_{j}-\overline{V}\right)\right)\left(I^{F}-I\right)\left(I^{F}\right)^{-1} \\
&=\left(I^{F}\right)^{-1}\left[\sum_{j=1}^{m}E\left(G_{ij}G_{ij}^{\top}\right)\right]\left(I^{F}\right)^{-1}+ \\
&+\left(I^{F}\right)^{-1}\left(I^{F}-I\right)\left(\sum_{j=1}^{m}\operatorname{var}\left(V_{j}-\overline{V}\right)\right)\left(I^{F}-I\right)\left(I^{F}\right)^{-1} \\
&=\left(I^{F}\right)^{-1}\left(mA+mB\right)\left(I^{F}\right)^{-1}
\end{aligned}
\tag{.13}
$$

where

$$
\begin{aligned}
A &= \left(I^{F}\right)^{-1}E\left(G_{ij}G_{ij}^{\top}\right)\left(I^{F}\right)^{-1} \\
B &= \left(I^{F}\right)^{-1}\left(I^{F}-I\right)\operatorname{var}\left(V_{j}-\overline{V}\right)\left(I^{F}-I\right)\left(I^{F}\right)^{-1}
\end{aligned}
\tag{.14}
$$

To find $E\left(G_{ij}G_{ij}^{\top}\right)$, we have:

$$
\begin{aligned}
E\left(G_{ij}G_{ij}^{\top}\right) &= E\left[\left(S^{F}\left(Z_{ij}\right)-\overline{S}_{i}^{F}\right)\left(S^{F}\left(Z_{ij}\right)-\overline{S}_{i}^{F}\right)^{\top}\right] \\
&= E\left[S^{F}\left(Z_{ij}\right)\left(S^{F}\left(Z_{ij}\right)\right)^{\top}\right]-2E\left[S^{F}\left(Z_{ij}\right)\left(\overline{S}_{i}^{F}\right)^{\top}\right]+E\left[\overline{S}_{i}^{F}\left(\overline{S}_{i}^{F}\right)^{\top}\right] \\
&= E\left[S^{F}\left(Z_{ij}\right)\left(S^{F}\left(Z_{ij}\right)\right)^{\top}\right]-2E\left[S^{F}\left(Z_{ij}\right)\left(\overline{S}_{i}^{F}\right)^{\top}\right]+E\left[\overline{S}_{i}^{F}\left(\overline{S}_{i}^{F}\right)^{\top}\right] \\
&= E\left[S^{F}\left(Z_{ij}\right)\left(S^{F}\left(Z_{ij}\right)\right)^{\top}\right]-E\left[S^{F}\left(Z_{ij}\right)\left(\overline{S}_{i}^{F}\right)^{\top}\right] \\
&= E\left[S^{F}\left(Z_{ij}\right)\left(S^{F}\left(Z_{ij}\right)\right)^{\top}\right]-\frac{1}{m}\sum_{j'=1}^{m}E\left[S^{F}\left(Z_{ij}\right)\left(S_{ij'}^{F}\right)^{\top}\right] \\
&= E\left[S^{F}\left(Z_{ij}\right)\left(S^{F}\left(Z_{ij}\right)\right)^{\top}\right]-\frac{1}{m}J
\end{aligned}
$$

where

$$J = \sum_{j'=1}^{m} E\left[S^F\left(Z_{ij}\right)\left(S^F_{ij'}\right)^\top\right]$$

When $j = j'$,

$$E\left[S^F\left(Z_{ij}\right)\left(S^F\left(Z_{ij}\right)\right)^\top\right] = \operatorname{var}\left(S^F\left(Z_{ij}\right)\right) = I^F$$

When $j \neq j'$, since $Z_{ij}(\beta_0)$ are independent draws for different j from the conditional density $p_{Z|C,GC(Z)}(z \mid C_i, G_{C_i}(Z_i))$, we have:

$$\begin{aligned}
E\left[S^F\left(Z_{ij}\right)\left(S^F\left(Z_{ij'}\right)\right)^\top\right] &= \operatorname{cov}\left[S^F\left(Z_{ij}\right), S^F\left(Z_{ij'}\right)\right] \\
&= E\left\{\operatorname{cov}\left[S^F\left(Z_{ij}\right), S^F\left(Z_{ij'}\right) \mid \mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right]\right\} + \\
&+ \operatorname{cov}\left(E\left[S^F\left(Z_{ij}\right) \mid \mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right], E\left[S^F\left(Z_{ij'}\right) \mid \mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right]\right) \\
&= 0 + \operatorname{var}\left(S\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right) \\
&= \operatorname{var}\left(S\left(\mathcal{C}_i, G_{\mathcal{C}_i}\left(Z_i\right)\right)\right) \\
&= I
\end{aligned}$$

Thus

$$\begin{aligned}
J &= E\left[S^F\left(Z_{ij}\right)\left(S^F_{ij}\right)^\top\right] + \sum_{j=j'}^{m} E\left[S^F\left(Z_{ij}\right)\left(S^F_{ij}\right)^\top\right] \\
&= I^F - (m-1)I
\end{aligned} \tag{.15}$$

and

$$\begin{aligned}
E\left(G_{ij}G_{ij}^\top\right) &= E\left[S^F\left(Z_{ij}\right)\left(S^F\left(Z_{ij}\right)\right)^\top\right] - \frac{1}{m}J \\
&= I^F - \frac{1}{m}\left[I^F - (m-1)I\right] \\
&= \frac{1}{m}\left\{mI^F - \left[I^F - (m-1)I\right]\right\} \\
&= \frac{1}{m}(m-1)\left(I^F - I\right)
\end{aligned}$$

By (.14) and (.15), we have:

$$mA = mE\left(G_{ij}G_{ij}^{\top}\right) = (m-1)\left(I^F - I\right) \tag{.16}$$

By a similar argument, we have:

$$m\,\mathrm{var}\left(V_j - \overline{V}\right) = (m-1)\,\mathrm{var}\left(q\left(\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right)\right) \tag{.17}$$

By (.13), (.16) and (.17), we have:

$$m(A+B) = (m-1)\left[\left(I^F - I\right) + \left(I^F - I\right)\mathrm{var}\, q\left(\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right)\left(I^F - I\right)\right]$$

and

$$E\left[n\sum_{j=1}^{m}\left(\widehat{\beta}_{nj}^{*} - \widehat{\beta}_{n}^{*}\right)\left(\widehat{\beta}_{nj}^{*} - \widehat{\beta}_{n}^{*}\right)^{\top}\right] \to \left(I^F\right)^{-1} m(A+B)\left(I^F\right)^{-1} \tag{.18}$$

By (.12) and (.18), we have:

$$\begin{aligned}
\widehat{\Sigma} &= m^{-1}\sum_{j=1}^{m}\left[n^{-1}\sum_{i=1}^{n}\frac{-\partial S^F\left(Z_{ij}\left(\widehat{\beta}_n^I\right), \widehat{\beta}_{nj}^{*}\right)}{\partial \beta^{\top}}\right]^{-1} + \left(\frac{m+1}{m}\right) n\sum_{j=1}^{m}\frac{\left(\widehat{\beta}_{nj}^{*} - \widehat{\beta}_{n}^{*}\right)\left(\widehat{\beta}_{nj}^{*} - \widehat{\beta}_{n}^{*}\right)^{\top}}{m-1} \\
&\to E\left[m^{-1}\sum_{j=1}^{m}\left(n^{-1}\sum_{i=1}^{n}\frac{-\partial S^F\left(Z_{ij}\left(\widehat{\beta}_n^I\right), \widehat{\beta}_{nj}^{*}\right)}{\partial \beta^{\top}}\right)^{-1}\right] + \\
&+ \left(\frac{m+1}{m(m-1)}\right) E\left[n\sum_{j=1}^{m}\left(\widehat{\beta}_{nj}^{*} - \widehat{\beta}_{n}^{*}\right)\left(\widehat{\beta}_{nj}^{*} - \widehat{\beta}_{n}^{*}\right)^{\top}\right] \\
&= \left(I^F\right)^{-1} + \frac{m+1}{m(m-1)}(m-1)\left(I^F\right)^{-1}\left[\left(I^F - I\right) + \left(I^F - I\right)\mathrm{var}\, q\left(\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right)\left(I^F - I\right)\right]\left(I^F\right)^{-1} \\
&= \left(I^F\right)^{-1} + \frac{m+1}{m}\left(I^F\right)^{-1}\left[\left(I^F - I\right) + \left(I^F - I\right)\mathrm{var}\, q\left(\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right)\left(I^F - I\right)\right]\left(I^F\right)^{-1}
\end{aligned}$$

Bibliography

[1] Hernando Ombao, Martin Lindquist, Wesley Thompson, and John Aston. *Handbook of neuroimaging data analysis*. Chapman and Hall/CRC, 2016.

[2] Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, 14(5):365–376, 2013.

[3] Anthony Steven Dick, Daniel A Lopez, Ashley L Watts, Steven Heeringa, Chase Reuter, Hauke Bartsch, Chun Chieh Fan, David N Kennedy, Clare Palmer, Andrew Marshall, et al. Meaningful associations in the adolescent brain cognitive development study. *NeuroImage*, 239:118262, 2021.

[4] Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper LR Andersson, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523–1536, 2016.

[5] Leah H Somerville, Susan Y Bookheimer, Randy L Buckner, Gregory C Burgess, Sandra W Curtiss, Mirella Dapretto, Jennifer Stine Elam, Michael S Gaffrey, Michael P Harms, Cynthia Hodge, et al. The lifespan human connectome project in development: A large-scale study of brain connectivity development in 5–21 year olds. *Neuroimage*, 183:456–468, 2018.

[6] Nora D Volkow, George F Koob, Robert T Croyle, Diana W Bianchi, Joshua A Gordon, Walter J Koroshetz, Eliseo J Pérez-Stable, William T Riley, Michele H Bloch, Kevin Conway, et al. The conception of the abcd study: From substance use to a broad nih collaboration. *Developmental cognitive neuroscience*, 32:4–7, 2018.

[7] Nora D Volkow, Joshua A Gordon, and Michelle P Freund. The healthy brain and child development study—shedding light on opioid exposure, covid-19, and health disparities. *JAMA psychiatry*, 78(5):471–472, 2021.

[8] Erik W van Zwet and Eric A Cator. The significance filter, the winner's curse and the need to shrink. *Statistica Neerlandica*, 75(4):437–452, 2021.

[9] Scott Marek, Brenden Tervo-Clemmens, Finnegan J Calabro, David F Montez, Benjamin P Kay, Alexander S Hatoum, Meghan Rose Donohue, William Foran, Ryland L Miller, Timothy J Hendrickson, et al. Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902):654–660, 2022.

[10] Junfei Zhao, Andrew X Chen, Robyn D Gartrell, Andrew M Silverman, Luis Aparicio, Tim Chu, Darius Bordbar, David Shan, Jorge Samanamud, Aayushi Mahajan, et al. Immune and genomic

correlates of response to anti-pd-1 immunotherapy in glioblastoma. *Nature medicine*, 25(3):462–469, 2019.

[11] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186, 2017.

[12] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[13] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295, 2015.

[14] Armin Schwartzman, Andrew J Schork, Rong Zablocki, and Wesley K Thompson. A simple, consistent estimator of snp heritability from genome-wide association studies. *Annals of Applied Statistics*, 2019.

[15] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569, 2010.

[16] Emmanuelle Génin. Missing heritability of complex diseases: case solved? *Human genetics*, 139(1):103–113, 2020.

[17] Peter M Visscher, Loic Yengo, Nancy J Cox, and Naomi R Wray. Discovery and implications of polygenicity of common diseases. *Science*, 373(6562):1468–1473, 2021.

[18] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[19] Timothy L Lash, Tyler J VanderWeele, and Kenneth J Rothman. Measurement and measurement error. *Modern epidemiology*, 4:287–314, 2021.

[20] James T Kennedy, Michael P Harms, Ozlem Korucuoglu, Serguei V Astafiev, Deanna M Barch, Wesley K Thompson, James M Bjork, and Andrey P Anokhin. Reliability and stability challenges in abcd task fmri data. *Neuroimage*, 252:119046, 2022.

[21] John R Cook and Leonard A Stefanski. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical association*, 89(428):1314–1328, 1994.

[22] Naisyin Wang, Xihong Lin, Roberto G Gutierrez, and Raymond J Carroll. Bias analysis and simex approach in generalized linear mixed measurement error models. *Journal of the American Statistical Association*, 93(441):249–261, 1998.

[23] Petros Dellaportas and David A Stephens. Bayesian analysis of errors-in-variables regression models. *Biometrics*, pages 1085–1095, 1995.

[24] Cathryn M Lewis and Evangelos Vassos. Polygenic risk scores: from research tools to clinical instruments. *Genome medicine*, 12(1):44, 2020.

[25] Samuel Thomas and Wanzhu Tu. Learning hamiltonian monte carlo in r. *The American Statistician*, 75(4):403–413, 2021.

[26] Ronald D Ruth. A canonical integration technique. *IEEE Trans. Nucl. Sci.*, 30(CERN-LEP-TH-83-14):2669–2671, 1983.

[27] Radford M Neal, S Brooks, A Gelman, GL Jones, XL Meng, et al. Handbook of markov chain monte carlo. *Press C, editor*, 22011, 2011.

[28] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 12 2007.

[29] Min Zheng, Jessecae K. Marsh, Jeffrey V. Nickerson, and Samantha Kleinberg. How causal information affects decisions. *Cognitive Research: Principles and Implications*, 5(1):6, 2020.

[30] Stefan Wager. Stats 361: Causal inference. Technical report, Technical report, Stanford University, 2020. URL: https://web. stanford. edu . . . , 2020.

[31] Gemma Hammerton and Marcus R. Munafò. Causal inference with observational data: the need for triangulation of evidence. *Psychological Medicine*, 51(4):563–578, 2021.

[32] Peter C Austin and Andreas Laupacis. A tutorial on methods to estimating clinically and policy-meaningful measures of treatment effects in prospective observational studies: A review. *The International Journal of Biostatistics*, 7(1):000010220215574679 1285, 2011.

[33] Stijn Vansteelandt and Oliver Dukes. Assumption-lean Inference for Generalised Linear Model Parameters. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):657–685, 07 2022.

[34] Weining Shen, Suyu Liu, Yong Chen, and Jing Ning. Regression analysis of longitudinal data with outcome-dependent sampling and informative censoring. *Scandinavian Journal of Statistics*, 46(3):831–847, 2019.

[35] M Alan Brookhart and Mark J Van Der Laan. A semiparametric model selection criterion with applications to the marginal structural model. *Computational statistics & data analysis*, 50(2):475–498, 2006.

[36] Peter C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424, 2011. PMID: 21818162.

[37] Ashkan Ertefaie, Nima S. Hejazi, and Mark J. van der Laan. Nonparametric Inverse-Probability-Weighted Estimators Based on the Highly Adaptive Lasso. *Biometrics*, 79(2):1029–1041, 07 2022.

[38] Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767, 2011.

[39] Heejung Bang and James M. Robins. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61(4):962–973, 12 2005.

[40] Jesse Hemerik, Jelle J. Goeman, and Livio Finos. Robust testing in generalized linear models by sign flipping score contributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):841–864, May 2020.

[41] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[42] Stephen R. Cole and Constantine E. Frangakis. The consistency statement in causal inference: A definition or an assumption? *Epidemiology*, 20(1), 2009.

[43] Anastasios A Tsiatis. Semiparametric theory and missing data. 2006.

[44] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

[45] Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 2021.

[46] Joseph P Romano and EL Lehmann. Testing statistical hypotheses, 2005.

[47] Suzie Cro, Tim P. Morris, Michael G. Kenward, and James R. Carpenter. Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: A practical guide. *Statistics in Medicine*, 39(21):2815–2842, 2020.

[48] Anna T. R. Legedza Cynthia M. DeSouza and Abdul J. Sankoh. An overview of practical approaches for handling missing data in clinical trials. *Journal of Biopharmaceutical Statistics*, 19(6):1055–1073, 2009. PMID: 20183464.

[49] Iosief Abraha, Antonio Cherubini, Francesco Cozzolino, Rita De Florio, Maria Laura Luchetta, Joseph M Rimland, Ilenia Folletti, Mauro Marchesi, Antonella Germani, Massimiliano Orso, Paolo Eusebi, and Alessandro Montedori. Deviation from intention to treat analysis in randomised trials and treatment effect estimates: meta-epidemiological study. *BMJ*, 350, 2015.

[50] National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials*. The National Academies Press, Washington, DC, 2010.

[51] Baptiste Leurent, Manuel Gomes, Rita Faria, Stephen Morris, Richard Grieve, and James R. Carpenter. Sensitivity analysis for not-at-random missing data in trial-based cost-effectiveness analysis: A tutorial. *PharmacoEconomics*, 36(8):889–901, 2018.

[52] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[53] Joseph G. Ibrahim and Geert Molenberghs. Missing data methods in longitudinal studies: a review. *TEST*, 18(1):1–43, 2009.

[54] Ping-Tee Tan, Suzie Cro, Eleanor Van Vogt, Matyas Szigeti, and Victoria R. Cornelius. A review of the use of controlled multiple imputation in randomised controlled trials with missing outcome data. *BMC Medical Research Methodology*, 21(1):72, 2021.

[55] Kang Hyun. The prevention and handling of the missing data. *Korean J Anesthesiol*, 64(5):402–406, 2013.

[56] Neil J Perkins, Stephen R Cole, Ofer Harel, Eric J Tchetgen Tchetgen, BaoLuo Sun, Emily M Mitchell, and Enrique F Schisterman. Principled Approaches to Missing Data in Epidemiologic Studies. *American Journal of Epidemiology*, 187(3):568–575, 11 2017.

[57] Roger A. Sugden. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 151(3):567–567, 1988.

[58] Peng Li, Elizabeth A. Stuart, and David B. Allison. Multiple Imputation: A Flexible Tool for Handling Missing Data. *JAMA*, 314(18):1966–1967, 11 2015.

[59] James R. Carpenter, James H. Roger, and Michael G. Kenward. Analysis of longitudinal trials with protocol deviation: A framework for relevant, accessible assumptions, and inference via multiple imputation. *Journal of Biopharmaceutical Statistics*, 23(6):1352–1371, 2013. PMID: 24138436.

[60] Peter C. Austin, Ian R. White, Douglas S. Lee, and Stef van Buuren. Missing data in clinical research: A tutorial on multiple imputation. *Canadian Journal of Cardiology*, 37(9):1322–1331, 2023/05/17 2021.

[61] Melissa J. Azur, Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49, 2011.

[62] Jonathan A C Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338, 2009.

[63] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011.

[64] Yu-Sung Su, Andrew Gelman, Jennifer Hill, and Masanao Yajima. Multiple imputation with diagnostics (mi) in r: Opening windows into the black box. *Journal of Statistical Software*, 45(2):1–31, 2011.

[65] Joseph L Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997.

[66] Donald B. Rubin. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6(1):34 – 58, 1978.

[67] JM Robins and N Wang. Inference for imputation estimators. *Biometrika*, 87(1):113–124, 03 2000.

[68] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[69] Nathaniel Schenker and Alan H Welsh. Asymptotic results for multiple imputation. *The Annals of Statistics*, 16(4):1550–1566, 1988.

[70] Jae Kwang Kim. Finite sample properties of multiple imputation estimators. 2004.

[71] Donald B Rubin. Multiple imputation. In *Flexible Imputation of Missing Data, Second Edition*, pages 29–62. Chapman and Hall/CRC, 2018.

[72] J. Kowalski and X.M. Tu. *Modern Applied U-Statistics*. Wiley Series in Probability and Statistics. Wiley, 2008.

[73] Katie Crist, Kelsie M. Full, Sarah Linke, Fatima Tuz-Zahra, Khalisa Bolling, Brittany Lewars, Chenyu Liu, Yuyan Shi, Dori Rosenberg, Marta Jankowska, Tarik Benmarhnia, and Loki Natarajan. Health effects and cost-effectiveness of a multilevel physical activity intervention in low-income older adults; results from the pep4pa cluster randomized controlled trial. *International Journal of Behavioral Nutrition and Physical Activity*, 19(1):75, 2022.

[74] Porchia Rich, Gregory A. Aarons, Michelle Takemoto, Veronica Cardenas, Katie Crist, Khalisa Bolling, Brittany Lewars, Cynthia Castro Sweet, Loki Natarajan, Yuyan Shi, Kelsie M. Full, Eileen Johnson, Dori E. Rosenberg, Melicia Whitt-Glover, Bess Marcus, and Jacqueline Kerr. Implementation-effectiveness trial of an ecological intervention for physical activity in ethnically diverse low income senior centers. *BMC Public Health*, 18(1):29, 2017.

[75] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[76] William H. Greene. *Econometric Analysis*. Pearson Education, fifth edition, 2003.